The combined Wordnet Bahasa

Francis Bond[⋆], Lian Tze Lim[◇], Enya Kong Tang[†] and Hammam Riza[‡]

[⋆]Nanyang Technological University, Singapore
[◇]KDU College Penang [†]Linton University College, Malaysia
[‡]Agency for the Assessment and Application of Technology, Indonesia
bond@ieee.org, liantze.lim@kdupg.edu.my, ektang@leg.edu.my, hammam.riza@bppt.go.id

This paper outlines the creation of an open combined semantic lexicon as a resource for the study of lexical semantics in the Malay languages (Malaysian and Indonesian). It is created by combining three earlier wordnets, each built using different resources and approaches: the Malay Wordnet (Lim & Hussein 2006), the Indonesian Wordnet (Riza, Budiono & Hakim 2010) and the Wordnet Bahasa (Nurril Hirfana, Sapuan & Bond 2011). The final wordnet has been validated and extended as part of sense annotation of the Indonesian portion of the NTU Multilingual Corpus (Tan & Bond 2012). The wordnet has over 48,000 concepts and 58,000 words for Indonesian and 38,000 concepts and 45,000 words for Malaysian.

1. Introduction[1]

This paper discusses the creation of a new release of the Open Wordnet Bahasa, an open-source semantic lexicon of Malay. The Wordnet is the result of merging with three wordnet projects for Malaysian and Indonesian, then adding data from other resources and finally by tagging a collection of Indonesian text.

Up until now, there has been no richly linked, broad coverage semantic lexicon for Malaysian and Indonesian. The Center of the International Cooperation for Computerization produced dictionaries for Malaysian and Indonesian (CICC 1994a,b) but the semantic hierarchy was limited to an upper level ontology for Indonesian. Multi-lingual lexicons such as **KIMD**, **FEM** or **KAMI** (Johns 2000; Lafourcade et al. 2003; Quah, Bond & Yamazaki 2001) do not include semantic relations: there is no way of knowing, for example, that a *harimau* 'tiger' is a kind of *karnivor* 'carnivore'. Building a wordnet, we can take advantage of the rich structure of the Princeton wordnet (Fellbaum 1998), and infer these relations for the Malay languages.

A wordnet is a semantic lexicon that follows the structure originally developed by the Princeton Wordnet of English (**PWN**) (Fellbaum 1998). Open class words (nouns, verbs, adjectives and adverbs) are grouped into synonym sets (**synsets** roughly equivalent to concepts) and these are linked by semantic relations such as hyponymy, meronymy and antonymy. Version 3.0 of the wordnet, has 117,700 synsets in all. Since the success of the English wordnet, wordnets have been developed for many languages and used in a wide variety of research (Bond & Paik 2012). Most wordnets are released under open licenses, which encourages their distribution, reuse and development.

[1] This research was supported in part by the joint JSPS/NTU grant on *Revealing Meaning Using Multiple Languages*, the Creative Commons Catalyst Grant: *Assessing the effect of license choice on the use of lexical resources* and the joint research agreement between Nippon Telegraph and Telephone Corporation and Nanyang Technological University. We would like to thank the anonymous reviewers, Ruli Manurung, Muhammad Zulhelmy bin Mohd Rosman, František Kratochvíl and David Moeljadi for their support and help.

Bond, Francis, Lian Tze Lim, Enya Kong Tang and Hammam Riza, 2014. 'The combined Wordnet Bahasa'. In Siaw-Fong Chung and Hiroki Nomoto, eds. *Current Trends in Malay Linguistics*. NUSA 57: 83–100. [Permanent URL: http://hdl.handle.net/10108/79286]

Wordnets have been used for research into measuring similarity of meaning and in a number of applications such as word sense disambiguation, information retrieval, automatic text classification, automatic text summarization, machine translation and automatic crossword puzzle generation.

We give an example of a synset, extended to include Malaysian and Indonesian, in Figure 1. This shows the entry for ***harimau*** $^{n}_{1}$.[2] The English wordnet gives many semantic relations: it is a hyponym of ***big cat*** $^{n}_{1}$ and in turn has hyponyms ***tiger cub*** $^{n}_{1}$, ***tigress*** $^{n}_{1}$ and ***Bengal tiger*** $^{n}_{1}$. It is a member of the ***genus Panthera*** $^{n}_{1}$ and has links to the Suggested Upper Merged Ontology (Niles & Pease 2001) as well as illustrations (from the Japanese wordnet project: Bond et al. 2009).

