Abs-264

Laurence Anthony (Waseda University, Japan), Kiyomi Chujo (Nihon University, Japan), and Kathryn Oghigian (Waseda University, Japan)

A freeware, open-source, web-based framework for distribution and analysis of single and parallel corpora

In recent years, an increasing number of corpora have been made available to corpus researchers. Many of these are released as raw texts that can work with standalone concordancers such as WordSmith Tools (Scott, 2010), ParaConc (Barlow, 2010), and AntConc (Anthony, 2010). However, more and more corpora are being released through dedicated Web-based interfaces. This trend is partly due to the convenience of an 'anytime-anywhere' environment and also due to the fast processing offered by server-based programs. A further and perhaps more important reason for the growth in Web-based tools is that they allow researchers to avoid restrictions associated with distributing copyrighted materials (Hemming and Lassi, 2010).

One major problem with releasing a corpus via a Web-based interface is the effort required to develop, test, and manage the server-side analysis tool. Each new project requires the corpus researcher to consider the corpus database architecture, the design of the interface, the development programming language, and ultimately the difficult task of coding itself. Some corpus projects are fortunate to have programmers and designers as part of the team, but this usually results in a long development time frame. More commonly, project members are forced to outsource the tool development to commercial software developers. In this case, the costs can become very high and the resulting tool inflexible.

In this paper, we introduce a freeware, open-source Web-based framework that allows corpus researchers to easily release their single or parallel corpora to the wider field with only minimal knowledge of servers, databases architectures, and programming languages. In essence, researchers download a setup file from a Website and drag and drop this into a folder on a standard hosting server. Then, after launching the setup file in a browser, a script proceeds to index the raw texts and setup the system for users. There are many advantages to such an open-source platform including: 1) corpus researchers no longer have to 'reinvent the wheel' creating a Web interface each time a new project is started, 2) the framework reduces development time and costs allowing researchers to focus their efforts on developing better corpora, 3) the framework gives complete control back to the corpus researchers, allowing them to tweak the system to their own needs, and 4) the corpus community can work together to improve the framework by adding different database architectures, attractive interface skins, and new functionality.

In the presentation, we will introduce a pilot version of the framework and show how it has already been used to create an effective environment for novice teachers and students at two university institutions, where they interact with single corpora of science and engineering texts, and parallel corpora of newspaper texts.

References

Anthony, L. (2010). AntConc (Version 3.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Barlow, M. (2010). MonoConc Pro (Version 2.2) [Computer Software]. Houston, US: Athelstan. Available from http://athel.com/

Hemming, C. and Lassi, M. (2010). Copyright and the Web as Corpus. Retrieved September 21, 2010, from http://hemming.se/gslt/copyrightHemmingLassi.pdf

Scott, M. (2010). WordSmith Tools (Version 5) [Computer Software]. Liverpool, UK: Lexical Analysis Software. Available from http://www.lexically.net/wordsmith/version5/index.html