Theoretical Issues for Corpus Linguistics Raised by the Study of Ancient Languages

Stanley E. Porter and Matthew Brook O'Donnell
University of Surrey Roehampton and OpenText.org

Corpus linguistics has focused largely upon the analysis of modern languages, usefully compiling large corpora for linguistic analysis. In the course of our study of ancient Greek from a corpus-based perspective, a number of issues regarding corpus linguistics have come to the fore. In this paper we wish to highlight and discuss three of these issues.

(1) ***Corpus size and compilation criteria*** The potential size of a corpus of a modern language, such as English, is virtually infinite, limited perhaps only by storage capacity and compiler effort. These limitations are less and less significant due to developments in technology and automated corpus building. However, the situation is very different for the study of a language such as ancient Greek, where the corpus size has been limited by historical accident. The number of texts is restricted to those that have for some reason been preserved from the ancient world. Early studies in corpus linguistics, faced with technological limitations, were forced to address issues of representativeness and sampling, similar to those facing compilers of a corpus of ancient texts. More recent corpus studies have tended to emphasize the size of the corpus over the contours of its composition. However, significant linguistic results, such as those from Svartvik's study of the English voice system, can result from a carefully compiled corpus of limited size.

(2) ***Annotation and levels of analysis*** In the light of the limited size of the available corpus of ancient Greek, it is necessary to include greater detail and levels of linguistic annotation in the corpus than is common in large modern language corpora. A limited corpus demands that as much information as possible be garnered from the available data. This involves a much more detailed analysis than the general lexical and morphological patterning that takes place in large-scale corpus studies. In order to facilitate this study, we have had to develop a consistent system of annotations of higher linguistic levels. This has not only generated much useful data, but forced us to examine a number of issues previously unaddressed by corpus annotators.

(3) ***Analysis of texts*** One of the strengths of corpus linguistics has been the observation of linguistic patterns across large samples of text. In contrast, the textual orientation of Classical and New Testament scholarship requires close attention to the linguistic features and function of individual documents as well as overall patterns in the language. This requires sensitivity to and the analysis of contextual and co-textual features and patterns. This micro-analysis is more suitable to this kind of limited corpus, and has provided useful data for studying various text-types and created new possibilities for the use of discourse analysis in corpus-based studies.