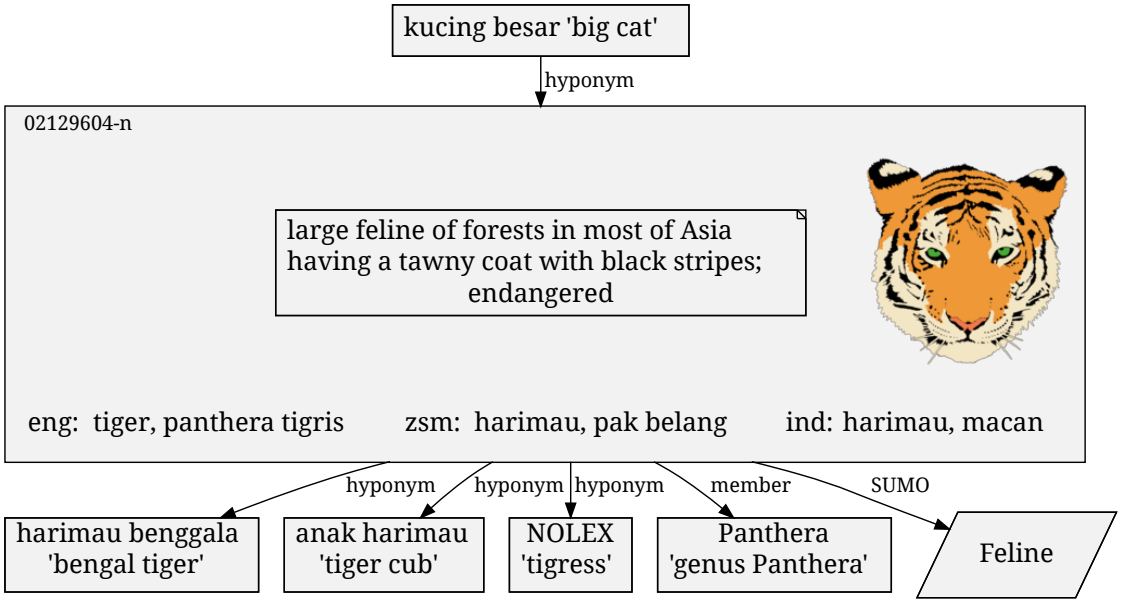

Figure 1. Wordnet Entry for *harimau* 'tiger'

We take three existing wordnets for the Malay languages: The Wordnet Bahasa (Nurril Hirfana, Sapuan & Bond 2011), the Malay wordnet (Lim & Hussein 2006) and the Indonesian wordnet (Riza, Budiono & Hakim 2010) and show that they can be combined to form a much richer resource. This combined wordnet is released under an open source license (the MIT license) 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. Bahasa Indonesia "the language of Indonesia" is very similar, and largely mutually intelligible. In this paper we will use **Malaysian** for standard Malay (the official language of Malaysia, ISO 639-3 code `zsm`), **Indonesian** to refer to the official language of Indonesia (`ind`) and **Malay** to refer to the generic Malay language that includes both (`msa`). Malay is the official language of four South Eastern Asian countries, namely Malaysia, Indonesia, Brunei and Singapore. Some people from

[2] Wordnet senses are shown in bold italics, with the part-of-speech (**n**oun, **v**erb, **a**djective, adve**r**b) as a superscript and the sense number a subscript.

the Philippines, Thailand, Burma, Sri Lanka, Cocos Island and Christmas Island also use it. There are about 40 million native Malay speakers worldwide.[3]

Originally orthographic conventions were quite different for Malaysian and Indonesian, with Malaysian based largely on English orthography and Indonesian on Dutch. The 1972 spelling reform harmonized the orthographic conventions, making the written forms very similar (Asmah Haji Omar 1975). Because of the enormous overlap in vocabulary we create a single wordnet for both languages. The vast majority of words are usable for both Malaysian 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 languages we can help to create a stronger lexical resource for the entire region.

The paper is organized as follows. In the following section we introduce the wordnets we will be working with. In Section 3 we discuss how we clean, merge and extend them to make one enhanced resource. In Section 4 we measure the coverage by annotating senses in a small corpus. We discuss the results, how the resource is being used and planned future work in Section 5 before finally concluding.

2. The wordnets

We will combine three wordnets, with a little more data from other resources. All three wordnets were built automatically, with limited manual clean-up.

The most common approaches to building a wordnet for a new language are automatic or semi automatic approaches. There are two main methods: **merge** and **extend** (Vossen 2005). The **merge** approach takes an existing monolingual resource and then maps it to the wordnet structure. The **extend** approach takes the synsets from Princeton WordNet (**PWN**), and then adds lemmas to them from the target language. This method allows the preservation of the original structure of the wordnet. All three wordnets opted for the extend approach, both because of its simplicity and because the resulting wordnet is automatically aligned to all other wordnets. Therefore, they take the basic semantic structure of English and add Malay words to the synonym-sets. For concrete nouns, this works very well: a ***tiger*** is an ***animal*** no matter what language we express it in: by annotating the synsets with Malay translations, we get the relation ship that *harimau* 'tiger' is a *binatang* 'animal'. For abstract nouns, verbs and adjectives we expect more differences between languages, but we leave representing these language specific differences for future work. For example, the closest concept to that expressed by the English word ***angry*** in Malay is ***marah***[4] but it is broader in meaning, including the emotion ***resentful***.[5] Identifying and expressing these subtle differences is a very hard task.

To account for some of the differences between Malay and English we have adopted two extra tags (inspired by the Basque wordnet (Pociello, Agirre & Aldezabal 2011)): `nolex` for a synset that is not-lexicalized in Malay or Indonesian (e.g. ***haircut***$^{n}_{1}$) and `autohypo` for a synset whose lemma is the same as its immediate hypernym (e.g. ***hen***$^{n}_{1}$, which for

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

[4] This is one of the words we currently link in the same synset.

[5] The differences in scope possibly related to cultural differences (Wierzbicka 1999:240), although that is beyond the current scope of our study.

Malay is the same as its hypernym ***chicken*** $^{n}_{2}$: Malay does not lexically distinguish between male and female animals).

In the following sections we introduce the wordnets and other resources we will use.

