

Digitising a Machine-Tractable version of Kamus Dewan with TEI-P5

Lian Tze LIM¹, Ruoh Tau CHIEW², Enya Kong TANG¹, RUSLI Abdul Ghani³, and NAIMAH Yusof³

¹(not affiliated)

²The Name Technology, Cyberjaya, Malaysia

³Dewan Bahasa dan Pustaka, Kuala Lumpur, Malaysia

ABSTRACT

Kamus Dewan is the authoritative dictionary for Bahasa Malaysia, containing a wealth of linguistic and cultural information about Bahasa Malaysia. It is currently available in print, as well as an searchable online dictionary. However, the online dictionary lacks advanced search capabilities that target specific fields within each headword and lemma entry. For these information to be targeted and extracted efficiently by computers, the macro- and micro-structures of Kamus Dewan entries need to be first annotated or marked up explicitly. We describe how TEI-P5 guidelines have been applied in this endeavour to make the Kamus Dewan more machine-tractable. We also give some examples of how the machine-tractable data from Kamus Dewan can be used for linguistic research and analysis, as well as for producing other language resources.

Keywords: Machine-tractable dictionaries, Language resources, Bahasa Malaysia, TEI

1 INTRODUCTION

Kamus Dewan (Hajah Noresah, 2004) is the authoritative dictionary for Bahasa Malaysia, containing a wealth of linguistic and cultural information about Bahasa Malaysia and the Malay Archipelago. The information fields in the entries' micro-structures include morphological variations, etymology, domain, register, regional usage; multiword expressions including phrases, idioms and proverbial sayings (*peribahasa*); glosses and examples.

Most electronic dictionaries, including the digital version of Kamus Dewan, allow searches by headwords. 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.

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, based on the Text Encoding Initiative (TEI) guidelines. By using specific lookups based on the extracted fields, specialised dictionaries on specific domains can be extracted as well.

This project will annotate the macro- and micro-structures of Kamus Dewan dictionary entries using TEI XML. The annotated fields will then be extracted into a MySQL database to facilitate more specific and targeted word lookups and analysis.

2 DIGITAL READINESS OF LEXICAL RESOURCES FOR NLP

Lexical resources provide fundamental information about lemmas and their senses of a language, to enable natural language processing (NLP) and computational linguistic (CL) analysis on text and utterances of the language. To aid the discussion, we use the following (much-simplified) categorisation of lexical resources in Figure 1 in terms of their digital readiness for NLP work (Figure 1).

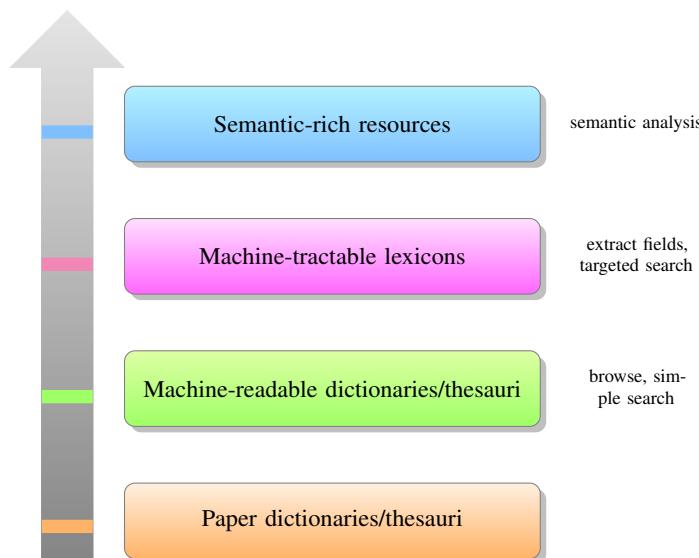


Figure 1. Types of lexical resources, based on digital readiness

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 2. Example entry from the printed Kamus Dewan

2.1 Paper dictionaries/thesauri

Paper dictionaries or thesauri are traditional dictionaries and thesauri printed on paper, for human consumption only. Text formatting effects such as bolds and italics, as well as punctuations, provide visual cues to help readers discern the various microstructure fields in an entry (see Figure 2). All derivations and meaning entries are organised by headwords, which must be looked up by some sorting order (e.g. alphabetical order for Latin-based scripts; radical- and/or stroke order for languages with ideograms like Chinese and Japanese, etc.)

For example, to look up '*mengandungi*', a reader must first look up the *kata akar* (root word) '*kandung*', and then scan through the entry's paragraphs to find the relevant sub-entry '*mengandungi*'. This may present some difficulties if the reader is unfamiliar with Bahasa Malaysia's morphological rules.

