Antoinette Renouf
Research and Development Unit for English Studies (RDUES)
University of Central England
Birmingham

Revisiting the 'Corp' in WebCorp

The web, for all its faults and with due caveats, presents the best hope of giving the maximum number of real people access to up-to-date real language use. Since 1998, we in RDUES have confronted the problem of establishing a tool, **WebCorp,** that can provide corpus linguists with the same quality of processed and analysed linguistic output that can be derived from orderly, finite corpora. We have reported on each stage of development at four ICAME conferences.

We have, in the light of personal experience and user feedback via our site http://www.webcorp.org.uk/ , succeeded in providing many of the retrieval and analysis facilities corpus linguists seek, but problems remain. Some of these relate to the nature of the web; for instance, the routine absence of information about language, date and author basic identifiers required for linguistic processing.

Another obstacle for **WebCorp** from the outset has been its reliance on commercial search engines as the gateway to web texts. This mediation has had several deleterious effects. First, it has inhibited the speed of **WebCorp** performance; secondly, the informational focus of the search engines has led to a linguistic naivety which we have had to address; and thirdly, their commercial evolution is characterised by unpredictable changes in service. Recently, for instance, Google's word count statistics have become unreliable, and its search parameters are precluding pattern search with wildcards.

From the outset, it was clear that the dependence of **WebCorp** on commercial search engines would be its Achilles' heel. In 2000, I thus established a relationship with a UK Search Engine, which brought us the dual benefit of fast access to web text to accelerate development, and training in cutting-edge search engine technology. The upshot is that we are now well on the way to creating our own tailored search engine, infrastructure that will ensure the survival and improved reliability and performance of the **WebCorp** tool.

This paper will pinpoint some linguistic and procedural problems in web text search and explain how they will be solved by replacing the commercial search engine with tailored web-search architecture.