2.1 The Malaysian wordnet

The first wordnet built for Malaysian was by Lim & Hussein (2006) who built a wordnet for Malay. The prototype was based on sense alignments produced by hand aligning definitions from the Kamus Inggeris Melayu Dewan, an English-Malaysian dictionary (**KIMD** Johns 2000) and the Princeton Wordnet (v1.6). Words whose definitions were aligned with a synset, were assigned to that synset. For example, the following entry from **KIMD** (1) is aligned with the wordnet synset with ID 10025218-n (2).[6]

(1) (***dot***, noun, 1, [small round spot, small circular shape], ⟨titik, bintik⟩). (**KIMD**)

(2) (10025218-n, ***dot***, noun, 1, [a very small circular shape]). (**PWN**)

This alignment allows one to simply add ***titik*** and ***bintik*** to synset ***dot*** $^{n}_{1}$[7] in the basic extend approach. Based on the aligned glosses, Lim & Hussein (2006) built 12,429 noun synsets and 5,805 verb synsets.

Lim & Hussein (2006) point out several issues with the resulting wordnet. One is that the coverage is incomplete, and very much depends on which words could be aligned. Another problem is the nature of the dictionary used. **KIMD** is a unidirectional English to Malay dictionary and not all the Malaysian equivalents it provides are valid lemmas. For example, **KIMD** provides ***orang, anggota, dan lain-lain yang tidak hadir*** (literally 'person, member, etc. who are not present') as the Malaysian equivalent for English ***absentee***. While this is valid as a definition, it is not lexicalized in Malaysian and is not suitable as a lemma of a synset.

Further, the wordnet is missing much useful information that the English wordnet has, including:

1. Malaysian definitions for the glosses
2. Verb frames (such as *Somebody —s something*)
3. Sense frequencies

2.2 The Indonesian wordnet

There have been two projects which built Indonesian wordnets. The first adopted the expand approach, and created a small prototype which was never released (Putra, Arfan & Manurung 2008). The second also used the expand approach, and then corrected entries using the infrastructure from the Asian Wordnet Project (Riza, Budiono & Hakim 2010). The Indonesian Wordnet at the Asian Wordnet currently has 33,726 synsets; 38,394

---

[6] The English headword is shown in bold italics; noun is the part of speech; 1 is the sense number: this is the first sense; the definition is in square brackets: []; the Malay lemmas are in angle brackets for KIMD: ⟨⟩.

[7] This represents the first sense of the noun ***dot***: the entry shown in (2).

words and 65,206 senses (word-synset pairs).[8] The lexicons used to expand were bilingual English-Indonesian.

We used the version of the lexicon from the Asian Wordnet Project (Riza, Budiono & Hakim 2010), which contained some 25,755 synsets. Indonesian lemmas were added by human translators using online dictionaries and MT systems as aids. The infrastructure from the Asian Wordnet Project was used as an online cooperative tool for the compilation (Sornlertlamvanich et al. 2008).

Riza, Budiono & Hakim (2010) noted that suitable Indonesian lemmas do not exist for all synsets.

2.3 Wordnet Bahasa

The third wordnet also took the expand approach, but used a multiple-pivot approach, aligning Malaysian to English using additional languages (French and Chinese) and semantic codes as extra information to constrain the translations (Nurril Hirfana, Sapuan & Bond 2011).

Wordnet Bahasa used two lexicons: **FEM** (Lafourcade et al. 2003:the French-English-Malaysian Lexicon), which contains entries with French, English and Malaysian as well as hypernyms in French; and **KAMI**, which contains Malaysian, English and Chinese as well as semantic classes from the Goi-Taikei ontology (Quah, Bond & Yamazaki 2001). These were linked with three wordnets: one for English (**PWN**), one for Chinese (Xu et al. 2008) and one for French.[9] To map between the Goi-Taikei ontology and wordnet, we used the mappings produced by CoreNet (Kang et al. 2010).

The construction broadly followed the matching through multiple pivot approach of Bond & Ogura (2007). Each Malaysian word in the lexicon, is linked through pivots to every synset that has the same part-of-speech. There are three pivots for this: the English term, the French or Chinese term and the hypernym. After linking through the terms, semantic classes are used to see if the hypernym is compatible with the synset's hypernyms.

Here is an example for the following entries:

(3) Entry in **FEM**

lexical entry	
Malaysian	busur
English	bow
French	arc
Part-of-Speech	noun
Hypernym	arme 'weapon'

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

[9] The French wordnet was made by merging the French Wordnet from EuroWordNet (Vossen 1998) and the new Wordnet Liberé du Français (WOLF: Sagot & Fišer 2008). As these map to different versions of the English WordNet, they were harmonized using the mappings from Daude, Padro & Rigau (2003).

(4) Entry in **KAMI**

lexical entry	
Malaysian	busur
English	bow
Chinese	弓
Part-of-Speech	noun
Hypernym	⟨940 : *worktool*⟩

(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[10] 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), they 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). They 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). There is thus a strong match to the correct synset.

When we come to the **KAMI** entry for {busur, bow, 弓} (4), they look up wordnet and also find one entry (5a) that matches in two languages, and several that match in only one (we only show 5b). Consulting 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.[11] We thus have a reasonable link to the correct synset, and only weak links to the others.

---

[10] This is in fact an error, it means 'archery' and should be in a different synset.

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

After matching all the candidates, thresholds were set based on the number of languages matched, the hypernym matching and the amount of ambiguity (for more details see Nurril Hirfana, Sapuan & Bond 2011). Finally the 5,000 most common synsets used in the British National Corpus[12] (Fellbaum & Vossen 2007) were hand checked. During this process, candidates that were only used in either Malaysian or Indonesian were marked as such. The default assumption was that a sense (synset-word) mapping can be used in either Malaysian or Indonesian.

The resulting Wordnet Bahasa had 19,207 synsets, 48,111 senses and 19,460 unique words (counting hand-checked and high-quality automatic candidates). 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.

