Introducing "CorpusStudio"

Erwin R. Komen
Radboud University Nijmegen

Corpus research can be done with a variety of programs, which usually are command-line oriented. Corpus research is often done with a kind of "brute force" method, where researchers directly invoke queries using Window's command prompt. Such methods are error-prone, and can lead to unrepeatable or irretrievable results.

The program "CorpusStudio" provides a shell between the user and the search engines. It supports working with syntactically annotated Penn-Treebank corpora using the CorpusSearch2 engine (Randall et al., 2005). It also facilitates searching XML coded Treebank corpora using the Xquery language (Boag et al., 2010).

The basic unit of CorpusSearch is the *Corpus Research Project*, which is an XML file containing general information about the project (date, author, goal), all queries that are used, the order in which the queries are to be processed, and the locations of the input and output files. Since the *corpus research projects* comprise all the data needed to perform a particular task on a (selectable) set of input files, they offer many advantages over the brute-force method. They allow for **repeatable** corpus research, they are a vehicle in **teaching**, constitute a form to hand in corpus research **assignments** etc.

CorpusStudio contains its own editors for definitions and queries. The "construction editor" allows users to define hierarchical query execution, where output or complement of one query can serve as the input for another. The query execution results in an HTML file, which contains a table with the quantitative results, and it also contains all the results of the queries, with or without preceding text lines, following text lines, syntactic breakdown of the results and so forth.

References

Boag, Scott, Don Chamberlin, Mary F. Fernández, Daniela Florescu, Jonathan Robie, and Jérôme Siméon. 2010. *XQuery 1.0: An XML Query Language (Second Edition)*: W3C Recommendation, <http://www.w3.org/XML/Query/#specs>.

Randall, Beth, Ann Taylor, and Anthony Kroch. 2005. *CorpusSearch 2*, <http://corpussearch.sourceforge.net/credits.html>.