2.2 Machine-readable dictionaries/thesauri

Machine-readable dictionaries (MRDs) or thesauri are digitised versions of the original paper-printed versions, and are the most common form of electronic dictionaries. This opens up the possibility of easier search: users can now access the headword '*kandung*' directly via search box. The contents of Kamus Dewan can be accessed online as a searchable MRD (<http://prpm.dbp.gov.my/>), as well as the Kamus Pro application by The Name Technology (<http://www.tntsbt.com/>).

Most MRDs retain the text formatting styles and punctuations from the original printed versions to serve as visual cues for differentiating the various fields in an entry, without actually differentiating the fields. Therefore, it would not be possible to easily identify whether an italicised text segment is a *peribahasa* or multi-word expression (MWE), an example usage of the lemma being described, or an utterance in a foreign language, without looking at other contextual visual cues (e.g. surrounding punctuations). This means mere text formatings in MRDs are insufficient to support advanced targeted look-ups, and unable to facilitate extraction of information to support NLP applications.

2.3 Machine-Tractable lexicons

Machine-tractable dictionaries or lexicons can be summarised as MRDs with machine-tractable structures, i.e. all fields and hierarchy of the entries are specifically marked and delineated, such that different information can be identified and extracted. For example, search terms can be scoped to information fields in the micro-structure such as spelling variations, derivations and phrases (or other types of MWEs); labels

based on usage, domain and register; syntactic information (e.g. parts-of-speech); senses (i.e. different meanings of the lemmas), glosses, translation equivalents, example sentences, etc. The hierarchical relations between headwords, homonyms, derivations and phrases can also be retained and made explicit. In particular, each sense of a lexical entry must be clearly delineated.

This is the level of digital-readiness to which we wish to bring the Kamus Dewan in this paper. No extra information is added to the content of the original printed dictionary — we only seek to make the macro- and micro-structures of the dictionary entries accessible by computers, using a standardised, unambiguous markup.

2.4 Semantic-Rich Resources

Machine-tractable lexicons can be further enriched with semantic information for each sense entry. This would be very useful for NLP tasks, such as text categorisation, sentiment analysis and information extraction. Some possible directions include sentiment polarity and scores (Baccianella et al., 2010; Chen and Skiena, 2014), semantic relations and networks (Miller et al., 1990; Bond et al., 2014), or some vectorial representation of the senses (Magnini et al., 2002; Patwardhan and Pedersen, 2006; Hirao et al., 2015).

Semantic-rich resources are outside the scope of this paper, although the machine-tractable version of the Kamus Dewan would be a good foundation (or at least, very beneficial) to the creation of such resources for Bahasa Malaysia.

3 STANDARDS FOR MODELLING AND MARKING UP DICTIONARIES

The Text Encoding Initiative (TEI; TEI Consortium, 2015) is a set of guidelines for electronic text encoding and interchange, by marking up (or creating) natural language texts with XML. TEI is maintained by the TEI Consortium, which aims to develop and maintain guidelines for the digital encoding of literary and linguistic texts. It is highly flexible: TEI covers a range of different texts, including prose, verses, books, dictionaries and performance texts. Annotators are not compelled to use all the proposed XML tags, or limited to the suggested set. The TEI schema can be trimmed or added to as needed, with the guidelines serving as a reference about the purpose of each XML tag (see Tutin and Véronis 1998; Erjavec et al. 2003; Dimitrova et al. 2002; Schneiker et al. 2009; Budin et al. 2012 for some case studies).

Other standards exists for marking up and annotating dictionaries and lexicons. For example, the Lexical Markup Framework (LMF; Francopoulo et al., 2009) was especially proposed for modelling MRDs and lexicons for use with NLP applications. While both standards include guidelines for modelling MRDs and lexicons, TEI is more popular in the social sciences, while LMF is more popular among computer scientists and NLP researchers, perhaps due to their different purposes.

LMF models the structure of MRDs for the express use of NLP applications, while TEI seeks to annotate existing texts. LMF therefore initially appeared to be the natural choice for marking up Kamus Dewan. However, we soon discovered that the structure of Kamus Dewan was better handled by TEI, as LMF has no mechanisms for modelling sub-entries, which abounds in Kamus Dewan as derived forms and phrasal constructions of root words. It would still be possible to convert the senses of Kamus Dewan entries to be LMF-compliant, but this requires changes to the entry-subentry hierarchy. In contrast, TEI simply adds extra annotation mark-ups to the original text and preserves the original structure. We therefore chose to use TEI in this project.

