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

Lim Lian Tze    Tan Ewe Hoe    Tang Enya Kong

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

11th International Conference on Translation
Outline

1 Electronic Dictionaries

2 Logical Annotation of Kamus Dewan
   • Annotating KD with TEI
   • Searching with TEI-annotated fields

3 Malay WordNet
   • Princeton’s English WordNet
   • A Malay WordNet Prototype

4 In Closing...
Electronic Dictionaries

- Converted/scanned/OCR-ed/etc from paper dictionaries
  - ✓ Speed up word look-ups (dictionary software, WWW interface)
  - ✓ Facilitate automatic spell-checking in computer applications

- Common search options:
  - By headword (base form)
    ⇒ returns entire paragraph entry under headword
  - Full-text indexed search
    ⇒ returns all headword paragraph entries, the body of which contains the search term
What if I’d like to see... 

- *Only* specific senses for a derived word, phrasal expression, etc?
- A list of all phrases, containing words originating from Jawa?
- Text formatting: visual cues for humans to distinguish entry fields

**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 ~.

- Insufficient to support advanced look-up in electronic dictionaries (e.g. phrases, example usages and latinate names of genus/species may all be italicised)
Outline

1. Electronic Dictionaries

2. Logical Annotation of Kamus Dewan
   - Annotating KD with TEI
   - Searching with TEI-annotated fields

3. Malay WordNet
   - Princeton’s English WordNet
   - A Malay WordNet Prototype

4. In Closing...
The Text Encoding Initiative (TEI) Guidelines

http://www.tei-c.org/

(own emphasis)

"The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation."
Annotating KD with TEI

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 ~.

<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>
</entry>
<entry>
  <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>
## Extracted Sense Records from TEI-annotated KD

(Simplified view)

<table>
<thead>
<tr>
<th>Headword</th>
<th>Orth. form(s)</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>kakek</td>
<td>kakek</td>
<td>datuk</td>
</tr>
<tr>
<td>kakek</td>
<td>kakek moyang</td>
<td>nenek moyang</td>
</tr>
<tr>
<td>kakek</td>
<td>kakek, kakek-kakek</td>
<td>orang lelaki yg tersangat tua</td>
</tr>
<tr>
<td>kakek</td>
<td>kakek, kakek-kakek</td>
<td>sudah tua benar (bkn orang lelaki)</td>
</tr>
</tbody>
</table>
What does ‘mengandungi’ mean?

- **Search by headword**
  - Look up ‘kandung’
  - Scan through entire entry paragraph until you see ‘mengandungi’
  - (Non-speakers unfamiliar with morphological rules?)

- **Full-text indexed search** (with computer)
  - Search for all entry paragraph texts containing ‘mengandungi’
  - May include ‘false’ results:
    - *krom* ... bahan pewarna yg *mengandungi* kromium...

- **Search by TEI-annotated fields**
  - Search for sense records with ‘mengandungi’ as an orthographic form

<table>
<thead>
<tr>
<th>Headword</th>
<th>Orth. form(s)</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>kandung</td>
<td>mengandung, <em>mengandungi</em></td>
<td>berisi, memuat</td>
</tr>
</tbody>
</table>
Other examples

- **Phrase search:** ‘kapur tohor’

<table>
<thead>
<tr>
<th>Headword</th>
<th>Orth. form(s)</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>kapur</td>
<td>kapur hidup, kapur kuripan, kapur mentah, kapur tohor</td>
<td>kapur yg belum dicampur dgn air, kalsium oksida</td>
</tr>
</tbody>
</table>

- **Automatically mining animal/plant names in various languages:**

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

  Gardenia augusta ≡ kacapiring, gardenia, common gardenia, cape jasmine, 椀子, クチナシ, Jazmín del Cabo . . .
Outline

1. Electronic Dictionaries

2. Logical Annotation of Kamus Dewan
   - Annotating KD with TEI
   - Searching with TEI-annotated fields

3. Malay WordNet
   - Princeton’s English WordNet
   - A Malay WordNet Prototype

4. In Closing...
Princeton’s English WordNet

- Developed by Princeton University Cognitive Science Laboratory (http://wordnet.princeton.edu, free license)
- Richer semantic content of lexical entries (sense level)
- Widely used by researchers in linguistics, cognitive science, artificial intelligence, etc.
- Basic unit: Synset = “synonym set”
- Represents a lexicalised concept by synonyms, gloss and relations to other synsets
Example English WordNet Synsets and Relations

