TESTING THE CORE LANGUAGE HYPOTHESIS: SOME EFFECTS OF TEXT MIXING

JAKE R. WILLIAMS AND PETER S. DODDS

Natural languages are full of rules and exceptions. Perhaps one of the most studied (quantitative) rules is Zipfs law [1], which states that the frequency of occurrence of a word is approximately inversely proportional to its rank. Though this law of ranks has been observed to hold across disparate texts and forms of data, it has been observed as recently as 2001 [2, 3] that multiple scaling regimes exist. These two regimes are purported [2, 4, 5] to be attributed to two forms of language the kernel (core) and unlimited (non-core) lexica. In light of a recently proposed model for the stochastic growth of such languages, we investigate the validity of the core/non-core hypothesis with empirical data. Our experiments focus on texts that exhibit this dual behavior, which we note are most commonly of mixed origin. We hypothesize and show strong evidence that the act of mixing texts leads to an effective decay of word introduction (such as with the aforementioned model). We then go on to measure this decay via several methodologies, applying our measurements to achive reasonable estimates of the location and severity of lower scalings.

Date: March 31, 2014.