4 USING TEI-P5 TO ANNOTATE KAMUS DEWAN ENTRIES

In this example, we will describe our experiences using the TEI-P5 guidelines for dictionaries to annotate the macro- and micro-structures of Kamus Dewan entries with examples.

4.1 TEI Default Text Structure

The overall structure of a TEI document is as follows:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <!-- .... -->
```

```
</teiHeader>
<text>
  <front>
    <!-- front matter of copy text, if any, goes here -->
  </front>
  <body>
    <!-- body of copy text goes here -->
  </body>
  <back>
    <!-- back matter of copy text, if any, goes here -->
  </back>
</text>
</TEI>
```

The TEI header `<teiHeader>` contains meta-information about the text, such as editors, revisions, etc. The `<text>` may contain `<front>` and `<back>` matters, but our main focus is the Kamus Dewan dictionary entries, which will go into the `<body>`.

The `<body>` element can have divisions `<div>` of any type. We separate each alphabet part of Kamus Dewan as its own division:

```
<body>
  <div type="part" n="a">
    <!-- all entries of root words starting with 'A' -->
  </div>
  <div type="part" n="b">
    <!-- all entries of root words starting with 'B' -->
  </div>
  ...
</body>
```

4.2 A Simple Entry

Our first example is a simple entry of the root word ‘apeks’ with a single sense, with two example usages:

apeks (apéks) bahagian puncak atau hujung sesuatu yg tirus: ~ *paru-paru*; ~ *daun*.

Here is the same entry with each information field annotated with TEI-P5 tags:

```
<entry xml:id="kd_entry.1413">
  <form>
    <orth>apeks</orth>
    <pron>apéks</pron>
  </form>
  <sense xml:id="kd_sense.3611" n="1">
    <def>bahagian puncak atau hujung sesuatu yg tirus</def>
    <cit type="example">
      <q xml:id="kd_example.1192">apeks paru-paru</q>
      <q xml:id="kd_example.1193">apeks daun</q>
    </cit>
  </sense>
</entry>
```

The tags and attributes used are explained below (extracted from the TEI-P5 guidelines at <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/>):

- <entry> single structured entry.
- <xml:id> a unique identifier within the entire dictionary.
- n traditional identifier of the relevant structural units, or to record the numbering of sections or list items in the copy text
- <form> groups all the information on the written and spoken forms of one headword.
- <orth> orthographic form of a dictionary headword.
- <pron> contains the pronunciation(s) of the word.
- <sense> groups together all information relating to one word sense in a dictionary entry, for example definitions, examples, and translation equivalents.
- <def> contains definition text in a dictionary entry.
- <cit> contains a quotation from some other document. In a dictionary it may contain an example text.
- <q> the example text itself should be enclosed in a <q> or <quote> element.

4.3 Homonyms, Spelling Variants and Foreign Words

Homonyms, such as those for ‘badam’, are modelled using the type= “hom” and n attributes of <entry>. The spelling variants are annotated as <form type= “variant”>.

badam I = buah ~ sj tumbuhan (buahnya berbentuk bujur), *Prunus spp*.

badam II = bunga ~ merah-merah pd kulit (tanda penyakit kusta).

```
<entry xml:id="kd_entry.1982" type="hom" n="I">
  <form>
    <orth>badam</orth>
  </form>
  <sense xml:id="kd_sense.5118" n="1">
    <form type="variant">
      <orth>buah badam</orth>
    </form>
    <def>sj tumbuhan (buahnya berbentuk bujur), _Prunus spp_</def>
  </sense>
</entry>
<entry xml:id="kd_entry.1983" type="hom" n="II">
  <form>
    <orth>badam</orth>
  </form>
  <sense xml:id="kd_sense.5119" n="1">
    <form type="variant">
      <orth>bunga badam</orth>
    </form>
    <def>merah-merah pd kulit (tanda penyakit kusta)</def>
  </sense>
</entry>
```

We can also mark a headword as foreign, which is italicised in the printed copy:

ala carte (Perancis) hidangan masakan yg tersenarai pd menu, yg boleh dipilih secara berasingan mengikut kesukaan pelanggan pd harga yg telah ditetapkan.

