Comparative study of register specific properties of Indonesian SMS and Twitter: implications for NLP

Claudia M. Brugman and Thomas J. Conners, Center for Advanced Study of Language

Abstract

This talk presents research in progress on specific dialect properties of Indonesian as produced on two electronic platforms: SMS and Twitter. What we'll show in this talk includes the following: Indonesian language properties differ not only between the more standard media such as speech and writing, vs. electronic media, but there is also a difference of the properties of language produced on these two platforms. Second, the properties that characterize the two electronic media, as against the media with longer histories, firmly plant SMS and Twitter as ways of recording spoken Indonesian, rather than being another medium for writing Indonesian. Third, we will talk about the ways that electronic Indonesian represent non-standard Indonesian, and the problem that that presents for performing NLP operations on these messages. We'll give some detailed examples of properties of actually-produced Indonesian and what it would take to smooth them over for NLP. We'll talk about how these special properties of ElectrIndo give clues as to the social identity and social relationships of the people in the communications, and we'll give some details about the differences between the language used in SMS and Twitter. We will wrap up with some implications for NLP and for future research on the intersection between variationist sociolinguistics and the problems of electronically produced communications. The implications go beyond differences between these two platforms in particular, but rather extend to other platforms whose production processes are more or less conversation-like.

This kind of study, which compares two electronic media, has rarely been done, and to our knowledge has been accomplished only on "major" languages such as English or German. This belies the huge populations of avid users of non-Indo-European languages. Our study included a preliminary examination of some indicators of lexical density and frequency, and clausal complexity, as well as describing orthographic variation in terms of its communicative value (e.g. as mimicking prosodic features of the corresponding spoken form). The analyses were performed on original corpora, collected in-country.