**Towards a Japanese-Indonesian parallel corpus: Translating the *Balanced Corpus of Contemporary Written Japanese***

Building a language resource is one of the earliest stages in linguistic research, in general, and Natural Language Processing (NLP), in particular, and well-researched languages (e.g., English, German) have plenty of language resources to work with. However, NLP research is still needed for under-resourced languages, such as Indonesian, in spite of the fact it is one of the most frequently spoken languages in the World (more than 230 million people). Moreover, in order to be useful, these resources should: i) contain a reasonable amount of data, since most research methods are based on statistics; ii) rely on robust processing tools (for cleaning, alignment, tagging, etc.); iii) needless to say, they should be publicly available to be easily accessed by as much as users as possible. So far, attempts to build Indonesian-English parallel corpora have been made, namely, IDENTIC Corpus[1] and PANL-BPPT Parallel Corpus[2]. However, they are restricted in size (IDENTIC contains 994,545 words, while PANL-BPPT has only 500,000 words) and they show some inconsistences in sentence alignment (some sentences appear twice) and/or in POS tagging (e.g., the word "tahu" is always tagged as verb).

This work introduce a project of building a Japanese-Indonesian parallel corpus translating the *Balanced Corpus of Contemporary Written Japanese* (BCCWJ)[3], a large-size (one hundred million words) corpus of Japanese compiled at National Institute for Japanese Language and Linguistics (NINJAL) and publicly released in 2011. The project is part of a bigger work to translate the corpus in several languages (i.e., English, Chinese, Italian, Indonesian), finally aimed to provide researchers interested in the target languages, proper multilingual data set to apply their linguistic studies. We are now working at the translation of a small-size data sample (around 7,000 Japanese characters) in the context of a feasibility study granted by NINJAL. The selected data are extracted from a collection of a Japanese bulletin board, i.e. Yahoo! *Chiebukuro* (3 samples of 84, 148, and 77 characters), and a blog, i.e. Yahoo! *Burogu* (3 samples of 85, 162, and 136 characters). We will use the translation output to test sentence alignment and segmentation, as well as to set-up the guidelines for text processing tasks as tokenization (e.g., Indonesian needs specific rules for reduplication and clitics), morphological annotation, and POS tagging.

---

1 http://www.lrec-conf.org/proceedings/lrec2012/summaries/644.html
2 http://panl10n.net/english/OutputsIndonesia2.htm
3 http://www.ninjal.ac.jp/english/products/bccwj/