```
<entry xml:id="kd_entry.508" type="foreign">
<form>
<orth>ala carte</orth>
</form>
<sense xml:id="kd_sense.1389" n="1">
<usg>Perancis</usg>
<def>hidangan masakan yg tersenarai pd menu, yg boleh dipilih secara berasingan mengikut kesukaan pelanggan pd harga yg telah ditetapkan</def>
</sense>
</entry>
```

4.4 Multiple Senses and Sub-senses

Given an entry with multiple senses, they can be discerned distinctly as the n-th **<sense>**:

abadi Ar 1. ada permulaan yg tiada pengakhiran (bkn masa, kehidupan, kenangan dsb): *kehidupan akhirat adalah kehidupan yg ~; 2. wujud atau berterusan utk selama-lamanya (sepanjang hayat dsb), tidak berkesudahan, kekal: keamanan yg ~; kasih sayang yg ~;*

```
<entry xml:id="kd_entry.8">
<form>
<orth>abadi</orth>
</form>
<sense xml:id="kd_sense.16" n="1">
<usg>Ar</usg>
<def>ada permulaan yg tiada pengakhiran (bkn masa, kehidupan, kenangan dsb)</def>
<cit type="example">
<q xml:id="kd_example.5">kehidupan akhirat adalah kehidupan yg abadi</q>
</cit>
</sense>
<sense xml:id="kd_sense.17" n="2">
<usg>Ar</usg>
<def>wujud atau berterusan utk selama-lamanya (sepanjang hayat dsb), tidak berkesudahan, kekal</def>
<cit type="example">
<q xml:id="kd_example.6">keamanan yg abadi</q>
<q xml:id="kd_example.7">kasih sayang yg abadi</q>
</cit>
</sense>
```

4.5 Usage Labels

<usg> is used to mark up various usage labels, including for etymology (e.g. ‘Ar’ for Arabic), domain (e.g. ‘Eko’ for economy); genre (e.g. ‘sl’ for *sastera lama* old literature):

adi I (Sanskrit) sl yg pertama, yg terutama, yg tertinggi: *pahlawan ~; pendekar ~.*

```
<entry xml:id="kd_entry.134" type="hom" n="I">
<form>
```

```

<orth>adi</orth>
</form>
<sense xml:id="kd_sense.378" n="1">
  <usg>Sanskrit</usg>
  <usg>sl</usg>
  <def>yg pertama, yg terutama, yg tertinggi</def>
  ...
</sense>
</entry>

antaboga † sj naga besar, hantu bumi.

<entry xml:id="kd_entry.1263">
  <form>
    <orth>antaboga</orth>
  </form>
  <sense xml:id="kd_sense.3328" n="1">
    <usg type="temporal">ark</usg>
    <def>sj naga besar, hantu bumi</def>
  </sense>
</entry>

```

At present, we do not differentiate between other different types of usages or labels (geographical, stylistic, domain and others), apart from the temporal *arcane* (†) label.

4.6 Cross-References

TEI-P5 can also model cross-references with the **<xr>** tag:

astana → **istana**.

```

<entry xml:id="kd_entry.1703">
  <form>
    <orth>astana</orth>
  </form>
  <xr>
    <ref>istana</ref>
  </xr>
</entry>

```

4.7 Subentries: Derived Forms and Phrasal Constructions

As derived forms and phrasal constructions from Bahasa Malaysia headwords (*kata akar*) have quite distinct, derivational meanings from the headwords, they are regarded as lemmas in their own right. They are therefore best modelled as sub-entries of the headword, using the **<re>** (related entry) tag. A **type** attribute can be included to indicate whether it is a derived or a phrase entry.

ala III Ar tinggi;

terala sl termulia, tertinggi: *barang lakunya ~ drpd raja-raja yg lain*.

```

<entry xml:id="kd_entry.504" type="hom" n="III">
  <form>
    <orth>ala</orth>
  </form>
  <sense xml:id="kd_sense.1384" n="1">
    <usg>Ar</usg>
    <def>tinggi</def>

```

```

</sense>
<re>
  <form type="derived">
    <orth>terala</orth>
    <usg>sl</usg>
  </form>
  <sense xml:id="kd_sense.1385" n="1">
    <def>termulia, tertinggi</def>
    <cit type="example">
      <q xml:id="kd_example.543">barang lakunya terala drpd
        raja-raja yg lain</q>
    </cit>
  </sense>
</re>
</entry>

```

badar IV; ~ sila = raja ~ sl sj kain putih yg halus.

