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)

NGOKO	MADYA	KRAMA
kowe 'you (2SG Pron)'	sampeyan 'you (2SG Pron)'	panjenengan 'you (2SG Pron)'
tuku 'to buy'	tumbas 'to buy'	mundhut 'to buy'
bungah 'happy'	bingah 'happy'	rena 'happy'

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:

/å/ dialect	/a/ dialect
sångå [sɔ · ŋɔ] ‘nine’	sanga [sa · ŋa] ‘nine’
åjå [ɔ · ɟɔ] ‘do not’	aja [a · ɟa] ‘do not’
tekå [tə · kɔ] ‘to come’	teka [tə · ka] ‘to come’

- It is written in the same spelling.

Dialect (3/4)

- Grammatically, there are also two dialects.

Different from Old Javanese (without particle)	**Similar to Old Javanese (with particle)**	**Old Javanese**
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.

Eastern Dialect	Central Dialect	Western Dialect
/å/ group	/å/ group	/a/ group
No possesive particle	No possesive particle	possesive particle

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 · gɔ · rɔ]
  - segara ‘sea’ (western (WEST) dialect) -- [sə · ga · ra]
  - segara ‘sea’ (eastern (EAST) dialect) -- [sə · gɔ · 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 worddnets.

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

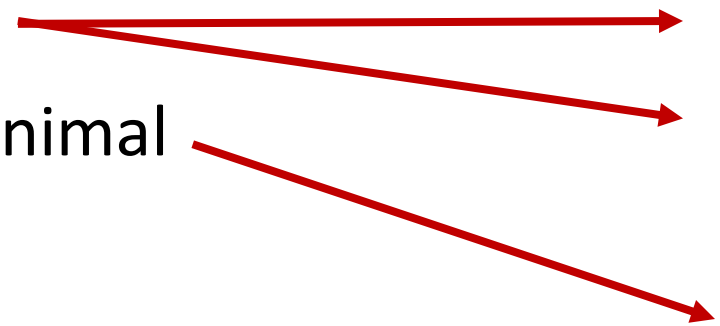
- goat, caprine animal -- (any of numerous agile ruminants related to sheep but having a beard and straight horns)

JAVANESE

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

Synset Translation

1) goat
2) caprine animal



a) wedhus ‘goat’ (plain word)
b) menda ‘goat’ (high polite word)
c) kambing ‘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.

  - koen -- (you, 2SG) + (EAST)
  - kowe -- (you, 2SG) + (CENTRAL)
  - 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

Bond, Francis, LianTze Lim, Enya Kong Tang and Hammam Riza. 2014. The combined Wordnet Bahasa. In Siaw-Fong Chung and Hiroki Nomoto, eds. *Current Trends in Malay Linguistics*. NUSA 57: 83–100.

Thoongsup, Sareewan, Kergrit Robkop, Chumpol Mokarat, Tan Sinthurahat, Thatsanee Charoenporn, Virach Sornlertlamvanich, Hitoshi Isahara. 2009. Thai WordNet Construction. Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009, pages 139–144, Suntec, Singapore, 6-7 August 2009.

Isahara, Hitoshi, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Proceedings of LREC 2008*.

Thank you

Matur nuwun