Using Twitter to study the evolution of conversational utterances

C. M. Alis and M. T. Lim

National Institute of Physics, University of the Philippines Diliman, 1101 Quezon City, Philippines

mlim@nip.upd.edu.ph

Twitter hosts millions of conversations daily, making it a natural platform for understanding conversations. In this work, we track how the length of utterances, or a single speaking turn, changes with time. We consider Twitter messages (or ‘tweets’) that begin with an @username to be an utterance. Figure 1 shows representative utterance length distributions for days from different years that fall on a Friday in December. While the natural utterance length distribution in terms of characters is typically unimodal [1], a typical distribution in Twitter is bimodal: the second peak being due to the 140-character limit. Errant spikes in Fig. 1 are due to spam.

Using a sample size of 229 million utterances spanning three years (September 2009-December 2012) collected through the Twitter Streaming API [2], we find that the median utterance length has been shortening at an average rate of 5.20 chars/year. We argue that this shortening is due to the increasing usage of jargon, including coined words.


Fig. 1. Representative utterance length distributions on Twitter.