```

<entry xml:id="kd_entry.1991" type="hom" n="IV">
  <form>
    <orth>badar</orth>
  </form>
  <re>
    <form type="phrase">
      <orth>badar sila</orth>
    </form>
    <sense xml:id="kd_sense.5153" n="1">
      <form type="variant">
        <orth>raja badar</orth>
      </form>
      <usg>sl</usg>
      <def>sj kain putih yg halus</def>
    </sense>
  </re>
</entry>

```

TEI-P5 allows nested `<re>`, which makes it ideal to model phrasal constructions of derived forms. In the example below, ‘*basahan*’ is a derived subentry of ‘*basah*’, while ‘*sahaja basahan*’ is a phrase subentry (which also happens to be a *peribahasa*) of ‘*basahan*’.

basah ...

basahan ... 3. sesuatu yg telah menjadi perkara biasa: *minuman keras sudah menjadi ~ kpd setengah-setengah orang; sahaja ~ prb* sudah menjadi kebiasaan berbuat sesuatu perbuatan yg tidak baik;

```

<re>
  <form type="derived">
    <orth>basahan</orth>
  </form>
  ...
  <sense xml:id="kd_sense.6957" n="3">
    <def>sesuatu yg telah menjadi perkara biasa</def>
  ...
<re>

```

```
<form type="phrase">
  <orth>sahaja basahan</orth>
</form>
<sense xml:id="kd_sense.6958" n="1">
  <usg>prb</usg>
  <def>sudah menjadi kebiasaan berbuat sesuatu perbuatan yg
    tidak baik</def>
</sense>
</re>
</sense>
</re>
```

5 APPLICATIONS

Source files of Kamus Dewan entries, formatted as HTML web pages, were marked up with TEI-compliant XML as described above, using a custom parser written in the Java programming language. To facilitate easier manipulation of the data, all TEI-annotated Kamus Dewan entries, lemmas and senses were also exported to a MySQL database. The database currently contains:

- 28 829 distinct root words (*kata akar*);
- 75 825 distinct orthographic forms (including derived forms, phrases), where 25 521 are multi-word expressions;
- 87 913 definitions;
- 30 604 examples.

The following subsections will provide some example applications now made possible by this machine-tractable version of Kamus Dewan.

5.1 Targeted Lookups

A few search procedures were implemented in the MySQL database to help facilitate advanced searches and lookups. For example, executing the procedure

```
CALL SEARCH_HEADWORD('tanak');
```

returns all definition entries listed under the headword ‘*tanak*’, including derived forms and phrases (Table 1).

Conversely, a user can also search for all definitions for ‘*mereka*’ — which may originate from different headwords (results in Table 2) — using the procedure call

```
CALL SEARCH_ORTHFORM('mereka');
```

The task of looking up phrases and MWEs is also made simpler, as a user would no longer need to find out which headword to look up first (Table 3):

