Abs-122

John Flowerdew (City University of Hong Kong)

Small corpora, larger corpora and discourse analysis in the study of lexical cohesion

Acknowledging that there may be degrees of gradience, Bednarek (2009) argues for a three-pronged approach to corpus study: small-scale corpus analysis, large-scale corpus analysis, and manual analysis of individual texts. This paper exemplifies this approach by showing some ongoing work on lexical cohesion. Small corpora are typically about 100,000 words in size. Their advantage is that they allow the research to become very familiar with the corpus content and thus facilitate discourse analysis and that hand tagging of features that would not be susceptible to automated tagging can be done. Their limitation is that the frequency data they provide may not be representative. The study reported in this paper employs a large small corpus (some 600,000 words). The corpus has been tagged using a semi-automated process. Once tagging has been done, quantitative and qualitative data can be instantly retrieved. The large number of instances of given patterns revealed by the concordancer allows for analysis of individual examples in context (discourse analysis). Where examples of minor categories are few in number, then reference corpora such as BNC and COCA can be used to verify if such patterns are significant or not. The procedures referred to above will be exemplified during the presentation, using the corpus in question, along with a concordancer.

Bednarek, M. (2009). Corpora and discourse: a three-pronged approach to analyzing linguistic data, HCSNet Summerfest ''08, Cascadilla Proceedings Project, Somerville, MA, USA, 19-24.