All three wordnets basically have the same structure: Malay lemmas added to the Princeton Wordnet structure, the main difference is in their coverage.

2.4 Comparison of the wordnets

We compare the three wordnets according to four criteria: size, precision, coverage of high frequency concepts, and coverage of an Indonesian Corpus. The results are given in Table 1 for all three wordnets and a merged wordnet created taking the union of all entries.

Table 1. Sizes and coverage of the wordnets (ind: Indonesian; zsm: Malaysian; msa: Malay)

Wordnet	Lang	Synsets	Words	Senses	Precision	Core	Corpus
Indonesian Wordnet	ind	27,506	30,358	57,560	67%	46.7%	69.0%
Malaysian Wordnet	zsm	23,953	23,833	48,996	83%	82.9%	76.1%
Wordnet Bahasa	msa	19,347	19,572	48,181	85%	97.7%	76.0%
Union	msa	52,805	66,364	146,463	78%	99.5%	85.9%

Synsets, words and senses are the numbers of synsets with at least one Malay word, the number of unique lemmas and the number of senses (synset-lemma combinations). Precision was tested by a bilingual English/Malay speaker[13] looking at a randomly selected sample of 100 senses and marking them as appropriate or not (and also whether they were a lemma or definition).

Core gives the percentage of synsets in the 5,000 most common synsets used in the British National Corpus[14] (Fellbaum & Vossen 2007) for which an entry existed. Because the distribution of word senses is Zipfian, we expect the most frequent senses to cover a disproportionally large percentage of actual tokens.

---

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

[13] A final-year undergraduate student at Nanyang Technological University who has taken some linguistics courses.

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

Finally, we tested coverage against the Indonesian portion of the NTU Multilingual Corpus (Tan & Bond 2012). This consists of 2,197 sentences taken from web pages introducing Singapore to tourists. It contains 58,058 tokens (including punctuation), tagged with parts of speech from the Indonesian Tagset I (Pisceldo, Manurung & Mirna 2009) (as well as parallel text in English, Chinese, Japanese, Korean and Vietnamese that was not used here). We carried out some morphological analysis to get the lemma: stripping the clitic *-nya* from nouns and the passive prefix *di-* from verbs. Any word tagged as a negative marker (mainly *tidak*), adjective, adverb, noun, proper name, foreign word or main verb was treated as an open class word:[15] there were 36,470 of these. We checked against the English wordnet to identify around 3,000 as English terms. The remainder we considered should have an entry in wordnet: we measure coverage by seeing if the word is in the wordnet, without checking that the correct sense is there.

The results show that the Indonesian wordnet is larger, but with more noise, and less coverage of common words. The Malaysian and Bahasa wordnets are more similar in terms of size, coverage and accuracy. Simply unifying all three gave a much better coverage, with a reasonable accuracy (78%). Surprisingly, there was very little overlap in senses: even if the wordnets had lemmas for the same synsets, they were different 94.7% of the time. This shows one of the weaknesses of opportunistic methods: even if a lemma is found for a synset, it may often be the case that some are missing: combining many resources helps to alleviate this problem.

3. Combining and expanding the wordnets

In order to make a more useful resource, with better cover for both Malaysian and Indonesian, we determined to merge the wordnets, cleaning up any problems as we found them. We also added some extra entries, as described below.

3.1 Cleaning

When we evaluated the precision of the wordnets, we noticed a few systematic errors, detailed below:

1. Sometimes the lemma is just the English definition (followed by semicolon)
   - 01739099-v according to a plan;
2. The definition is included in the lemma (in brackets)
   - 09500936-n phonix [burung dalam legenda Arab] ‘Phoenix [bird of Arab legend]’
3. The lemma is a Malay definition (preceded by the domain in brackets)
   - 13996211-n (Rusia) kebebasan ‘(Russia) freedom’
4. The lemma is a Malay definition (not a word)
   - 00059127-n tindakan melarikan diri dari sesuatu

We removed all entries where the Malay lemma was the same as the English definition. We used regular expressions to identify definitions in brackets, and split these entries into a lemma and definition. Finally, if the Malay lemma was four or more words, we

[15] POS tag: `neg, prn, jj, rb, fw, vb.*, nn.*`: we tried to match the slightly idiosyncratic choice of words covered by **PWN**.

Table 2. Data from the Unicode Common Locale Data Repository

Type	Tag	Eng	Ind	Zsm	Synset
language	ca	Catalan	Katalan	Catalonia	06967529-n
territory	HR	Croatia	Kroasia	Croatia	08815858-n
day	mon	Monday	Senin	Isnin	15163979-n
month:gregorian	8	August	Agustus	Ogos	15212455-n

reclassed it as a definition. In this way, we ended up with 8,200 definitions (and lost some lemmas).

3.2 Merging

After the cleaning, the wordnets were merged. We mapped the Malaysian Wordnet to **PWN** version 3.0 using the mappings provided by Daude, Padro & Rigau (2003); the Indonesian and Bahasa wordnets were both based on version 3.0.

For each entry, we had to decide if it was used in Indonesian, Malaysian or both. If an entry was in both the Malaysian and Indonesian wordnets, then we marked it as both. If it appeared only in Malaysian or Indonesian, then we searched for it in the Indonesian and Malaysian wikipedias. If it was a single word and appeared seven or more times or if it was a multi-word expression and appeared at least once then we marked it as being usable in both varieties. The thresholds were arrived through manual inspection of different values, and are slightly conservative. If we could not find enough evidence for its use in the other variety, we left as Malaysian if it came from the Malaysian wordnet and Indonesian if it came from the Indonesian wordnet or followed the information in the Bahasa Wordnet.

