Digitising Dictionaries for Advanced Look-up and Lexical Knowledge Research in Malay

LIM Lian Tze, TAN Ewe Hoe and TANG Enya Kong
{liantze, ewehoe, enyakong}@cs.usm.my

Unit Terjemahan Melalui Komputer
School of Computer Sciences
Universiti Sains Malaysia
Penang, Malaysia

Abstract

Electronic dictionaries need not be mere OCR digitised versions of their paper-form counterparts: they can be made more computer-tractable to facilitate more meaningful operations and data exchange. For instance, explicitly annotating different fields in a dictionary entry allows more targeted look-ups, as we will show using *Kamus Dewan* as an example. Dictionary data can also be re-organised to enable semantic-based search. The wordnet lexical database is one such model, for which we created a prototype for the Malay language. As both the proposed annotated *Kamus Dewan* and Malay WordNet are compiled according to established standards and guidelines, the data can be aligned with similar lexical resources of other languages. This provides a means for mutual sharing, interchange and enrichment of lexical data and knowledge between Malay and other languages.

1 Introduction

Dictionaries contain rich lexical knowledge for a language. The advent of Information Technology has seen many paper dictionaries being digitised, thus greatly speeding up word look-ups by human users, either using a local computer or via an online interface on the World Wide Web (WWW). Such electronic dictionaries can also be integrated into various office productivity applications to facilitate automatic spell-checking.

Most electronic dictionaries allow searches by headwords (stemmed or otherwise), and sometimes by a full-index text search of the full entry of the headword. Entries returned from a search are usually presented with formatting effects (e.g. bold/italic typefaces, larger font sizes) so that human users may distinguish each field (gloss text, example usage, etc). However, these formatting effects serve only as stylistic presentations and do not distinguish the fields or their structure explicitly. For example, the example usage of a word, a scientific name for an organism and a subentry for a phrasal expression containing the same word may all be italicised, without further

annotation of which is which.[1] Such problems prevent users from performing more targeted searches, as well as other computer applications from fully utilising the data in dictionaries. This can be overcome by annotating the field and structure of dictionary entries explicitly, as we will show with XML-annotated examples from *Kamus Dewan* [1], based on the Text Encoding Initiative guidelines [6, 7] in Section 2.

Taking things one step further, the content of existing dictionaries can be re-organised to produce lexical resources that are more semantics-oriented. Entries in conventional paper dictionaries are organised alphabetically by the headwords as this provides the most efficient manual look-up method for humans. In contrast, computers do not have such limitations, thus searches based on different organisational models are just as efficient. The Computational Linguistics community has produced various (computer-readable) lexical resources based on different models, mostly for the English language. Similar resources may be created for the Malay language to encourage further lexical data and knowledge exchange with other languages. As an example, we have created a Malay WordNet prototype based on Princeton's WordNet [4, 11], which is also aligned with wordnet systems of other languages and described in Section 3.

2 Logical Annotation of *Kamus Dewan*

Kamus Dewan (KD) [1], published by Dewan Bahasa dan Pustaka, is the most authoritative dictionary for the Malay language. A commercial electronic version is available from The Name Technology Sdn. Bhd.,[2] while an online look-up interface can be found at `http://www.karyanet.com.my`. In this section, we demonstrate how searches based on these electronic versions of KD can be further enhanced by annotating the logical structure of KD entries.

2.1 Annotating KD with TEI

The Text Encoding Initiative (TEI) Guidelines are "an international and interdisciplinary standard that enables libraries, museums, publishers, and individual scholars to represent a variety of literary and linguistic texts for online research, teaching, and preservation" [6, 7]. The Guidelines provide schemas for annotating different texts, including prose, print dictionaries and drama, using eXtended Markup Language (XML), and have been applied in many projects for various languages.[3]

As an example, consider the following entry for *'kakek'* from *Kamus Dewan* (KD):

kakek (kakék) Id **1.** datuk; ~ *moyang* nenek moyang; **2. = kakek-kakek** a) orang lelaki yg tersangat tua: *kelihatan seorang ~ datang tergopoh-gapah;* b) sudah tua benar (bkn orang lelaki): *suaminya sudah ~.*

Figure 1: KD entry for *'kakek'*, as formatted in the paper version

[1] Although punctuation markers placed before the italicised text may help in distinguishing them, occasional ambiguities still occur.

