Corpus Browser: an intuitive and interactive corpus tool

XXX
XXX

In this software demonstration, we will present Corpus Browser, an interactive corpus exploration tool. Corpus Browser is an open source application programmed in Java that can run on Microsoft Windows, Linux/Unix, and Macintosh computers. It requires corpus data (such as from the BNC sampler or the MICASE) which have been preprocessed into a MySQL database.

First, the user loads such a pre-processed corpus into Corpus Browser, which provides a directory structure of the modes/registers/genres of the corpus in a window bar, from which the user can freely pick corpus parts to investigate. Second, when the user has defined corpus parts to investigate, a window on the bottom of the screen allows to define further subsections for analysis within the chosen corpus parts. Third, using Corpus Browser's menus and point-and-click interface, the user can perform many core corpus-linguistic functions, which can all be applied to three kinds of elements: words (e.g., "open"), tags (e.g., "AJ0"), and matches of regular expressions (e.g., "<w AV0>[A-Za-z]+ly <w AJ0>[^<]+?").

In the current stable development version, the following functions are already available:

(1) **frequency lists**: the frequencies (absolute and relative) of elements in a (part of the) corpus;
(2) **concordances**: the sentence context in which an element is used in a (part of the) corpus;
(3) **collocation displays**: the ten most frequent collocates of an element within a sentence (e.g., the collocates of the word "open") or the ten most frequent collocates of an element that match another element (e.g., the collocates of the word "open" that are nouns) with their frequencies (absolute and relative) in numbers and barplots;
(4) **dispersion**: a dispersion plot of words or tags within a (part of the) corpus plus a measure quantifying the degree of dispersion (*DP*).

Many of these functions allow to place text and graphics results to be compared next to each other for an immediate and intuitive comparisons of, say, near synonyms or words' dispersion in corpus parts.

The key features of Corpus Browser are how easy the exploration of corpus data becomes, how fast searches can be conducted and changed and their results can be displayed, and how well Corpus Browser can be extended to include many other corpora and formats of annotation and make them available for fast browsing and evaluation. While Corpus Browser can be applied to many different purposes, one we find particularly relevant is in the corpus-based second language teaching context. Generating frequency lists and concordances and identifying collocations often requires considerable time, and once the composition of the corpus (parts) is changed, everything has to be done again. Within Corpus Browser, changes regarding the search words, the composition of the corpus parts, and additional filtering options can usually be implemented with a few mouse clicks, and the corresponding changes in the output are available within a few seconds. Apart from general research applications, Corpus Browser is therefore an ideal tool for fast and dynamic corpus exploration in the intermediate to advanced second language learner classrooms.