"But this formula doesn't mean anything!"

Ylva Berglund, Uppsala University
Oliver Mason, University of Birmingham

When working with quantitative measures of even simple statistics, one is sometimes confronted by colleagues who fail to see the significance of such studies, claiming there was no meaning in a statistical formula. Even when formulae are used to measure some textual parameter, the results may be difficult to interpret. What does it actually mean if a text has an average word length of 4.67 and an TTR of 0.34?

In this paper we start from earlier research into measuring the performance of non-native speakers (or rather writers). We have previously shown (Proc of TaLC 2000, to appear) that automatic stylistic assessment can be used to distinguish between different kinds of texts, both those of different genres and, with even better results, native versus non-native authorship. Taking into account a variety of 'surface' parameters which were measured for essays written by Swedish learners of English at University level, we will now look at a possible mapping between a number of numerical values and the quality of the essay, which cannot be directly quantified itself. Using PCA and cluster analysis we try to demonstrate what gives an English text the 'Englishness', and how non-native speakers develop their language ability in this 'textual parameter space'. We also hope to show that even if a formula as such does not mean anything by itself, quantitative measures are a valuable means that can be used to enhance our understanding of intuitive judgements about texts.