Lexical diversity measurement using subsample entropy: formalism and evaluation

Loïc Jeanson[1], Guillaume Guex[2], Aris Xanthos[3]

[1]University of Lausanne – loic.jeanson@unil.ch

[2]University of Lausanne – guillaume.guex@unil.ch

[3]University of Lausanne – aris.xanthos@unil.ch

Abstract

Lexical diversity (LD) has been computed in various ways. To alleviate the notorious dependence of LD measures on sample length, resampling strategies have proved most successful. They ultimately boil down to the basic idea of averaging LD measures over multiple subsamples taken from the original sample. Over the last decade, there has been an increased interest in using entropy estimators to measure LD. Entropy has the particularity of accounting for differences in relative frequencies between types, but it shows a non-negligible dependence to sample length. In this contribution, we assess the usefulness of employing a resampling strategy to construct a new LD measure called "subsample entropy". We systematically evaluate its robustness and sensitivity in comparison to established LD indexes, using text samples varying in length and style. We conclude that subsample entropy fares comparatively well due to its remarkable robustness and overall versatility.

Keywords: Lexical diversity, entropy, subsample entropy, measure, index, sample length, resampling, robustness, sensitivity, evaluation, Zipf-Mandelbrot's law.