The Web as a Source of Linguistic Information

Antoinette Renouf
Research and Development Unit for English Studies, University of Liverpool.

However large and up-to-date the available electronic text corpora are, there will always be aspects of the language which are too rare or too new to be evidenced in them. In fact, the WWW is the largest existing repository of texts across a range of textual domains. It is not surprising that individual corpus linguists have increasingly hit upon the idea of querying the standard web search engines in order to retrieve the more recondite or newly-minted instances of language use. Whilst this strategy can yield useful linguistic results, the standard engines are not designed for the purpose, and the procedure is prohibitively slow and the output requires extensive post-editing. Last year, the Research and Development Unit for English Studies at Liverpool moved on from being such users, taking on board the needs of the community and beginning to develop 'WebCorp', an Internet search system which allows on-line access to web texts as linguistic rather than information sources. A demonstration tool is available at: http://www.webcorp.org.uk. This paper will report on the research initiative and highlight some the issues involved.