Creating Javanese Wordnet: Between Challenge and Opportunity

Totok Suhardijanto
Department of Linguistics, Faculty of Humanities
University of Indonesia
The language

• **Number of speakers:** 84,3 million (Ethnologue 2003)

• **Classification:** Malayo-Polynesian (branch of Austronesian)

• **Location:**
  a) Three provinces in Java island (Central Java, Yogyakarta, and East Java);
  b) Javanese settlements in Sumatera, Kalimantan, Sulawesi, Papua, and Maluku;
  c) Malaysia, Singapore, the Netherlands, Suriname, and New Caledonia.

• **Grammar:** an agglutinative language in which grammatical relations are expressed by affixes
### Style

- **Register (Style):**
  1. ngoko (informal/low),
  2. madya (neutral/plain),
  3. krama (polite formal/high)

<table>
<thead>
<tr>
<th>NGOKO</th>
<th>MADYA</th>
<th>KRAMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>kowe ‘you (2SG Pron)’</td>
<td>sampeyan ‘you (2SG Pron)’</td>
<td>panjenengan ‘you (2SG Pron)’</td>
</tr>
<tr>
<td>tuku ‘to buy’</td>
<td>tumbas ‘to buy’</td>
<td>mundhut ‘to buy’</td>
</tr>
<tr>
<td>bungah ‘happy’</td>
<td>bingah ‘happy’</td>
<td>rena ‘happy’</td>
</tr>
</tbody>
</table>
Dialect (1/4)

• Three main dialects:
  1. Eastern Dialects:
     • Malang,
     • Surabaya,
     • Banyuwangi
  2. Central Dialects:
     • Semarang,
     • Solo-Yogyakarta
  3. Western Dialects:
     • Banyumas,
     • Tegal,
     • Cirebon,
     • Indramayu,
     • Banten
Dialect (2/4)

- Phonetically, in general, there are only two dialects:

<table>
<thead>
<tr>
<th>/å/ dialect</th>
<th>/a/ dialect</th>
</tr>
</thead>
<tbody>
<tr>
<td>sångå [sɔ · ηɔ] ‘nine’</td>
<td>sanga [sa · ŋa] ‘nine’</td>
</tr>
<tr>
<td>åjå [ɔ · ɉɔ] ‘do not’</td>
<td>aja [a · ja] ‘do not’</td>
</tr>
<tr>
<td>tekå [tə · kɔ] ‘to come’</td>
<td>teka [tə · ka] ‘to come’</td>
</tr>
</tbody>
</table>

- It is written in the same spelling.
Dialect (3/4)

• Grammatically, there are also two dialects.

<table>
<thead>
<tr>
<th>Different from Old Javanese (without particle)</th>
<th>Similar to Old Javanese (with particle)</th>
<th>Old Javanese</th>
</tr>
</thead>
</table>
| anakku  
child + I (1SG pos)                   | anake inyong  
child + e part + I (1SG)             | wĕka ni nghulun  
child + ni particle + I (1SG)         |
Dialects (4/4)

- Although different in some vocabularies, basically all three main dialects can be classified into two groups.

<table>
<thead>
<tr>
<th>Eastern Dialect</th>
<th>Central Dialect</th>
<th>Western Dialect</th>
</tr>
</thead>
<tbody>
<tr>
<td>/å/ group</td>
<td>/å/ group</td>
<td>/a/ group</td>
</tr>
<tr>
<td>No possessive particle</td>
<td>No possessive particle</td>
<td>possessesive particle</td>
</tr>
</tbody>
</table>
Which Dialect for Wordnet?

• Which dialect should be chosen for creating Javanese Wordnet?

• In first step, we start from Solo-Yogya dialect which is regarded as the standard dialect of javanese; it is comprehensible by speakers of other dialects.

• Solo/Yogya dialect is an /å/ dialect, but not different in terms of spelling with other dialects.