```
CALL SEARCH_ORTHFORM('hilang kabut teduh hujan');
```

Etymologists and linguists can also search for specific labels — for example, Table 4 shows partial results from a search for lemmas originating from Jawa (*Jw*) old literature (*sl*).

5.2 Lexicography Analysis

As the definitions and examples have now been explicitly tagged, they can be regarded as a corpus, which lends itself to various analysis which may give further insights to Bahasa Malaysia lexicographic practice.

For example, we extracted the fifty most frequent words¹ (not including prepositions, conjunctions, infinitives, etc.) used in definitions (Table 5). We can also be more specific in our purpose and look specifically for ‘genus’ terms, by searching for the patterns ‘*sj...’* (‘a kind of ...’) and ‘... *yg*’ (‘... that which is’); which gives Table 6.

Table 1. Lookup results for all senses of lemmas under headword ‘tanak’

Orth. forms	Usage	Definition
bertanak		(sedang) memasak nasi
bertanak		yg ditanak
bagai bertanak di kuali	<i>prb</i>	bermurah-murah kpd orang lain sehingga mendatangkan kesusahan kpd diri sendiri
nasi bertanak		nasi yg ditanak (bukan dikukus)
mempertanak		menanak (nasi)
menanak		memasak nasi (dlm periuk, kawah, dll)
menanak		memasak sesuatu dgn merebusnya sahaja
ditanaknya semua berasnya	<i>prb; Mn</i>	perihal orang yg suka memperlihatkan kepandaian atau kebijaksanaannya di hadapan orang ramai
menanak kentang		merebus kentang
menanak minyak		memasak santan kelapa utk dijadikan minyak
menanakkan		menanak utk
penanak; pertanak		orang yg menanak, tukang masak
petanakan		sesuatu yg ditanak, masakan
sepenanak; sepenanak nasi		waktu yg lamanya serupa dgn lama orang menanak nasi (lebih kurang 20 minit)
jurutanak; tukang tanak		orang yg memasak, tukang masak
minyak tanak		minyak kelapa
tanak-tanakan		bermain masak-masak (bkn anak-anak)

Table 2. Search results for all senses of lemmas with orthographic form ‘mereka’

Headword	Pron.	Ortho. forms	Definition
mereka	meréka	mereka; mereka itu	kata ganti diri ketiga (utk bilangan yg banyak), orang-orang itu
reka	réka	mereka; mereka-reka	menyusun (memasang, mengatur, mengarang) baik-baik
reka	réka	mereka; mereka-reka	mencari akal (daya, upaya, ikhtiar)
reka	réka	mereka; mereka-reka	memikirkan (sesuatu), merancang, merencanakan
reka	réka	mereka; mereka-reka	membayangkan (dlm angan-angan), mencita-citakan
reka	réka	mereka; mereka-reka	menduga, mengagak-agakkan, mengira-ngirakan

Table 3. Search result for the *peribahasa* ‘hilang kabut teduh hujan’

Headword	Ortho. forms	Usage	Definition
kabus	hilang kabut teduh hujan	<i>prb</i>	mendapat kesenangan setelah menderita

Table 4. Partial search result for lemmas with Jawa (*Jw*) old literature (*sl*) origin

Headword	Ortho. forms	Usage	Definition
adipati	adipati	<i>Jw; sl</i>	raja, kepala daerah
aji	aji	<i>Jw; sl</i>	raja, ratu
aji	aji mahkota	<i>Jw; sl</i>	raja yg merdeka
aji	kakang aji	<i>Jw; sl</i>	panggilan permaisuri kpd raja
andeka	mengandeka	<i>Jw; sl</i>	bertitah
angur	angur	<i>Jw; sl</i>	lebih baik ... (drpd), biarlah, remaklah
limpung	limpung	<i>Jw; sl</i>	senjata yg tajam
lir	lir	<i>Jw; sl</i>	seperti (umpama)
lir	sang lir sari	<i>Jw; sl</i>	yg spt bunga (gadis yg elok)
pakanira	pakanira	<i>Jw; sl</i>	tuan, engkau

Table 5. Fifty most frequent words in *Kamus Dewan* definitions

Word	Freq.	Word	Freq.	Word	Freq.	Word	Freq.
sesuatu	7782	barang	1437	sangat	1044	boleh	876
tidak	7188	mempunyai	1412	lebih	1037	baik	873
orang	6923	alat	1345	dapat	1024	tanah	864
tumbuhan	3869	hati	1340	sudah	1022	mata	850
pokok	3218	seseorang	1304	dibuat	1010	anak	836
tempat	2340	besar	1304	ada	994	bahan	829
air	1732	bahagian	1294	bagi	967	atas	810
kecil	1585	keadaan	1287	laut	966	kain	792
perbuatan	1578	menjadikan	1257	kayu	954	diri	786
bunyi	1524	membuat	1235	biasanya	950	melakukan	786
menjadi	1517	ikan	1165	benda	900	telah	786
kata	1466	sama	1062	burung	887		
digunakan	1440	banyak	1052	wang	885		

Table 6. Fifty most frequent ‘genus’ words in *Kamus Dewan* definitions

Word	Freq.	Word	Freq.	Word	Freq.	Word	Freq.
tumbuhan	3463	kata	83	tanah	60	angin	49
orang	2205	wang	81	barang	55	minyak	47
ikan	788	makanan	79	tali	54	pegawai	47
sesuatu	753	perahu	79	baju	53	ubat	44
burung	497	unsur	79	keadaan	53	kayu	44
alat	361	bahagian	76	bekas	52	seseorang	44
kain	173	batu	75	kapal	52	minuman	43
binatang	163	perempuan	73	bakul	52	perbuatan	43
penyakit	161	apa	69	buah	51	nasi	42
kuih	129	air	67	serangga	51	kawasan	41
bahan	123	tempat	65	siput	51	pekerjaan	41
permainan	122	anak	64	pokok	51		
benda	110	surat	61	hantu	50		

Going through the list, we see that both ‘*binatang*’ (163 occurrences) and ‘*haiwan*’ (12 occurrences) — both mean ‘animal’ — are used. While this may at first come across as an inconsistency, this finding actually reveals an interesting development of Bahasa Malaysia: ‘*binatang*’ was initially neutral, but picked up derogatory connotations in the 1980s. Thereafter, ‘*haiwan*’ was used for new entries starting with the third edition of *Kamus Dewan* in 1994. (The word ‘*binatang*’ in existing entries were retained for reasons of exactly this lexicography historical detail.)

5.3 Multilingual Botanical and Zoological Checklist

The Malay archipelago has a very rich biodiversity. It is therefore unsurprising that *Kamus Dewan* contains a huge number of names for flora and fauna. Many (if not most) definitions of these names include the scientific names in Latin. For example, here is the annotated entry for the headword ‘*adas*’, with the scientific names annotated with `<name xml:lang="la">`:

```
<entry xml:id="kd_entry.125">
  <form>
    <orth>adas</orth>
  </form>
  <sense xml:id="kd_sense.339" n="1">
    <form type="variant">
      <orth>adas landi</orth>
      <orth>adas pedas</orth>
    </form>
    <def>sj tumbuhan (herba), <name xml:lang="la">Foeniculum
      vulgare</name></def>
  </sense>
  <sense xml:id="kd_sense.340" n="2">
    <form type="variant">
      <orth>adas cina</orth>
      <orth>adas manis</orth>
    </form>
    <def>sj tumbuhan (herba), <name xml:lang="la">Anethum
      graveolens</name></def>
  </sense>
</entry>
```

Using the scientific names as a pivot, we can then align the Malay flora and fauna names to their translations in other languages, creating a multilingual checklist.²

The Catalogue of Life (Roskov et al., 2015) is an online database of the world’s known species of animals, plants, fungi and micro-organisms. The Catalogue of Life 2015 Annual Checklist (CoL2015) contains more than 1.6 million species (84 % coverage) from 154 databases, and is available for download as an MySQL database. The data contains some common names in a number of languages (most notably English), though not always available for all species. No Bahasa Malaysia common name is available in CoL2015. CoL2015 was also used to check for typographical errors in the scientific names from *Kamus Dewan*, by searching for close matches with Levenshtein distance of less than 3 single-character editing actions.

We also used CoL2015 for looking up the unique accepted name for each species from *synonyms*. Scientific nomenclature can change as study and research progresses, and may vary based on region and domain, etc. All synonyms are still recorded for reference purposes: for example, legislators may need to refer to previous conventions and cases. (Not all species were found in the CoL2015 data for download; the Col2016 release is expected to cover more botanical species with the inclusion of more source databases.)

¹We used the Python NLTK library (<http://www.nltk.org/>) and MySQL to query and process the text.

²A checklist is a list of species names.

We then used the accepted scientific names to look up their English common names from WordNet (Miller et al., 1990; Bond and Paik, 2012): some example alignments are shown in Table 7. As many definitions for flora and fauna in the *Kamus Dewan* comprise only the scientific name, the definition from WordNet lends further description about the item.

Table 7. Example aligned Bahasa Malaysia and English common names from WordNet via scientific names

B. M'sia (KD-TEI)	Scientific name	English (WN)	Definition (WN)
bayam duri	<i>Amaranthus spinosus</i>	thorny amaranth	erect annual of tropical central Asia and Africa having a pair of divergent spines at most leaf nodes
bayan lepas	<i>Psittacula krameri</i>	ring-necked parakeet	African parakeet
bebaru	<i>Hibiscus tiliaceus</i>	balibago; mahagua; mahoe; majagua; purau	shrubby tree widely distributed along tropical shores; yields a light tough wood used for canoe outriggers and a fiber used for cordage and caulk; often cultivated for ornament
bebésaran	<i>Morus alba</i>	white mulberry	Asiatic mulberry with white to pale red fruit; leaves used to feed silkworms
belatik	<i>Padda oryzivora</i>	Java finch; Java sparrow; ricebird	small finch-like Indonesian weaverbird that frequents rice fields

	A	C	D	E	F	G	H	I	J	K	L
1	Thumbnail	Scientific Name	English	B. Malaysia	B. Indonesia	Chinese	Japanese	Korean	Thai	Arabic	French
44		<i>Psittacula krameri</i>	Rose-ringed Parakeet; ring-necked parakeet	bayan lepas		红领绿鹦鹉	ワカヘンセイインコ			باراكيت أحمر؛ بنبي، منه؛ البراكيت؛ الأحمر، بنبي منه؛ Perruche collier الباراكيت	Perruche collier;
45		<i>Morus alba</i>	white mulberry	bebésaran; besaran; kertau; mulberi		白桑；白桑椹		뽕나무；오디； 뽕나무꽃			Mûrier blanc; Murier blanc
46		<i>Padda oryzivora</i>	Java Sparrow; Java finch; Java rice sparrow; Java rice bird; ricebird	Burung Ciak Jawa; belatik	Gelatik Jawa	禾雀；爪哇禾雀；爪哇雀；灰文鸟； 灰文鸟；文鸟；文鸟	ブンチョウ；手乗りブンチョウ； 手乗り文鳥；文鳥	문조		چوارا سبارو	Padda de Java; Moineau de Java; Calfat de Java; Padda oryzivore
			Belimbing besi; Pokok Belimbing Besi;				スタート	카림볼라；			

Figure 3. A multilingual checklist compiled from the TEI-annotated *Kamus Dewan*, WordNet and Wikidata