[2] `http://www.tntsb.com`

[3] `http://www.tei-c.org/Applications/`

This entry can be annotated as follows, using the TEI schema for print dictionaries:[4]

```
<entry>
 <form>
  <orth>kakek</orth>
  <pron>kakék</pron>
  <etym>Id</etym>
 </form>
 <sense n="1">
  <def>datuk</def>
   <re>
    <form><orth><oRef/> moyang</orth></form>
    <sense><def>nenek moyang</def></sense>
   </re>
 </sense>
 <sense n="2">
  <form>
   <lbl>=</lbl>
   <orth>kakek–kakek</orth>
  </form>
  <sense n="a">
   <def>orang lelaki yg tersangat tua</def>
   <eg>kelihatan seorang <oRef/> datang tergopoh–gapah</eg>
  </sense>
  <sense n="b">
   <def>sudah tua benar (bkn orang lelaki)</def>
   <eg>suaminya sudah <oRef/></eg>
  </sense>
 </sense>
</entry>
```

Figure 2: KD entry for *'kakek'* annotated with TEI

Thus clearly delineating each field in the entry, and what they represent. We reproduce here the TEI guidelines' descriptions of the tags used in Figure 2:

- <entry> contains a reasonably well-structured dictionary entry.
- <form> groups all the information on the written and spoken forms of one headword.
- <orth> gives the orthographic form of a dictionary headword.
- <pron> contains the pronunciation(s) of the word.
- <etym> encloses the etymological information in a dictionary entry.
- <sense> groups together all information relating to one word sense in a dictionary entry.
- <def> contains definition text in a dictionary entry.
- <eg> contains an example text containing at least one occurrence of the word form, used in the sense being described.
- <re> contains a dictionary entry for a lexical item related to the headword, such as a compound phrase or derived form, embedded inside a larger entry.

[4] http://www.tei-c.org/P4X/DI.html

- <lbl> contains a label for a form, example, translation, or other piece of information, e.g. abbreviation for, contraction of, literally, approximately, synonyms, etc.
- <oRef> indicates a reference to the orthographic form(s) of the headword.

The nested nature of the tags also indicate the applicable scope for each piece of information. In addition, Latin scientific names (or other proper nouns if required) of organisms can also be explicitly tagged as <term> (or other appropriate tags), as in this example for *'kacapiring'*:

kacapiring sj tumbuhan (pokok dan bunganya), bunga cina, bunga susu, bunga susun kelapa, *Gardenia augusta*.

```
<entry>
  <form><orth>kacapiring</orth></form>
  <sense>
    <def>sj tumbuhan (pokok dan bunganya), bunga cina, bunga susu, bunga susun kelapa,
        <term lang="lat">Gardenia augusta</term></def>
  </sense>
</entry>
```

Figure 3: KD entry for *'kacapiring'*, with its scientific name explicitly annotated.

See the TEI guidelines for print dictionaries for the full list of tags and their purposes.

A KD Parser tool was programmed to semi-automatically[5] parse sample KD entries (formatted as Microsoft Word documents) and to annotate their logical structure in XML. Although the current XML tag set used does not conform to the TEI starndards, various computer programs are available to help transform the KD Parser output to TEI-compliant files. We also exported the annotated KD data as a "flat" list of senses to streamline the look-up operations described in the next subsection. For example, 4 orthogonal-form–sense records are obtained from the *'kakek'* entry above:

Headword	Orth. form(s)	Gloss
kakek	kakek	datuk
kakek	kakek moyang	nenek moyang
kakek	kakek, kakek-kakek	orang lelaki yg tersangat tua
kakek	kakek, kakek-kakek	sudah tua benar (bkn orang lelaki)

Table 1: List of sense records from KD entry headed by *'kakek'*. Only the headword, orthographic form(s) and gloss applicable to each sense are shown for brevity.

2.2 Searching for Sense Records using TEI-annotated Fields

Once the fields in each entry are annotated explicitly, the KD contents can now be queried in a more targeted manner.

For example, to look up the definition for *'mengandungi'* in the unannotated electronic version of KD, a user would either have to a.) look up the entry headed by *'kandung'* and read through

[5] i.e. some human checking and validation is required to pre-process the files and to correct annotation errors.

the entire entry paragraph until she finds the information for *'mengandungi'*, or b.) search for *all* headword entries where the text contain the word *'mengandungi'* – which will also return irrelevant records such as ***krom*** *... bahan pewarna yg mengandungi kromium... '* – and browse through all records until the relevant one is found.