**synset = synonym set**

### English Synsets containing noun “plant”

1. *(plant #n#1, works #n#1, industrial _plant #n#1)* — buildings for carrying on industrial labor
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
4. *(plant #n#4)* — an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience

**hypernymy** *(refinery)* is-a-kind-of *(plant, works, industrial plant)*

**meronymy** *(sleeve, arm)* is-part-of *(garment)*

**cause** *(pain, anguish, hurt)* causes *(suffer)*

**entailment** *(buy, purchase)* entails *(pay)*,
*(choose, take, select, pick out)*
Wordnets in other languages

http://www.globalwordnet.org/gwa/wordnet_table.htm

EuroWordNet  Dutch, German, French, Spanish, Italian, Czech, Estonian
BalkaNet  Romanian, Bulgarian, Turkish, Slovenian, Greek, Serbian
Others  Afrikaans, Arabic, Chinese, Hindi, Tamil, Turkish...

No Malay WordNet
A Malay WordNet Prototype
(Lim and Hussein 2006)

Using existing dictionary sense-to-English WordNet alignment data (produced manually by UTMK team of linguists and translators):

**English WordNet Synset**

\[(\text{plant}^n_2, \text{flora}^n_2, \text{plant\_life}^n_1)\]—a living organism lacking the power of locomotion

**Relevant KIMD Senses**

\text{plant} n. 1. living organism with leaves and root, \textit{tumbuh-tumbuhan}.

\text{flora} n. \textit{flora}.

**Generated Malay WordNet Synset**

\[(\text{tumbuh-tumbuhan, flora})\]—a living organism lacking the power of locomotion
Malay WordNet Synset Relation Examples

- Refer to relations between English WordNet synsets
- Copy selected relations over to Malay WordNet where possible

**hypernymy** (leksikon, kamus) *is-a-kind-of* (rujukan)

**meronymy** (juri) *is-member-of* (tribunal, pengadilan)
   (roti) *is-part-of* (sandwic)
   (tepung) *is-substance-of* (roti)

**troponymy** (menconteng) *is-one-way-to* (melukis)

**cause** (mengajar) *causes* (belajar, mengaji)

**entailment** (mendengkur, mengeruh) *entails* (tidur)
Example Uses of Malay WordNet

For human users

- like a thesaurus, with specific relation names
- an interesting way of learning/exploring new words
- Aligned to wordnets of other languages ⇒ Multilingual look-up

<table>
<thead>
<tr>
<th>English</th>
<th>Catalan</th>
<th>Spanish</th>
<th>Malay</th>
</tr>
</thead>
<tbody>
<tr>
<td>dictionary, lexicon</td>
<td>diccionari, lèxic</td>
<td>diccionario, léxico</td>
<td>kamus, leksikon</td>
</tr>
<tr>
<td>a reference book containing an alphabetical list of words with information about them</td>
<td>Obra de referència on es recullen alfabèticament les paraules</td>
<td>Obra de referencia donde se recogen alfabeticamente las palabras</td>
<td>Malay</td>
</tr>
</tbody>
</table>

...
Example Uses of Malay WordNet

For computers

- processing/analysing natural language texts
- sense tagging: selecting most likely sense for a lexical item occurrence
- automatic “context-sensitive” look-up

Sense-tagging in Malay
Dia mengalihkan buah gajah\textsubscript{n.1} nya dari papan catur.

*gageh (n.)*

1. a chess piece (bishop)
2. five-toed pachyderm (elephant)

(Plenty of existing algorithms/tools for English WordNet ⇒ reuse with Malay WordNet)
Outline

1 Electronic Dictionaries

2 Logical Annotation of Kamus Dewan
   - Annotating KD with TEI
   - Searching with TEI-annotated fields

3 Malay WordNet
   - Princeton’s English WordNet
   - A Malay WordNet Prototype

4 In Closing...
Conclusion... 

- IT can help enhance existing Malay dictionaries
  - to provide more advanced, targeted and meaningful search operations
  - to facilitate data exchange with resources of other languages
  - to support translators by providing multilingual, semantic-rich lexical resources

- For accuracy and comprehensive coverage, we need
  - lexicographer expertise
  - comprehensive input data sources
  - (computer technologies may help by generating a “draft” version for human experts to work on)

Thank You