WorldWideWeb-based interface to the Helsinki Corpus

Mark Catt
att@chass.utoronto.ca

The Centre for Computing in the Humanities and Social Sciences at the University of Toronto has developed a WorldWideWeb-based interface to the Helsinki Corpus (diachronic part). The interface is a powerful text-query utility that supports complex text searches. Named 'Patterweb/HC', the utility allows the user to restrict searches to texts matching one or more of the many features and properties that the compilers of the corpus encoded with the texts: date of composition, form (prose/verse), genre (letter, sermon, scientific handbook, etc) and so on. Moreover, the utility's query interface is flexible and powerful, supporting boolean and proximity searches as well as 'pattern-matching'. Patterweb/HC (http://www.chass.utoronto.ca:8080/english/helsinki/patterweb.html) formats results in a readable format, translating native COCOA-formatted character codes to HTML and augmenting query output with bibliographic references and a summary of textual features.

Patterweb/HC is based on the 'Patterweb' utility originally developed for Ian Lancashire's 'Early Modern English Dictionaries Database' (http://www.chass.utoronto.ca:8080/english/emed/emedd.html). Like its precursor, Patterweb/HC exploits the text-manipulation functions of two powerful software packages: OpenText Corporation's 'Pat' text search engine and the UNIX-based scripting language 'Perl'.

The design and development of Patterweb/EMEDD is described in Mark Catt's 'Patterweb: A User-Friendly Query/Display Program for Electronic Dictionary Databases', originally presented at the Learned Societies Congress (1996) and available on-line at

http://www.chass. utoronto.ca:8080/~mcatt/learneds/paper.html.

Owing to copyright restrictions, the corpus search functions are available only to University of Toronto users. However, the Patterweb/HC site is visible to all and provides a sample query output page that illustrates its capabilities.

122