In contrast, the annotated KD will allow queries that return sense records where *'mengandungi'* is one of the applicable orthographic forms i.e.

Headword	Orth. form(s)	Gloss
kandung	mengandung, mengandungi	berisi, memuat

Similarly, a search for the compound phrase *'kapur hidup'* (or any one of its equivalent orthogonal forms) will immediately yield

Headword	Orth. form(s)	Gloss
kapur	kapur hidup, kapur kuripan, kapur mentah, kapur tohor	kapur yg belum dicampur dgn air, kalsium oksida

In the same way, lexicographers and linguistics researchers can quickly derive "sub-dictionaries" from the annotated KD for more detailed study, e.g. by searching for records having a specific etymology source or subject field, are marked as idioms (*peribahasa*), or even a list of Malay common names for plants and animals to be matched against their names in other languages on the basis of their scientific names. For instance, a computer program can be written to discover that *'kacapiring'* (in the earlier example) is known as 'gardenia' or 'cape jasmine' in English[6] by searching the Internet with the keyword 'Gardenia augusta', which had been explicitly marked in the *'kacapiring'* entry. Such multilingual terminology lists will be helpful in sharing and exchanging literature and research resources on biodiversity.

2.3 Alternative Storage Format

The TEI format can also be considered as an alternative storage format for KD data. This is because the structure of the explicitly-annotated KD data can be managed systematically using database systems. As TEI is a special type of XML, exports to various output formats, including HTML, formatted plain text, word processor documents and PDF can be done flexibly using configurable computer tools (see also `http://www.tei-c.org/Software/` for TEI-specific tools).

3 Malay WordNet

Researchers from the Computational Linguistics and Natural Language Processing fields have proposed and developed alternative lexical resources that are richer in semantic content, to support and drive further research in those fields, as well as development of various computer applications that are required to process natural language texts. As most of these lexical resources are available only for a few languages (most notably English), constructions of similar resources for

[6] `http://www.floridata.com/ref/g/gard_aug.cfm`

the Malay language will provide some interesting tools and models for researchers in this region to work with. We introduce the WordNet lexical database as one such example.

3.1 Princeton's English WordNet

WordNet [4, 11] is a lexical database system for English, designed based on psycholinguistic principles. It organises word senses on a semantic basis, rather than by their orthographic forms. This is done by grouping nouns, verbs, adjectives and adverbs into sets of synonyms, and then defining various relations between the synonym sets (synsets). WordNet is one of the best-known and most popular online lexical resource for research due to its wide coverage and free availability.

As an example, a search for 'plant' would give the following 4 noun synsets:[7]

1. **(plant#n#1, works#n#1, industrial plant#n#1)** – buildings for carrying on industrial labor; "they built a large plant to manufacture automobiles"
2. **(plant#n#2, flora#n#2, plant life#n#1)** – a living organism lacking the power of locomotion
3. **(plant#n#3)** – something planted secretly for discovery by another; "the police used a plant to trick the thieves"; "he claimed that the evidence against him was a plant"
4. **(plant#n#4)** – an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience

where the first synset consists of three members: 'plant#n#1', 'works#n#1' and 'industrial plant#n#1' which are synonyms of each other, and has the gloss 'buildings for carrying on industrial labor', as well as the example usage 'they built a large plant to manufacture automobiles'.

Synsets are connected to each other by various relations defined in WordNet. Here are a few examples:

- hypernymy (is-a): **(refinery#n#1)** *is-a* **(plant#n#1, works#n#1, industrial plant#n#1)**
- meronymy (part-of): **(sleeve#n#1, arm#n#6)** *part-of* **(garment#n#1)**
- entailment: **(buy#v#1, purchase#v#1)** *entails* **(pay#v#1)**
- cause: **(pain#v#2, anguish#v#2, pain#v#3)** *causes* **suffer#v#3**

In other words, WordNet is similar to a thesaurus when used by human readers, with explicit types of relations connecting particular word senses. Thiss offers users another way to explore word meanings by navigating the lexical network. In particular, synonymy and hypernymy (is-a) are the two most important relations in WordNet, where members in a synset can substitute for each other in a context, and the is-a relation hierarchy provides a kind of linguistic ontology for English.

