A code-switching corpus for Indonesian-German based on the web forum *kaskus.co.id*

Hani Priandini
Universität Hamburg
priandini.hani@gmail.com

Mariska Ajeng Harini
Universität Hamburg
mariskaajeng@gmail.com

Heike Zinsmeister
Universität Hamburg
heike.zinsmeister@uni-hamburg.de

Tag
Datum
Zeit
Raum

Conversational code switching (CS) is defined as "the juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or subsystems" (Gumperz 1982, 59). Intrasentential CS is characterized by "elements from two (or more) language varieties in the same clause, but only one of these varieties is the source of the morphosyntactic frame for the clause." (Myers-Scotton 1997, 3.4). CS is also observed on the subword level. The use of CS has been linguistically analyzed for its grammatical structure (e.g. Muysken 2000) as well as for its sociolinguistic implications (e.g. Grosjean 1982). It poses great challenges for many natural language processing applications such as parsing, machine translation, speech recognition, and information extraction. CS annotated corpora exist, e.g., for the language pairs Hindi-English (Vyas et al. 2014) and Turkish-German (Çetinoğlu & Çöltekin 2016).

In this poster, we present the InDeu corpus, the first CS corpus for Indonesian-German. The corpus was sampled and annotated for performing linguistic analyses. In a secondary step, it was semi-automatically converted to the format of the shared task of language identification in CS data. It comprises 586 posts from the subforum *Germany* of the Indonesian webforum *Kaskus.co.id* and contains 1384 manually annotated switches. We extended Muysken (2000)'s subclasses by our own more fine-grained CS tagset for linguistic analyses. Analyzing parts of speech showed that nouns were inserted most frequently, which confirms findings in other languages. Furthermore, the user status and location had an influence on the CS types.

References: • Çetinoğlu, Ö. & Ç. Çöltekin. 2016. Part of Speech Annotation of a Turkish-German Code-Switching Corpus. In Proceedings of LAW-X. • Grosjean, F. 1982. Life with two Languages: an Introduction to Bilingualism. Harvard University Press. • Gumperz, J. 1982. Discourse strategies. Vol. 1. Cambridge University Press. • Myers-Scotton, C. 1997. Duelling languages: Grammatical structure in codeswitching. Oxford University Press. • Muysken, P. 2000. Bilingual speech. A typology of code-mixing. Cambridge University Press. • Vyas, Y. et al. 2014. POS Tagging of English-Hindi Code-Mixed Social Media Content. In Proceedings of EMNLP.