Linguistic Search Engine

Adam Kilgarriff
ITRI, University of Brighton, Lewes Road, Brighton, BN2 4GJ
E: adam@itri.brighton.ac.uk T: 01273 642919 F: 01273 642908

We propose to build a linguistic search engine, similar in overall design to Google or AltaVista but meeting the specifications and requirements of researchers into language.

Language scientists and technologists are increasingly turning to the web as a source of language data, because other resources are not large enough, because they do not contain the types of language the researcher is interested in, or simply because it is free and instantly available. The default means of access to the web is through a search engine such as Google. While the web search engines are dazzlingly efficient pieces of technology and excellent at the task they set themselves, for the linguist they are frustrating. The search engine results do not present enough instances (Google sets a limit of 2000) or enough context for each instance (Google generally provides a ca 10-word fragment), they are selected according to criteria which are, from a linguistic perspective, distorting, and they do not allow searches to be specified according to linguistic functions such as lemmatization and word class.

Each of these goals could straightforwardly be resolved, but approaches to Google to date have gone unanswered. However this suggests a better solution: rather than depend upon existing search engines, it would be possible to set up a linguistic search engine, dedicated to linguists' interests. Then the kinds of processing and querying would be designed explicitly to meet linguists' desiderata, without any conflict of interest or 'poor relation' role. Once this is set up, large numbers of possibilities open out. All those processes of linguistic enrichment and `linguistic data mining' which have been applied with impressive effect to smaller corpora could be applied to the web so that web searches could be specified in terms of linguistically interesting units such as lemmas, word classes, and constituents (e.g. noun phrase) rather than strings. Thesauruses and lexicons could be developed directly from the web. The way would be open for further anatomizing of web text types and domains, both a topic of interest in itself and one where strategies would be needed so that web-based lexical resources could be developed for specific text types or domains, or so that the biases of the web could be countered to provide 'general languages' resources from the web. All of this can potentially be done for all of the many languages for which there is ample data on the web.

The web, teeming as it is with language data, of all manner of varieties and languages, in vast quantity and freely available, is potentially a fabulous linguists' playground. The Linguistic Search Engine will bring that dream closer to reality.