A robust statistical model of word frequencies

Michael Ingelby and Serge Sharoff
University of Leeds, United Kingdom.

For the purposes of language teaching or automatic language processing it is important to know how frequent a word is. However, a simple procedure counting the number of times a word occurs in a collection of texts leads to many unfortunate artefacts because some words occur too often in a small number of texts leading to frequency bursts. Our task in this paper is to introduce a statistical model which uses methods from robust statistics to estimate the frequencies of words in a collection of texts.

Keywords: Robust statistics; word frequencies; core lexicon.