Using the web as a source of linguistic data: experiences, problems and perspectives

Marco Baroni, Bologna

Forschungskolloquium Korpuslinguistik, Humboldt-Universität zu Berlin, Nov 24, 2004

Despite certain obvious drawbacks (e.g., lack of control, sampling, documentation etc.), there is no doubt that the World Wide Web is a mine of language data of unprecedented richness and ease of access (Kilgarriff and Grefenstette, 2003).

An increasing body of studies has shown that simple algorithms using Web-based evidence are successful at many linguistic tasks, often outperforming sophisticated methods based on smaller but more controlled data sources (e.g., Turney 2001, Keller and Lapata 2003).

Most Internet-based linguistic studies use one of three strategies to access Web data:

- Collect frequency data directly from an existing search engine (e.g., Turney 2001);

- Use a search engine to find relevant pages, and then retrieve the pages to build a corpus (e.g., Baroni and Bernardini 2004);

- Build a corpus by spidering the web (e.g., Terra and Clarke 2003).

In this talk, I will overview these three strategies, presenting some of the results achieved with each method, and discussing their inherent problems.

I will conclude the overview by suggesting that direct web-spidering is the only viable long term strategy. However, this strategy will succeed on a large scale only if linguists from different institutions will pool forces and resources to build and maintain a `linguist's search engine" (Kilgarriff 2003, Fletcher 2004).