Key category analysis of a spoken corpus for EAP

This paper describes the statistical comparison procedure followed to compare a Needs-Driven Spoken Corpus (NDSC) for EAP (English for Academic Purposes) use (Jones 2003) with the British National Corpus Spoken Sampler, as the normative corpus, in an attempt to reveal overrepresented grammatical categories in the former for subsequent in-depth investigation.

The program used for the statistical analysis was *Wmatrix* (Rayson 2003), which uses corpus annotation techniques to enable the researcher to analyse more data in a shorter period of time than if he /she were using other tools. Word-class and semantic field tags are used to find key grammatical categories in the NDSC by comparing the two corpora. This type of key category analysis reduces the overload on the analyst compared to a key word analysis.

The paper will also report on the findings of this statistical analysis and examples of overrepresented categories in different genres in the NDSC will be provided. Some of these categories will be analysed in context in the final part of the paper.

Dr. Martha Jones, University of Nottingham
Dr. Paul Rayson, Computing Department, Lancaster University
Prof. Geoffrey Leech, Department of Linguistics and MEL, Lancaster University

Jones, M. A. (2003) The development and exploitation of a Needs-Driven Spoken Corpus for students of English for Academic Purposes. PhD Thesis, Lancaster University.

Rayson, P. (2003). Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. Ph.D. thesis, Lancaster University.