  • segara ‘sea’ (central/Solo-Yogya (CESY) dialect) -- [sə · ɡɔ · rɔ]
  • segara ‘sea’ (western (WEST) dialect) -- [sə · ɡɔ · rɔ]
  • segara ‘sea’ (eastern (EAST) dialect) -- [sə · ɡɔ · rɔ]
Resources

• Dictionaries:
  • 36K words from Bausastra Jawa (Standard Javanese Dictionaries) (available online)

• Wikipedia
  • Influenced and interfered by Indonesian
  • After 90s, Javanese texts have been interfered and mixed by Indonesian vocabularies.

• Texts
  • Digitalized literatures, articles, etc. (sastra.org)
Toward Javanese WordNet

• Two methods:

1) Merge
   • Build Javanese taxonomies (synsets and relation) and then mapping them to the Princeton WordNet (PWN) structure

2) Expand
   • Translate Javanese words directly to PWN’s by adding Javanese head words to PWN
   • This method allows the preservation of original WN structure
   • It is simple and the result could be automatically aligned to all other wordnets.
Synset Translation (1)

1) How to deal with a concept that has not an equivalent Javanese lexicon? Compound word or phrase is acceptable or not?
2) Which one is better: synset-to-synset or word-to-word equivalent?
Synset Translation

• Concepts with no equivalent words:
  1. colt -- (a young male horse under the age of four)
  2. mare -- (female equine animal)
  3. stallion -- (uncastrated adult male horse)

• In this case, phrase or compound word will be adopted to use as the equivalent lexical item.
Synset Translation

PWN

• goat, caprine animal -- (any of numerous agile ruminants related to sheep but having a beard and straight horns)

JAVANESE
a) wedhus ‘goat’ (plain word)
b) menda ‘goat’ (high polite word)
c) kambing ‘caprine animal’
Synset Translation

1) goat
   a) wedhus ‘goat’ (plain word)
   b) menda ‘goat’ (high polite word)
   c) kambing ‘caprine animal’

2) caprine animal
Synset Translation

• Translation principles:
  1) *Single word* refers to a lexicalized (PWN) concept in Javanese;
  2) *Compound/phrase* is represented some concepts with no equivalent lexical items in Javanese;
  3) Synset-to-synset equivalent is used for finding Javanese synset that is compatible to PWN synset
  4) Word-to-word equivalent is used for finding the right Javanese word that is compatible to PWN word in each synset
Other problems

1. How to deal with dialect?
   • We can use tags feature that is provided in WN to annotate a local/dialectal word.
     • \textit{koen} -- (you, 2SG) + (EAST)
     • \textit{kowe} -- (you, 2SG) + (CENTRAL)
     • \textit{rika} -- (you, 2SG) + (WEST)
Other problems

2. How to deal with style/register?
   • We can also use tags feature to annotate a word in a style.
     • kowe -- (you, 2SG) (CENTRAL) + (LOW)
     • sampeyan -- (you, 2SG) (WEST) + (HIGH)
     • panjenengan -- (you, 2SG) (WEST) + (POLITE HIGH)
Other problems

3. How to deal with Javanese concepts that do not exist in the PWN?
   • In our future plan, for any concepts that do not exist in the PWN, new synsets will be added.
   • For example:

PWN
   • rice -- (grains used as food either unpolished or more often polished)
JAVANESE
• beras  ‘rice grain’ (plain word)
• wos    ‘rice grain’ (polite word)
• sega   ‘cooked rice’ (plain word)
• sekul  ‘cooked rice’ (polite word)
• pari   ‘rice plant’ (plain word)
• pantun ‘rice plant’ (polite word)
• upa    ‘a single cooked rice seed’
• menir  ‘broken rice’
Conclusion

1. Javanese WordNet can be develop by using expand method because of its simplicity and possiblity to link to all other WN
2. Resources: *bausastra jawa*, wikipedia, digitalized documents
3. Register (style) and dialect issues: tags or annotation
4. No lexical items in Javanese => *compound, phrase*
5. No concepts in WN (because of different cultures) => adding new synset
References


Thank you

Matur nuwun