[7] We use the notation *word#p#i* to mean the *i*-th sense of *word* having the part-of-speech *p*, e.g. 'plant#n#1' indicates the first sense of the noun 'plant'. The 6 verb synsets containing 'plant' are not shown here.

WordNet has also been used in a large number of research involving linguistics, cognitive science, artificial intelligence and other fields.[8]

3.2 A Malay WordNet Prototype

There has been much effort and interest in building wordnet systems in languages other than English, and the Global WordNet Association (GWA)[9], a non-commercial organisation, provides a platform for discussing, sharing and connecting these wordnets of different languages. Wordnets for various languages are currently listed as available on the GWA website, with many of them aligned to each other (including English, Arabic, French, Dutch, German, Spanish, Italian, Czech, Greek, Estonian, Greek, Romanian, etc).

As there was no wordnet system available for the Malay language, we attempted to build a prototype Malay WordNet using exising bilingual dictionary data (see [3] for a more detailed description):

1. A list of sense records were first produced from the *Kamus Inggeris-Melayu Dewan* (KIMD) [2], an English–Malay bilingual dictionary.
2. A subset of these sense records were manually aligned with the closest matching English WordNet synset, by our linguists and translators at our research group.
3. Malay synsets were then created based on the KIMD–English WordNet alignments.
4. Wherever possible, relations between the English synsets were copied over to the Malay synsets.

Here are sample synsets containing the noun *'nota'* from the Malay WordNet prototype, with English WordNet synset glosses retained:

1. **(nota#n#1, catatan#n#2)** – a brief written record
2. **(nota#n#2, anotasi#n#1)** – a comment or instruction (usually added)
3. **(peringatan#n#4, nota#n#3, surat#n#3, sebaris dua#n#1)** – a short personal letter

As well as some synset relations:

- hypernymy: **(leksikon#n#1, kamus#n#1)** *is-a* **(rujukan#n#5)**
- meronymy: **(roti#n#1)** *part-of* **(sandwic#n#1)**
- entailment: **(mendengkur#v#1, mengeruh#v#1)** *entails* **(tidur#v#2)**
- cause: **(mengajar#v#4)** *causes* **(belajar#v#2, mengaji#v#3)**

[8] See `http://lit.csci.unt.edu/~wordnet/` for a comprehensive list of papers, and `http://wordnet.princeton.edu/links` for a list of computer tools built around WordNet.

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

By applying WordNet::Similarity [5] – a suite of computer programs that measures the similarity score between any two word senses based on their locations in the English WordNet – but using our Malay WordNet prototype instead, we were able to produce a prototype Malay sense-tagger which correctly selects the sense of *'gajah'* as 'chess piece' (as opposed to the 'animal' sense) for the occurrence in the Malay sentence *'Dia mengalihkan buah gajahnya dari papan catur.'* Likewise, many other computer tools using WordNet – which were previously only applicable to English texts – will also be available for Malay once the complete Malay WordNet is in place.

3.3 Alternative Approach to Constructing Malay WordNet

The Malay WordNet prototype was meant to be that: a prototype for demonstration and explorative purposes. Several shortcomings of the prototype and its semi-automatic construction approach were identified:

- KIMD is a uni-directional, English-to-Malay dictionary. As such, many of the given Malay translation equivalents are not valid Malay collocations i.e. not lexicalising a concept in Malay. Many such Malay phrases, such as *'menyebabkan terkorban'*, were erroneously included in the Malay WordNet prototype.
- The KIMD–English WordNet manual alignment was actually undertaken for other purposes and not for establishing a strict English–Malay equivalence list. Therefore, the Malay synsets derived were just approximations, and contain many "false" members.
- In many cases, the sense distinctions and synset structures have an English bias: an English word might be perceived to have two separate senses due to usage scenarios or grammatical construction, but both senses might be perceived to correspond to one single Malay sense.
- As the Malay WordNet prototype is essentially translated from the English WordNet, many lexical gaps are unaccounted for. Culture- or language-specific concepts and words, like *'pantun'*, *'songkok'*, *'mencincang'*, *'kebaya'*, are not included.

