Abs-55

Eros Zanchetta (University of Bologna), Marco Baroni (University of Trento), and Silvia Bernardini (University of Bologna)

Corpora for the masses: the BootCaT front-end

This presentation introduces the BootCaT front-end, a graphical interface for the BootCaT toolkit (Baroni & Bernardini 2004).

The application implements an iterative procedure to bootstrap specialized corpora and terms from the web requiring only a list of seeds as input (a seed is a term that is expected to be typical of the domain of interest).

The front-end is a "wizard" that guides users through the BootCaT procedure allowing them to create a web corpus in a few minutes (theoretically, it should be possible to create a corpus in 6 minutes and 40 seconds).

Unlike other similar existing GUIs (such as JBootCaT, WeBoCa and WebBootCaT), BootCaT front-end is under active development (new features are added on a fairly regular basis) and is available for free. The application is cross-platform and runs on Windows, Mac and Linux.

New features currently being considered include: Unicode support, inclusion of non-HTML files (i.e. pdf, doc) in the corpus, exclusion of specific sites/domains, exclusion of documents that do not conform to specific licenses (i.e. Creative Commons).

References

M. Baroni and S. Bernardini (2004), "BootCaT: Bootstrapping corpora and terms from the web". Proceedings of LREC 2004.