Building a Tokenizer for Indonesian

David Moeljadi and Hannah Choi
Division of Linguistics and Multilingual Studies
Nanyang Technological University, Singapore

Tokenization or word segmentation is the task of separating out or tokenizing words from running text (Jurafsky and Martin, 2009, p. 81). The output tokens can be the input for further processing such as syntactic analysis, semantic annotation, or text mining. This study deals with tokenization for standard, formal Indonesian text: how the affixes, clitics, and reduplication should be treated; which affixes should be split from the stem and why, by using Wordnet Bahasa (Bond et al., 2014) and Python programming language (`www.python.org`). This is illustrated as follows.

Input	*Berikanlah*	*permata-permatamu*	*yang*	*termahal*	*kepadaku*
	give-IMP	jewel-REDUP=2SG	REL	SUP-expensive	to=1SG
Output	*berikan*	*permata mu*	*yang*	*ter mahal*	*kepada ku*

The evaluation or judgment criteria I set for splitting the affixes are: (1) Productivity of the affixes. The more productive an affix is, the more possibility it has to be split; (2) Consistency and predictability of meanings. The more consistent and predictable the meaning of an affix is, the more possibility it has to be split; and (3) Necessity of adding subtle meanings to Wordnet. Regarding the productivity, consistency, and predictability of the affixes, I refer to Kridalaksana, (1989), Voskuil, (1996), Sneddon et al., (2010), Alwi et al., (2014), and *Kamus Besar Bahasa Indonesia Edisi Kelima* (Amalia, 2016). The general rules are: (1) split the affixes from a word form if it is not in Wordnet, (2) productive affixes which have consistent, predictable meanings should be split (based on the judgment criteria above), and (3) there is no need to add very subtle meanings to Wordnet, e.g. collective meaning "in a group of ...". This tokenizer will be used for preprocessing in INDRA (Indonesian Resource Grammar), a computational grammar for Indonesian (Moeljadi, Bond, and Song, 2015) and sense annotation in Wordnet.

References

Alwi, Hasan et al. (2014). *Tata Bahasa Baku Bahasa Indonesia*. 3rd ed. Jakarta: Balai Pustaka.

Amalia, Dora, ed. (2016). *Kamus Besar Bahasa Indonesia*. 5th ed. Jakarta: Badan Pengembangan dan Pembinaan Bahasa.

Bond, Francis et al. (2014). "The combined Wordnet Bahasa". In: *NUSA: Linguistic studies of languages in and around Indonesia* 57, pp. 83–100.

Jurafsky, Daniel and James H. Martin (2009). *Speech and Language Processing*. 2nd ed. New Jersey: Pearson Education, Inc.

Kridalaksana, Harimurti (1989). *Pembentukan Kata dalam bahasa Indonesia*. Jakarta: Gramedia.

Moeljadi, David, Francis Bond, and Sanghoun Song (2015). "Building an HPSG-based Indonesian Resource Grammar (INDRA)". In: *Proceedings of the GEAF Workshop, ACL 2015*, pp. 9–16.

Sneddon, James Neil et al. (2010). *Indonesian Reference Grammar*. 2nd ed. New South Wales: Allen & Unwin.

Voskuil, Jan (1996). *Verb Taxonomy in Indonesian, Tagalog and Dutch*. Holland Institute of Generative Linguistics.