[*BAGL* 10 (2021) 17–47]

Synchronic Corpora and Ancient Languages: Theoretical Considerations for Designing a Corpus for Koine Greek

Nicholas List
University of Cambridge, Cambridge, UK

Abstract: Corpus linguistic research often necessitates large amounts of data (especially for Natural Language Processing tasks), yet this is exactly where ancient language corpora are most deficient. Many ancient language corpora increase their size by extending the temporal coverage of the corpus, allowing for diachronic analysis over an enlarged dataset. Because of this, less attention has been given to the compilation of synchronic corpora for ancient languages. Since temporal demarcation must be strictly controlled, other means of increasing corpus size must be explored. This paper considers a number of important theoretical considerations for the construction of a corpus for Koine Greek, including representativeness, size, and temporal coverage. While this study does present a corpus for Koine Greek, its primary aim is to foreground the particular theoretical challenges that face linguists engaged in synchronic corpus design for ancient languages. (Article)

Keywords: corpus design, corpus linguistics, Koine Greek, synchronic.

1. *Introduction*

Epigraphic (dead) languages pose their own peculiar challenges for corpus design. Compared to their modern counterparts, the literary evidence for ancient languages is in general rather meagre, subject to the vicissitudes of the historical record. Because of this, many corpus linguistic analyses of epigraphic languages like Ancient Greek or Latin have often been diachronic in focus, tracing semantic change of lexemes over a period of more than a thousand years.[1] Less attention, however, has been given to the challenges of building *synchronic* corpora for epigraphic languages, that is to say, corpora that aim to capture a representative sample of a language at a particular point in time.

This study highlights a number of important theoretical issues that relate to the design of a corpus for Koine Greek. While all corpus projects must grapple with key corpus issues such as representativeness and size, the particularities of constructing a corpus for an ancient language pose their own theoretical challenges. I take up these issues in section 3 of this paper. In section 4, I detail the key specifications for designing a synchronic corpus. While this study does present a corpus for Koine Greek (4.3 and 6), the primary aim of this study is to foreground the particular theoretical challenges that face linguists engaged in synchronic corpus design for ancient languages.

2. *What is Koine Greek?*

Before dealing directly with these theoretical issues, we must first clearly define the population that that corpus intends to represent.[2] That is to say, we must begin by delineating the chronological boundaries of Koine Greek. The term “Koine” was used by the ancients to refer to “the common dialect” (ἡ κοινὴ διάλεκτος) spoken throughout the Roman empire.[3] As a lingua franca, Koine has at times been characterized as simply a trade language; a vulgar devolution from the literary elegance of the Attic and Ionic of a previous era.[4] Hence the designation

[1] Cf. Mambrini, “Nominal vs Copular Clauses”; Perrone et al., “GASC”; Rodda et al., “Panta rei”; Toufexis, “One Era’s Nonsense.”

[2] Biber, “Representativeness,” 243–4.

[3] Colvin, “Koine Dialect,” 3796.

[4] Colwell (“The Greek Language,” 480) divides what he calls “Hellenistic Greek” into two categories: “literary” and “Koine.” Similarly, Köstenberger et al. (*Going Deeper*, 21) consider “vulgar Greek” “interchangeable” with the term Koine. This notion probably reflects some of the ancients’ own opinions of the common dialect, seen in the reactionary

"Koine" can be a cause of confusion when analyzing the language in terms of register; the idea of a "literary koine" may appear to some as a contradiction in terms, since "common" suggests a lowest-common-denominator approach to language description. Of course, the Greek of the Hellenistic and Roman periods is well attested in the higher literary registers, Polybius and Plutarch being two unambiguous examples.[5] This means that "Koine" cannot be solely comprised of sub- or non-literary registers, and instead encompasses the full spectrum of spoken and written registers.[6] The negative connotation of Koine Greek as "vulgar Greek" is the result of a misunderstanding of what constitutes a "koine" (in its sociolinguistic sense).