3.3 Expanding

After merging the three wordnets, we then added information on languages, territories and dates from the Unicode Common Locale Data Repository, on person's names from the English wordnet, some translations from Wikipedia and a few entries we found missing in corpora.

3.3.1 Unicode Common Locale Data Repository

The Unicode Common Locale Data Repository (CLDR),[16] contains the standard names of languages, territories and dates to be used as building blocks for software to support the world's languages. It is designed to be used for tasks such as choosing languages or countries by name. The Open Multilingual Wordnet Project (Bond & Foster 2013) extracted this data for each language and linked it to wordnet, using the Princeton wordnet as a pivot. This gives entries for language and territory names, as well as days of the week and months. Interestingly there are significant differences in Malaysian and Indonesian for these proper names, we give some examples in Table 2, where the type is the kind of data, and the tag is the tag used by the CLDR.

[16] `http://cldr.unicode.org/`

```
<page>                                  <page>
  <title>Marikh</title>                   <title>Laut Kaspia</title>
  <text>...                               <text>...
    {{Infobox Planet                       [[Kategori:Tasik di Eropah|Kaspia]]
    ....                                   [[Kategori:Tasik di Rusia|Kaspia]]
    }}                                     [[Kategori:Tasik di Asia|Kaspia]]
    ...                                     ...
    [[en:Mars]]                             [[en:Caspian Sea]]
    [[es:Marte (planeta)]]                  [[es:Mar Caspio]]
    ....                                    ....
  </text>                                 </text>
</page>                                 </page>
```

Figure 2. Excerpts from Malaysian and Indonesian Wikipedia article dumps

There were about 500 entries for each language, and these were added to the wordnet, so it now has the names of most major territories, languages, months and weekdays.

3.3.2 Person names

Finally, as modern Malaysian uses the Latin alphabet, it is normal for foreign names to be written as is: for example, *Sherlock Holmes* or *Albert Einstein*. We therefore decided to add all instances of people, either real or fictional, who are already in the Princeton wordnet, to the wordnets. Specifically, this involved all hyponyms of ***person***$^{n}_{1}$ and ***fictional character***$^{n}_{1}$ which are leaf nodes that are immediately dominated by an `instance hypernym` link.

3.3.3 Wikipedia article title translations

Each Wikipedia article contain links to articles written in different languages about the same topic, thus providing a free repository of multilingual translations (often about named entities). Wikipedia articles also contain useful information such as infoboxes and page categories, as shown in Figure 2.

We added entries to the combined wordnet by looking up Indonesian and Malaysian Wikipedia article titles as follows:

1. For the title for each Indonesian and Malaysian Wikipedia article, look up its corresponding English title in **PWN**.
2. If only one synset is found, map the Indonesian/Malaysian title to it.
3. If multiple synsets are found, we compare the hypernyms chain of each synset to the semantic type and categories of the Wikipedia article. The first synset whose hypernym chain contains the semantic type or one of the categories, is chosen as the synset to be mapped to.

This resulted in the following additions:

- 8,480 new mappings
- 3,725 new synsets
- 732 new Malay entries (i.e. used in both Malaysian and Indonesian)
- 2,109 new Malaysian entries
- 5,473 new Indonesian entries

3.3.4 Corpus-based additions

Finally, we also added by hand the most common unknown words from the corpus (all open class words not found in the combined wordnet with a frequency greater than four). These included entries such as ***populer*** $^{a}_{1}$ (Ind.) 'popular'. There were approximately 500 of these.

3.4 The combined wordnet

Table 3. Sizes of the wordnets (ind: Indonesian; zsm: Malaysian)

Wordnet	Lang	Synsets	Words	Senses
Indonesian Wordnet	ind	27,506	30,358	57,560
Malaysian Wordnet	zsm	23,953	23,833	48,996
Wordnet Bahasa	ind	19,316	19,522	48,111
	zsm	19,347	19,572	48,181
Combined	ind	48,689	58,541	133,005
	zsm	38,736	45,664	114,025

The combined wordnet has 8,200 definitions;
85,315 of the senses are shared between Indonesian and Malay.

The size of the combined wordnet is shown in Table 3, separated into Malaysian and Indonesian to allow comparison. It is released under an open source license (MIT) and can be looked up on-line or downloaded from `http://wn-msa.sourceforge.net/`. The new wordnet has much better coverage than any of the component wordnets, with almost twice as many synsets.

The extended combined wordnet covers 90.6% of the open class tokens in the corpus, an improvement of 4.7% over the simple merge. Roughly half of the remaining unknown words are proper nouns.

4. Sense tagging

The next stage in the creation of the expanded wordnet is to verify its coverage on some text. We chose to annotate text from the NTU Multilingual Corpus (Tan & Bond 2012). The main reason for this was to allow further research in comparing sense distributions across languages: the subset we used consists of web pages introducing Singapore to tourists, originally in English but translated into Chinese, Japanese, Indonesian, Korean and Vietnamese. In this paper we only discuss the Indonesian, but the Chinese, English and Japanese portions have also been tagged. The Indonesian text consists of 2,197 sentences. It contains 58,058 tokens (including punctuation). There are 36,470 open class words. When we tag with wordnet senses, we can also tag multiword expressions, so a word can potentially be part of two different entries. Because of this, there are a total of 38,102 potential synsets. Looking at types, around 7,300 unique open-class words appear in the corpus, and they are tagged with 6,100 distinct senses.

