Diachronic Internet Corpus of English (DICE): Fully Searchable Online Corpus of Historical Texts from the Net

Mark Kaunisto[1]

Abstract

In recent years, more and more attention has been paid to the Internet as a source of research materials. As regards historical studies of English, making use of Internet materials has largely been a "do-it-yourself" practice. Several archives of electronic texts are available, but selecting useful texts and downloading them can be a time-consuming process. Furthermore, there are a number of problems involved in examining historical texts on the Internet, e.g. insufficient bibliographic information about the source of the electronic edition, and the heterogeneity of the editorial practices behind the electronic versions. Assessing the "quality" of the texts found can be burdensome.

The Diachronic Internet Corpus of English (DICE), currently being compiled at the University of Tampere, aims to help scholars by integrating electronic texts suitable for linguistic study from different sources on the Internet. Drawing texts from a number of web sites – representing a variety of interests, including economics, medicine, psychology, religion, to name but a few – DICE includes texts from the sixteenth century to the twentieth century. So far the size of the corpus is over 10 million words. The DICE web site also has its own search engine, which enables word and phrase searches from the texts. The searches can be directed to a subset of all the texts included, e.g. according to the year of publication, sex of the author, place of publication, or genre.

In addition to a demo on the functional features of the DICE search engine, the paper will include a discussion of some of the practical problems involved in compiling the corpus (e.g. copyright issues and corpus structure), and the possibility of including the contents of some already existing archives in DICE in the future.

[1] *e-mail*: mark.kaunisto@uta.fi