As such, we propose to follow this alternative construction methodology [10] for future versions of Malay WordNet:

1. Develop a core wordnet of about 5000 synsets for Malay manually. The EuroWordNet [9] and BalkaNet [8] projects have identified 1024 and 5000 Common Base Concepts respectively. These base concepts were chosen on the basis of occupying high positions in the English WordNet hypernymy hierarchy, and having many relations to other synsets.
    - Translate the 5000 base concepts[10] (defined as English WordNet synsets together with relations between them) into Malay.
    - Add Local Base Concepts, which are specific to the Malay language and/or culture.
    - Add other necessary hypernymy and horizontal relations.
2. Validate core wordnet and ensure most frequent words are included.

[10] `http://www.globalwordnet.org/gwa/gwa_base_concepts.htm`

3. Extend the core wordnet downwards (semi)-automatically:
    - Use automatic techniques for more specific concepts (e.g. types of colours, food, etc).
    - Add specific domains, derivational words, 'easy' translations etc.
    - Add equivalence relations to WordNet.
4. Validate entire wordnet.

Such an approach is more time consuming and labour intensive than the one describe in the previous section. Nevertheless, wordnet systems thus produced will have the advantage of maintaining language- and culture-specific patterns and structures, while still having a core (the 5000 base concepts) that is semantically compatible and comparible to wordnets of other languages, using English WordNet as an interlingual index). [10]

4 Conclusion

We have described how current computer technology can help in enhancing existing Malay dictionaries so that more advanced, targeted and meaningful search operations for Malay lexical knowledge can be facilitated. One possibility is by explicitly annotating the logical structure and fields of KD entries with TEI, an international standard for literary text annotation. Malay lexical resources that are richer in semantic content can also be constructed, e.g. a Malay WordNet. As both the TEI standard and the wordnet model are now widely used among researchers working in different fields, nations and languages, the TEI-annotated KD and the Malay WordNet can serve as mediums for exchanging and sharing various resources with other communities and languages, as well as supporting computer tools and human researchers in translating those resources to (and from) Malay.

We also note that the preparation of the TEI-annotated KD and Malay WordNet will still require lexicographic and linguistic expertise, as well as comprehensive data input sources, to ensure good quality and coverage. The role of computer technologies is to help alleviate the tediousness of such data preparation work, and to improve the efficiency by perhaps first (semi-)automatically producing a draft version of data for human experts to improve upon. As such, we look forward to collaborations with the linguistics and lexicography community to build more computerised lexical resources for the Malay language.

Acknowledgements

The list of KIMD sense records was produced by Dr Guo Cheng-Ming, while the manual KIMD–WordNet alignment work was undertaken by all UTMK linguists and translators as part of an IRPA RMK-e8 project (ref. 305/PKOMP-612704). Nur Hussein was responsible for much of the programming work while compiling the Malay WordNet prototype. The development of the KD Parser and the Malay WordNet prototype was funded by MIMOS Sdn. Bhd. We are

grateful to Dewan Bahasa dan Pustaka for the use of sample *Kamus Dewan* and *Kamus Inggeris-Melayu Dewan* data for research, as well as providing information on the internal structure of these dictionaries.

References

[1] *Kamus Dewan*. Dewan Bahasa dan Pustaka, Kuala Lumpur, Malaysia, 2004.

[2] *Kamus Inggeris Melayu Dewan*. Dewan Bahasa dan Pustaka, Kuala Lumpur, Malaysia, 2000.

[3] Lian Tze Lim and Nur Hussein. 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*, pages 13–16, Bangkok, Thailand, October 2006. Best Paper Award.

[4] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography (special issue)*, 3(4):235–312, 1990.

[5] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. WordNet::Similarity – measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, San Jose, CA, July 2004.

[6] C. M. Sperberg-McQueen and L. Burnard, editors. *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium, 2002.

[7] Text Encoding Initiative Consortium. The Text Encoding Initiative, 2007. URL `http://www.tei-c.org`.

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

[9] Piek Vossen. EuroWordNet: A multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index. *Special Issue on Multilingual Databases, International Journal of Linguistics*, 17(2), 2004.

[10] Piek Vossen. Building wordnets. PowerPoint presentation, 2006. URL `http://www.globalwordnet.org/gwa/BuildingWordnets.ppt`.

[11] WordNet. WordNet: a lexical database for the English language, 2007. URL `http://wordnet.princeton.edu/`.