Tagging is done with a web interface, and the text is gone through sequentially, so that the annotators can see the context for each sentence. We give an example of a tagged sentences in (6: sentence number 1118). The first line is the Indonesian sentence, the

second line is the lemma (for open class words only) and the third line is the wordnet synset assigned by the annotator (shown here using the English synset label). Although we show the synsets with English labels, we could also have shown them with Malay (or Chinese, Japanese, Finnish, …), the actual tag is the synset ID which is shared across languages. There is one (discontinuous) multiword expression *di bawah* 'below'. One word is tagged as **p**: closed class part of speech. In this sentence the word *semtara* is a conjunction, and thus does not get a tag, although there are also potentially open class senses, such as *sementara* 'temporary'. Note that lemmatization only changes the form in two places: the reduplicated *toko-toko* 'shops' is reduced to *toko* 'shop' and *atasnya* to *atas*.

(6) *Sejumlah restoran lokal dan internasional yang populer berada*
sejumlah restoran lokal internasional populer berada
batch**$^{n}_{3}$ ***restaurant$^{n}_{1}$ ***local**$^{a}_{2}$ ***international***$^{n}_{1}$ ***popular***$^{a}_{1}$ ***be***$^{v}_{3}$

di lantai bawah , sementara toko-toko pakaian memenuhi lantai
di$_i$ lantai bawah$_i$ sementara toko pakaian memenuhi lantai
below$^{a}_{1i}$ ***floor***$^{n}_{2}$ ***bottom***$^{a}_{1}$ p ***shop***$^{n}_{1}$ ***clothes***$^{n}_{1}$ ***fill***$^{n}_{3}$ ***floor***$^{n}_{2}$

atasnya .
atas .
upper$^{a}_{1}$

'A number of popular local and international eateries take up the lower floors, while clothes shops dot the higher levels.'

The two words marked with asterisks were those where the word was in the lexicon, but not with the correct sense. The wordnet had *sejumlah* 'batch process' as a verb, but not as a noun. Similarly, it has the sense for *lokal* that meant 'affecting only a restricted part or area of the body' but not the sense that meant 'of or belonging to or characteristic of a particular locality or neighborhood'. In these cases, the concept itself existed in the wordnet, but was not linked to this word: the annotator simultaneously links it and tags it. These gaps are extremely hard to find, not just for wordnets but for any lexical resource (Baldwin et al. 2004). Because of this, we think it is essential to tag a corpus as we develop the wordnet, and to keep it in sync with the wordnet, in the same way that grammarians create a treebank along with their grammars: without empirical testing you cannot know the coverage (Oepen, Flickinger & Bond 2004).

Approximately 20% of the open class tokens were not in the combined wordnet. Of these 10% were names and or English expressions, such as *fried hokkien mee* or *ultra-modern*. 6% of the remainder had entries in the English wordnet, and could simply be linked. Only about 4% (150 tokens) really needed entirely new entries. An example is *sate* 'satay: skewers of meat grilled over charcoal served with raw onions and cucumber'. We intend to add these to the Wordnet Bahasa and, where appropriate (as in this case), the Princeton Wordnet. Only 2.5% of tokens had entries in the wordnet but without the correct sense: of these perhaps half (300 tokens) need new concepts not in the Princeton Wordnet, the rest are simply linked as they are tagged.

We are currently in the final stages of checking the annotation and adding the missing senses to the wordnet. When this is completed, the corpus will be released both along with the wordnet and as part of the NTU Multilingual corpus. We will also use the corpus to calculate sense frequencies. This will be used both in the lexicon interface, where the

most frequent sense will be listed first, and for word sense disambiguation, where most-frequent sense is a very strong baseline: most of the time, a word is used with its most common meaning.

5. Discussion

The three individual wordnets combined here all lacked enough coverage to be fully useful on their own, with over 25% of the tokens in a corpus not covered. Merging and extending the dictionaries and combing them with English has increased this coverage to over 90%. Further, sense tagging a corpus has allowed us to fill in missing gaps and add frequency information to the senses. We can thus calculate the standard baseline for word sense disambiguation: most frequent sense.

Therefore, even though the wordnet described here is far from complete in coverage or richness of information, it is a significant advance in the creation of lexical resources for Indonesian and Malay.

5.1 Availability

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 between wordnet and different ontologies. Many of these have been released freely, some of these we were granted permission to use especially for this research. Granting access to resources makes possible entirely new applications and so should be encouraged.

The combined wordnet is released as version 1.0 of the Wordnet Bahasa under the MIT license[17] (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 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.

The combined wordnet uses the existing infrastructure from the Japanese and Bahasa wordnet projects. We show a screenshot in Figure 3. It has been successfully included in the open multilingual wordnet[18] where it is linked to wordnets in 25 other languages. There are interfaces to access it for JAVA, Perl, Python, Scheme, Ruby and more.

