Dr Paul Rayson
UCREL, Computing Department
Lancaster University, UK
paul@comp.lancs.ac.uk

Keywords are not enough

Abstract:
This talk reports the development of a new kind of methodology and tool called Matrix (Rayson 2003) for advancing the statistical analysis of electronic corpora of linguistic data. The standard corpus linguistic research process model identifies the research question (and the linguistic features) early in the study. In recent years corpora have been increasingly annotated with linguistic information. We define data-driven corpus research, and describe the Matrix tool which assists in finding candidate research questions. Matrix allows the macroscopic analysis (the study of the characteristics of whole texts or varieties of language) to inform the microscopic level (focussing on the use of a particular linguistic feature) as to which linguistic features should be investigated further. By integrating part-of-speech tagging (Garside and Smith, 1997) and semantic field tagging (Rayson et al, 2004) in a profiling tool, the Matrix technique extends the keywords procedure (e.g. in WordSmith tools, see Scott 2000) to identify key grammatical categories and key concepts. Matrix will be exemplified by the comparison of UK 2001 general election manifestos of the Labour and Liberal Democratic parties. Currently, it has been tested on restricted levels of annotation and only on English language data.

References:

Garside, R., and Smith, N. (1997) A hybrid grammatical tagger: CLAWS4, in Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora.* Longman, London, pp. 102-121.

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

Rayson, P., Archer, D., Piao, S. L., McEnery, T. (2004). The UCREL semantic analysis system. In *proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 25th May 2004, Lisbon, Portugal, pp. 7-12.

Scott, M. (2000). Focusing on the text and its key words. In Burnard, L. and McEnery, T. (eds.) *Rethinking language pedagogy from a corpus perspective: papers from the third international conference on teaching and language corpora.* Peter Lang, Frankfurt, pp. 104 – 121.