CHAPTER THREE
CORPUS LINGUISTICS:
LEXICOGRAPHICAL DATA COLLECTION,
THEORY, AND ANALYSIS

This chapter builds upon the theoretical foundation of the previous chapter, with specific interest in the application of corpus linguistic techniques designed to provide evidence for predictable patterns of language (particularly in Hellenistic Greek). Four major sections will form the content of the chapter: (1) an introduction to the general concerns and priorities of corpus linguistics, (2) the provision of the theoretical support for a statistically sound and representative corpus for Hellenistic Greek, (3) an engagement of quantitative and qualitative measures as a means for identifying lexicogrammatical constraints to lexical function, (4) and a description of the relation between these concerns and systemic functional monosemy.

INTRODUCING CORPUS LINGUISTICS:
GENERAL PRINCIPLES AND GOALS

Corpus linguistics is a fundamentally functional linguistic technique that employs digitally compiled and stored texts and enables the processing and analyzing of significant patterns in natural language. These selected texts form a corpus from which particular linguistic phenomena can be observed and described. While it may be better to call corpus linguistics a set of techniques rather than a theory,¹ several theoretical assumptions drive these techniques.

Corpus linguistics is functional insofar as it is concerned with language as a naturally occurring phenomenon, and suggests that the naturally occurring utterances should be the primary focus for any linguistic investigation.²

² See Porter and O’Donnell, “Representative Papyri,” 289. Given that this has been stated as a fundamental pillar of corpus research it will be important to maintain the notion
Functionalism in this sense has necessarily moved in the direction of corpus analysis as computers have increased their memory capacity, speed, and complexity to allow for the storage and analysis of immense quantities of texts and annotations. As corpus linguist John Sinclair suggests, “linguistics has been formed and shaped on inadequate evidence and in a famous phrase ‘degenerate data.’” Computer-aided corpus linguistics is an attempt to provide a critical mass of data evidence to allow for sophisticated linguistic description (which may, in fact, overturn formerly accepted principles and conclusions). A further dimension of functionalism is that systemic assumptions pervade corpus studies as typical patterns establish themselves as instantial representations of systemic choices. Corpus observations give hard evidence for oppositional choices that characterize language systems.

Part and parcel with a functional perspective is the view of language as a social phenomenon. Wolfgang Teubert, in an introductory essay on corpus principles, is very adamant about this point, and attempts to distance his work (namely, corpus linguistics) from more cognitive, Chomskyan

of functionalism in subsequent theorizing and especially application. Michaela Mahlberg suggests that due to the increasing size of corpora (thus an allegedly higher level of the reliability of conclusions), sensitivity to functionality in certain contexts of use has fallen by the wayside (Michaela Mahlberg, “Lexical Items in Discourse: Identifying Local Textual Functions of sustainable development,” in Michael Hoey, Michaela Mahlberg, Michael Stubbs and Wolfgang Teubert [eds.], Text, Discourse and Corpora: Theory and Analysis [Studies in Corpus and Discourse; London: Continuum, 2007] 191). This is a devastating observation and thus must be acted upon accordingly.

---


4 Michael Stubbs suggests that corpus linguistics makes two major contributions to linguistic study: (1) it provides “many new and surprising facts about language use,” and (2) it may “help to solve paradoxes that have plagued linguistics for at least a hundred years” (Michael Stubbs, “On Texts, Corpora, and Models of Language,” in Michael Hoey, Michaela Mahlberg, Michael Stubbs and Wolfgang Teubert [eds.], Text, Discourse and Corpora: Theory and Analysis [Studies in Corpus Linguistics; London: Continuum, 2007] 127).
