Reclassifying the Leipzig Corpora Collection for Malay/Indonesian

It is often claimed that linguistic description and theorizing should be based on naturalistic data. The use of corpora is promoted as a way of realising this idea. In reality, however, the range of research one can conduct varies significantly depending on the quality and quantity of the corpora available in the language s/he works on. The idea of web corpora, whose data is collected by automated web crawling, helps alleviate the problem of quantity disparity in corpus linguistics in less studied languages with a large speaker population such as Malay/Indonesian. However, the quality is not necessarily guaranteed. This study reports our attempt at partially resolving the quality problem through the reclassification of the Malay/Indonesian subcorpora of the Leipzig Corpora Collection (LCC; Quasthoff et al. 2006), using a language-specific language identification approach.

LCC has four distinct categories for Malay/Indonesian: `msa`, `ind`, `ind-id`, `ind-bn`. These language code names suggest that they represent Malaysian Malay, Indonesian and Brunei Malay, Indonesian, and Brunei Malay respectively. This is the case for some subcorpora, but not for others. For instance, `msa_newscrawl_2011` contains Brunei Malay data from *Pelita Brunei*. Likewise, `ind_mixed_2012` contains a number of Malaysian Malay sentences. In order to make LCC more reliable, we have reclassified the data based on linguistic properties.

The reclassified version has three regional variety categories: `zsm` (Malaysian/Singapore/ Brunei Malay), `ind` (Indonesian) and `msa` (indeterminate). The last category contains sentences whose regional variety cannot be determined either by their linguistic forms or by the country domain of the website from which they were extracted.

Language identification was carried out page by page according to the following algorithm:

(1) a. (i) For each sentence, count the frequency of the words in the list of spelling differences between Malaysia and Indonesia (Nomoto et al. 2014). Identify the sentence as the language with the higher frequency.

(ii) Count the numbers of `zsm` and `ind` sentences contained in the page. Identify all the sentences in the page as the language with the higher frequency.

b. If (1a) fails, repeat (1a) by replacing the list of spelling differences with the lists of 10,000 most frequent words in `zsm` and `ind`.

c. If (1b) fails, identify the language of the page based on the country domain, i.e. `.my`, `.sg` and `.bn` as `zsm`, `.id` as `ind`, and other domains as `msa`.

(2) *Example* (source: `http://artikel.sabda.org/book/export/html/21`)
Di ayat yang ke 6 dikatakan **bahawa** Yesus melihat keadaan orang sakit itu dan Yesus tahu **bahwa** dia sudah lama sakit, lalu Yesus bertanya kepadanya maukah engkau sembuh? Dibawahnya, **tampak** dua orang yang sedang beristirahat. Diberkatilah orang yang mengandalkan TUHAN, yang menaruh harapannya pada TUHAN!

Sent 1	(1a-i) `zsm` 1 : `ind` 1 → ?	Sent 3	(1a-i) `zsm` 0 : `ind` 0 → ?
Sent 2	(1a-i) `zsm` 0 : `ind` 1 → `ind`	Sent 1–3	(1a-ii) `zsm` 0 : `ind` 1 → `ind`

We note that some spelling differences are more reliable diagnostics than others. This is because as the naturalness increases, speakers' use of substandard spellings also increases. For example, *bahawa*, which is standard in `zsm` but not in `ind`, is often found in otherwise Indonesian sentences. By contrast, *iaitu*, which is similarly standard only in `zsm`, is rarely used in Indonesian. We will compare the results of our language-specific language identification method and those of general language identification techniques proposed in past VarDial (Workshop on NLP for Similar Languages, Varieties and Dialects) shared tasks.

References

Nomoto, Hiroki, Nahoko Yamashita, and Ayano Osaka. 2014. Senarai komprehensif perbezaan ejaan Malaysia dan ejaan Indonesia. In *Gogaku Kenkyuujo Ronshuu 19*, 21–31. Tokyo: Tokyo University of Foreign Studies. [A comprehensive list of spelling differences between Malaysia and Indonesia].

Quasthoff, Uwe, Matthias Richter, and Christian Biemann. 2006. Corpus portal for search in monolingual corpora. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 1799–1802. Genoa.