Building an open concordancer for Malay/Indonesian

In this talk, we introduce a new open online concordancer for Malay/Indonesian that we built, MALINDO Conc, and compare it with the existing open concordancers. MALINDO Conc was designed as a common tool among researchers of Malay/Indonesian that is free of charge, easy to use and yet allows moderately sophisticated search queries.

MALINDO Conc was modelled after the Malay Concordance Project (`http://mcp.anu.edu.au/`), an open online concordancer for Classical Malay. It thus inherits some of the good features of the latter. First, MALINDO Conc intends to include **any variety of Malay** across the archipelago. The existing open concordancers, i.e. Korpus DBP (`http://sbmb.dbp.gov.my/korpusdbp/SelectUserCat.aspx`) and SEALang Library Corpora (`http://sealang.net/malay/corpus.htm`, `http://sealang.net/indonesia/corpus.htm`), on the other hand, deal with a particular geopolitical variety of Malay, which is either Malaysian/Singaporean/Bruneian Malay or Indonesian.

Secondly, MALINDO Conc allows **morphological search**. Thus, one can search the corpus for forms with a particular morphological profile. Some possible queries include:

- Keyword: root = *fikir* & prefix = *meN-* or *di-* or Ø & suffix = *-kan* or Ø & circumfix = Ø
  (inflected forms of *fikir* and *fikirkan*)
- Keyword: root = unspecified & circumfix = *ber-. . . -kan*
  (*ber-. . . -kan* verbs)
- Keyword: root = unspecified & prefix = *meN-* & reduplication = full
  (*meN-X-X* and *X-meN-X* verbs)
- Keyword: surface form = *ingin*
  Collocate: Find collocate = between R1 and R2 & root = unspecified & prefix = *di-*
  (*ingin* + *di-* verb, *ingin* + word (e.g. *untuk*) + *di-* verb)

The morphological search function expands the range of investigations one can conduct using corpora. Korpus DBP and SEALang Library Corpora only allow simple keyword search. Since they do not support queries using regular expressions (but "*" and "?" wild cards in Korpus DBP), one's search must be based on a particular lexical item, limiting possible corpus-based studies mostly to lexical ones. Morphological search makes it possible to refer to abstract classes, including those mentioned in the list above. Furthermore, the syntactic category of an affixed word is often predictable from the outermost affix in it. Therefore, MALINDO Conc can be used for morphosyntactic studies too.

Thirdly, MALINDO Conc accept **contributions from users**. Currently, MALINDO Conc's corpus consists only of the reclassified version of the Leipzig Corpora Collection (Goldhahn et al. 2012; Nomoto et al. under review). In the future, we will also include in the MALINDO Conc's corpus, data collected by others as well as ourselves, especially spoken data and data from regional Malay varieties.

On top of the three features above, MALINDO Conc has the following two features that are not found in the Malay Concordance Project. First, MALINDO Conc is **localized**. Specifically, the user interface and manual are provided in Indonesian (and also in Malay in the future). Localization is important because most of the users of MALINDO Conc will be speakers of Malay/Indonesian. Secondly, search results are **downloadable**. Both of these features are found with Korpus DBP, but not with SEALang Library Corpora.

References

Goldhahn, Dirk, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

Nomoto, Hiroki, Shiro Akasegawa, and Asako Shiohara. under review. Reclassification of the Leipzig Corpora Collection for similar languages: Malay and Indonesian.