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

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)
    $\Rightarrow$ returns entire paragraph entry under headword
  - Full-text indexed search
    $\Rightarrow$ 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

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>
```

Annotating KD with TEI (cont.)

```
  <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)

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)

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

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

Other examples

- Phrase search: 'kapur tohor'

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

- 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>
```

Floridata: **Gardenia augusta**
#171 **Gardenia augusta** Common Names: gardenia, cape jasmine Family: Rubiaceae (madder Family). Plant1 from Floridata: click for Plant Profile ...
www.floridata.com/ref/g/gard_aug.cfm - 20k - Cached - Similar pages - Note this

【梔子】幸せの香り。梔子ってどんな効能があるの？｜デザインネイチャー
クチナシ（梔子、英名 Common gardenia、学名"**Gardenia augusta**"、シノニム:"Gardenia jasminoides"）は、アカネ科・クチナシ属の常緑低木。 情報元：wikipedia URL:http://ja.wikipedia.org. 梔子: 控えめで甘い香りがすごく癒される。 ...
www.cixtow.com/dnature/クから始まるハーブ_/梔子.html - 15k -

Gardenia, Jazmín del Cabo - Gardenia jasminoides - [Translate this page]
Nombre científico o latino: Gardenia jasminoides. - Sinónimo: **Gardenia augusta**. - Nombre común o vulgar: Gardenia, Jazmín del Cabo. - Familia: Rubiaceae. ...
fichas.infojardin.com/arbustos/gardenia-jasminoides-gardenia-jazmin-del-cabo.htm - 71k -

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

Outline

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 sceince, 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

(plant$_{\#n\#2}$, flora$_{\#n\#2}$, plant_life$_{\#n\#1}$) – a living organism lacking the power of locomotion

Relevant KIMD Senses

plant n. 1. living organism with leaves and root, *tumbuh-tumbuhan*.
flora n. *flora*.

Generated Malay WordNet Synset

(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

dictionary, lexicon
a reference book containing an alphabetical list of words with information about them

English

diccionari, lèxic
Obra de referència on es recullen alfabèticament les paraules

Catalan

diccionario, léxico
Obra de referencia donde se recogen alfabéticamente las palabras

Spanish

kamus, leksikon

Malay

...

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**$_{n.1}$nya dari papan catur.

gajah *(n.)*

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

(Plenty of existing algorithms/tools for English WordNet
$\Rightarrow$ reuse with Malay WordNet)

Outline

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