---

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

[18] `http://compling.hss.ntu.edu.sg/omw/`

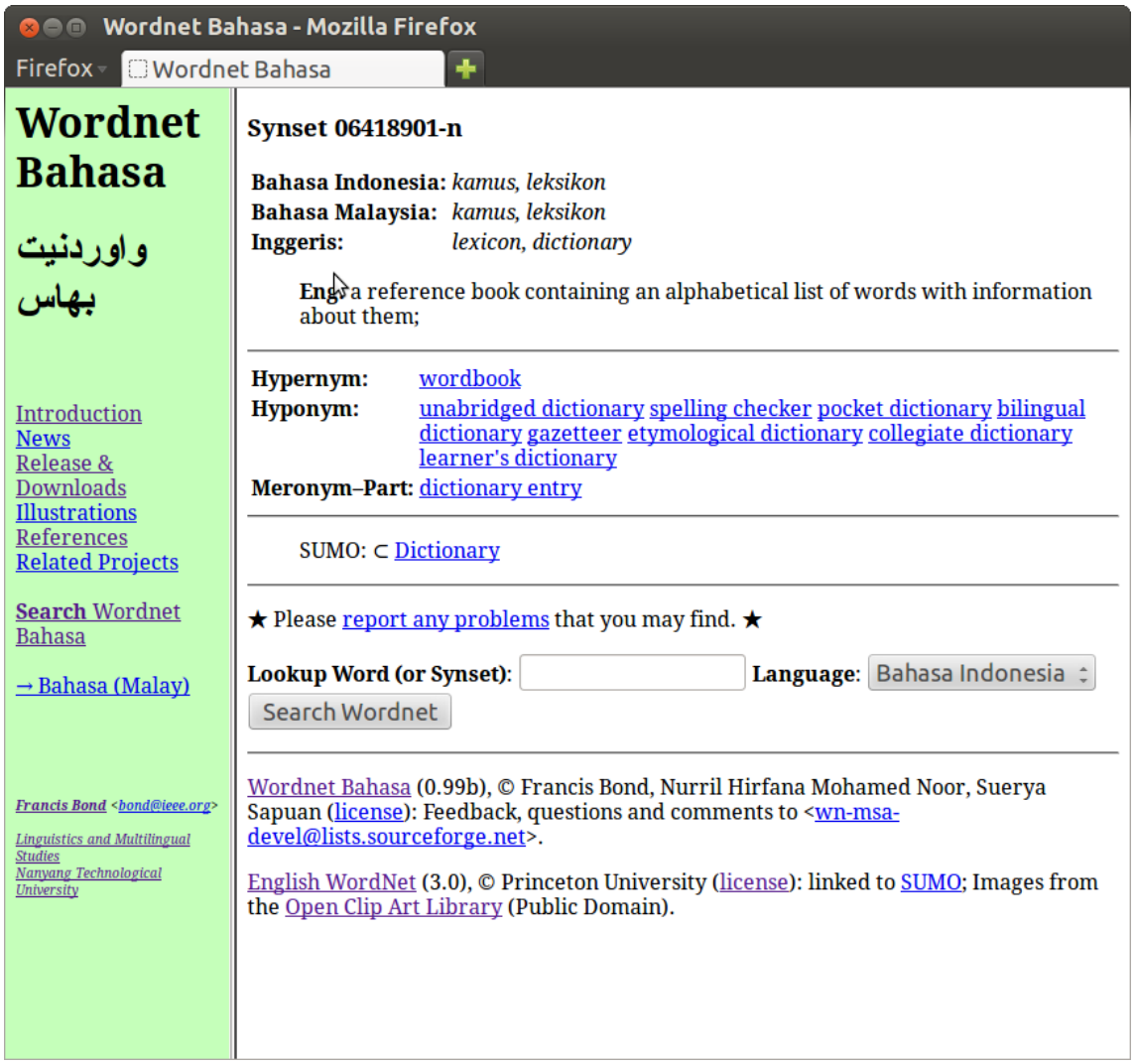

Figure 3. Screenshot for the Wordnet Entry for *kamus* 'dictionary'

5.2 Use of the resource

The combined wordnet Bahasa has been used to support research in a variety of topics. It has been used to help analyze Malay Tweets (Saloot, Idris & Mahmud 2014) as well as for general Malay semantic processing (Chu et al. 2014). Through the Open Multilingual Wordnet it is being made available in version 3.0 of the *Natural Language Processing Toolkit* (Bird, Klein & Loper 2009), an extremely popular textbook.[19]

In more linguistic research, it has been used to model the decompositional semantics of pronouns (along with analyses for Chinese, English and Japanese: Seah & Bond 2014). Finally, it has been linked to the **Semantic Domains** from SIL International (Muhammad Zulhelmy, Bond & Kratochvíl 2014).[20] The semantic domains are designed to aid in the rapid construction and subsequent organization of lexicons for languages which may have no dictionary at all. We are using them to create a lexicon and wordnet for Abui.[21]

5.3 Further work

This is only one step toward creating a complete wordnet for Malaysian and Indonesian: much more can be done to improve it. We intend to continue our research on the Wordnet Bahasa in multiple locations in Indonesia, Malaysia and Singapore, so that we can all contribute to a single rich lexical resource. Concretely, we would like to add more Malaysian and Indonesian definition sentences to make the wordnet more accessible to Malay speakers. Simultaneously, we wish to tag more corpora with this WordNet in order to get more frequency information and further check for gaps in coverage.

In addition to straightforward increases in terms of size and accuracy, we intend to enrich the structure of the wordnet in the following ways. 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.

Currently we under-specify the language for most entries in our master database, and output two fully specified versions of the dictionary (Malaysian and Indonesian) for applications. As there is a great deal of overlap, 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 Malaysian or Indonesian as required. We hope that this can be extended to cover other dialects of Malay, beyond the two standard varieties.

Finally, as we add words, synsets and relations not in English, the structure will move away from that of the Princeton Wordnet. We would like to build a graph containing only the Malay structures (using, for example, the approach of Vincze & Almázi (2014) to

---

[19] `http://www.nltk.org/howto/wordnet.html`

[20] "SIL International is a [Christian] faith-based nonprofit organization committed to serving language communities worldwide as they build capacity for sustainable language development." `http://sil.org`

[21] ISO 639-3 abz: a language spoken by approximately 16,000 speakers in the central part of the Alor Island in Eastern Indonesia.

ignore nodes not lexicalized in Malay) and investigate its structure and connectivity along the lines of Steyvers & Tenenbaum (2005).

