Is text length a linguistic variable? Evidence from social media

Aatu Liimatta (University of Helsinki)

When text length within a corpus varies dramatically, normalized word frequencies are an unreliable indicator of linguistic variation. In order to reduce this effect in statistical corpus studies, it is common to simply remove texts shorter than some threshold from the data. For example, studies on Wikipedia often exclude the shortest articles (cf. Hiltunen 2014). In many published corpora, only a sample of fixed length is included from each text to lessen the effect of text length on token frequencies (cf. e.g. Francis and Kučera 1964). Ideal text length has been the subject of discussion for a long time (cf. e.g. Biber 1993).

However, texts have a structure: a short text is not equivalent to a sample of similar length from a longer text. Limiting in any way the length of texts considered in a study means that that shorter texts and their unique features are effectively ignored in studies. This is not so problematic with many of the more "traditional" genres included in corpora, as they tend to be made up of longer texts to begin with. However, if we want to study variation within various CMC genres, and particularly between the smallest units of social media communication, i.e. between single social media comments, the issue becomes more noteworthy, as social media comments typically consist of very short texts or texts of varying length.

In this paper, I will focus on Reddit, the third-largest English-language social media platform; preliminary data indicates that 50% of Reddit comments are under 20 words long. Sharing many similarities with traditional online forum platforms, Reddit contains discussions on a wide variety of topics. Moreover, Reddit effectively does not limit the length of a comment (unlike e.g. Twitter), which makes it possible to compare comments of very different length.

Starting with the hypothesis that text length is determined in part based on the purpose of the text, and so is related to the register of the text, we should be able to observe differences in the distribution of register features such as word length or first-person pronouns and their collocation patterns in texts of different length. These differences will be analyzed using statistical corpus methods and provide evidence for whether or not text length in itself is a relevant register feature in social media comments, a genre which is characterized by highly varying text length. Preliminary results based on Reddit data indicate that the distribution of first-person pronouns correlates to a degree with text length, with shorter comments generally containing more first-person pronouns.

References

Biber, D., 1993. Representativeness in corpus design. Literary and Linguistic Computing, 8(4), 243-257.

Francis, W. N. and Kučera, H., 1964. *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown)*. Providence, Rhode Island: Brown University.

Hiltunen, T., 2014. Choice of national variety in the English-language Wikipedia. Tyrkkö, J. and Leppänen S. (Eds.), *Texts and Discourses of New Media.* Helsinki: VARIENG. Retrieved from http://www.helsinki.fi/varieng/series/volumes/15/hiltunen/.