Taking this a step further, we also queried Wikidata³ for common names in different languages, Wikimedia also provides CreativeCommons-licensed images for each Wikipedia entry about the species, which can be retrieved using Wikidata queries. Figure 3 shows a sample of the resultant multilingual checklist. Such a resource would help enrich the multilingual nomenclature in botanical and zoological work in the region.

³HTTP interfaces for querying Wikidata can be found at <http://www.wikidata.org/w/api.php> and https://wdq.wmflabs.org/api_documentation.html.

6 CONCLUSION

We have described our experiences in creating a machine-tractable version of Kamus Dewan by annotating the macro- and micro-structures of its entries, using TEI-P5 XML tags and guidelines for electronic dictionaries. The annotated data allows researchers and linguists to access the rich cultural and lexicographic contents in Bahasa Malaysia in more flexible and targeted ways, opening up possibilities in discovering new insights into the language, as well as creating language technology tools for Bahasa Malaysia.

ACKNOWLEDGEMENTS

This work is partly supported by the MSC Malaysia Innovation Voucher scheme from the Malaysian Multimedia Development Corporation.

REFERENCES

- Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC 2010)*, volume 10, pages 2200–2204.
- Bond, F., Lim, L. T., Tang, E. K., and Riza, H. (2014). The combined Wordnet Bahasa. *NUSA: Linguistic studies of languages in and around Indonesia*, 57:83–100.
- Bond, F. and Paik, K. (2012). A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71, Matsue, Japan.
- Budin, G., Majewski, S., and Mört, K. (2012). Creating lexical resources in TEI P5: a schema for multi-purpose digital dictionaries. *Journal of the Text Encoding Initiative*, (3).
- Chen, Y. and Skiena, S. (2014). Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 383–389.
- Dimitrova, L., Pavlov, R., and Simov, K. (2002). The Bulgarian dictionary in multilingual lexical data bases. *Cybernetics and Information Technologies*, 2(2):33–42.
- Erjavec, T., Evans, R., Ide, N., and Kilgarriff, A. (2003). From machine readable dictionaries to lexical databases: the Concede experience. In *Proceedings of the 7th International Conference on Computational Lexicography (COMPLEX'03)*, Budapest, Hungary.
- Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., and Soria, C. (2009). Multilingual resources for NLP in the Lexical Markup Framework (LMF). *Language Resources and Evaluation*, 43(1):57–70.
- Hajah Noresah, b. B., editor (2004). *Kamus Dewan*. Dewan Bahasa dan Pustaka, Kuala Lumpur, Malaysia.
- Hirao, T., Wariishi, N., Suzuki, T., and Hirokawa, S. (2015). Vector similarity of related words in the Japanese Word Net. In *Proceedings of the 4th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 142–147. IEEE.
- Magnini, B., Strapparava, C., Pezzulo, G., and Gliozzo, A. (2002). Comparing ontology-based and corpus-based domain annotations in WordNet. In *Proceedings of the First International WordNet Conference*, pages 21–25, Mysore, India.
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography (special issue)*, 3(4):235–312.
- Patwardhan, S. and Pedersen, T. (2006). Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together*, volume 1501, pages 1–8.
- Roskov, Y., Abucay, L., Orrell, T., Nicolson, D., Kunze, T., Culham, A., Bailly, N., Kirk, P., Bourgoin, T., DeWalt, R., Decock, W., and De Wever, A., editors (2015). *Species 2000 & ITIS Catalogue of Life, 2015 Annual Checklist*. Species 2000: Naturalis, Leiden, the Netherlands.
- Schneiker, C., Seipel, D., Wegstein, W., and Prätor, K. (2009). Declarative parsing and annotation of electronic dictionaries. In *Proceedings of the 6th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2009)*, pages 122–132, Milan, Italy.
- TEI Consortium, editor (2015). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. [Last modified 2015-10-04].
- Tutin, A. and Véronis, J. (1998). Electronic dictionary encoding: Customizing the TEI guidelines. In *Proceedings of the 8th EURALEX International Congress*, pages 363–374, Liège, Belgium.