6. Conclusions

We have produced a single wordnet that combines Standard Malaysian and Indonesian into a single semantic lexicon, only marking those entries where the Malaysian language and Indonesian language were differentiated. It covers over 85% of the open class tokens in typical text. This wordnet will serve as a platform for further work in the Malay languages and will be further extended through cooperation in Malaysia, Indonesia and Singapore.

References

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

Baldwin, Timothy, Emily M. Bender, Dan Flickinger, Ara Kim & Stephan Oepen. 2004. Road-testing the English resource grammar over the British National Corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, 2047–50. Lisbon.

Bird, Stephen, Ewan Klein & Edward Loper. 2009. *Natural language processing with Python*. California: O'Reilly. (`www.nltk.org/book`).

Bond, Francis & Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association or Computationa Linguistics (ACL-2013)*, 1352–1362. Sofia.

Bond, Francis, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi & Kyoko Kanzaki. 2009. Enhancing the Japanese WordNet. In *The 7th Workshop on Asian Language Resources (ACL-IJCNLP 2009)*, 1–8. Singapore.

Bond, Francis & 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).

Bond, Francis & Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global Wordnet Conference (GWC 2012)*, Matsue. 64–71.

Chu, Benjamin, Qiang Liu, Rohana Mahmud, Arun Anand, Weng Onn Kow & Dickson Lukose. 2014. Malay semantic text processing engine. In *eKNOW 2014, the Sixth International Conference on Information, Process, and Knowledge Management*, 38–43. Barcelona.

CICC. 1994a. Research on Indonesian dictionary. Tech. Rep. 6—CICC—MT53 Center of the International Cooperation for Computerization Tokyo.

CICC. 1994b. Research on Malaysian dictionary. Tech. Rep. 6—CICC—MT54 Center of the International Cooperation for Computerization Tokyo.

Daude, Jordi, Lluis Padro & 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.

Fellbaum, Christine (ed.). 1998. *WordNet: An electronic lexical database*. Massachusetts: MIT Press.

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

Johns, A. H. (ed.). 2000. *Kamus Inggeris Melayu Dewan*. Kuala Lumpar: Dewan Bahasa dan Pustaka.

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

Lafourcade, M., G. Seſasset, L. Metzger, A. Rahman & 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 & 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*, 13–16.

Muhammad Zulhelmy bin Mohd Rosman, Francis Bond & František Kratochvíl. 2014. Bringing together over- and under-represented languages: Linking wordnet to the SIL semantic domains. In *Proceedings of the 7th Global Wordnet Conference (GWC 2014)*, 40–48. Tartu.

Niles, Ian & Adam Pease. 2001. Towards a standard upper ontology. In Chris Welty & Barry Smith (eds.), *Proceedings of the 2nd International Conference On Formal Ontology In Information Systems (FOIS-2001)*, Maine.

Nurril Hirfana Mohamed Noor, Suerya Sapuan & Francis Bond. 2011. Creating the open Wordnet Bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, 258–267. Singapore.

Oepen, Stephan, Dan Flickinger & Francis Bond. 2004. Towards holistic grammar engineering and testing — grafting treebank maintenance into the grammar revision cycle. In *Beyond shallow analyses — formalisms and statistical modelling for deep analysis (workshop at IJCNLP-2004)*, Hainan Island. `http://www-tsujii.is.s.u-tokyo.ac.jp/bsa/`.

Pisceldo, Femphy, Ruli Manurung & Adriani Mirna. 2009. Probabilistic part-of-speech tagging for bahasa Indonesia. In *Third International MALINDO Workshop*, Singapore.

Pociello, Elisabete, Eneko Agirre & Izaskun Aldezabal. 2011. Methodology and construction of the Basque wordnet. *Language Resources and Evaluation* 45(2). 121–142.

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

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

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

Sagot, Benoît & 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.

Saloot, Mohammad Arshi, Norisma Idris & Rohana Mahmud. 2014. An architecture for Malay Tweet normalization. *Information Processing & Management* 50(5). 621–633.

Seah, Yu Jie & Francis Bond. 2014. Annotation of pronouns in a multilingual corpus of Mandarin Chinese, English and Japanese. In *10th Joint ACL–ISO Workshop On Interoperable Semantic Annotation*, Reykjavik.

Sornlertlamvanich, Virach, Thatsanee Charoenporn, Kergrit Robkop & Hitoshi Isahara. 2008. KUI: Self-organizing multi-lingual wordnet construction tool. In Attila Tanács, Dóra Csendes, Veronika Vincze, Christiane Fellbaum & Piek Vossen (eds.), *4th Global Wordnet Conference: GWC-2008*, 417–427. Szeged, Hungary.

Steyvers, Mark & Joshua B. Tenenbaum. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science* 29. 41–78.

Tan, Liling & Francis Bond. 2012. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). *International Journal of Asian Language Processing* 22(4). 161–174.

Vincze, Veronika & Attila Almázi. 2014. Non-lexicalized concepts in wordnets: A case study of English and Hungarian. In *Proceedings of the 7th Global Wordnet Conference (GWC 2014)*, 118–126. Tartu.

Vossen, Piek (ed.). 1998. *Euro wordnet*. Kluwer.

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

Wierzbicka, Anna. 1999. *Emotions across languages and cultures: Diversity and universals*. Cambridge: Cambridge University Press.

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