The development of Koine Greek has been well documented in a number of places.[7] The prehistory of Koine begins with the unification of the Greek peninsula under Philip II of Macedon (with the decisive victory at Chaeronea in 338 BCE and the subsequent conference at Corinth the following year). When Philip was assassinated in 336, the plans to expand East were taken up by his son, Alexander the Great. The next decade of conquests saw the fall of the Persian Empire, and by Alexander's death in 323, the military commander had secured a domain that stretched from the Adriatic in the West to Punjab in the East, modern-day Tajikistan in the North to Libya in the South.[8] Of course, the sheer size of the realm could never have been

movement of the Second Sophistic, which "endeavored to check the further progress of this 'Common' (i.e., unclassical Attic) Greek and revive the ancient pure Attic" (Jannaris, *An Historical Greek Grammar*, 7). Jannaris ("The True Meaning of the Κοινή," 96) has, however, shown that the ancients never used the term "Koine" in the sense of "vulgar," only using it to refer to "that national literary dialect which is free from all dialectal and even poetical admixture, a form of style best represented in the orators" (similarly Browning, "Language of Byzantine Literature," 106).

[5] Horrocks, *Greek*, 96.

[6] Cf. Adrados, *A History of the Greek Language*, 175–202.

[7] Brixhe, "Linguistic Diversity," 228–52; Bubenik, "The Rise of Koine," 342–45; Bubenik, "Koine, Origins of," 217–85; Colvin, "Koine Dialect," 3796–8; Horrocks, *Greek*, 79–123; Porter, "The Greek Language," 99–130.

[8] Missiou, "The Hellenistic Period," 325.

sustained, and Alexander's generals fought to secure the various regions of the recently conquered territory for themselves, a process that stabilized by 277/6 BCE, forming the city-states of the Hellenistic Period.

Alongside this geographic expansion and political upheaval came changes to the linguistic landscape of the Mediterranean. There is evidence for a number of Greek dialects in use around the Greek islands and the mainland from the eighth to fourth centuries BCE, the major categories including Attic-Ionic, Aeolic, Arcado-Cypriot, West Greek (Doric), Northwest Greek, and Pamphylian.[9] A form of the Attic dialect was formally adopted as the administrative language of the Macedonian Empire under Philip II (often referred to as "Great Attic").[10] As Geoffrey Horrocks explains, this would prove to be most decisive in securing the place of the Greek language in the subsequent Macedonian conquests:

> This Atticization of the Macedonian aristocracy was to be the crucial factor in the future history of the Greek language, since, continued Athenian cultural prestige notwithstanding, the emergence of Great Attic as a true national language (the Koine) would surely have been long delayed, or even prevented altogether, without the substitution of the military and political power of Macedonia for the declining influence of Athens.[11]

The widespread conquests of Alexander established the Greek of the Macedonian aristocracy as the language of communication and administration, which came to supplant many of the local languages.

> This pattern of domination in large part through linguistic unification was a pattern developed early and continued after the death of Alexander—the four succeeding Hellenistic kingdoms, including the Ptolemies and the Seleucids, and later the Romans, continued the

[9] Colvin, *A Brief History*, 98–109.

[10] Coined by Albert Thumb, "Großattisch" (cf. Silk, *Standard Languages*, 20n70).

[11] Horrocks, *Greek*, 80.

> same patterns. Greek thus became the lingua franca, or the language of common interaction, throughout the Greco-Roman world.[12]

The koine that began to emerge in the Hellenistic Period was an admixture of Attic and Ionic, though some Doric elements have also been detected by Bubenik.[13] Typical of all koines, the emergent dialect exhibited both reduced and more regularized features. Stanley Porter lists a number of reduced features, including a reduced specificity of a number of prepositions, the loss of prominence of the middle voice (and its replacement with the passive), the near-total disappearance of the optative mood and dual number (the dual form being restricted to specific vocabulary). In contrast, a higher degree of regularity can be seen in pronunciation and orthographic conventions (e.g. more consistent use of final *nu*).[14] These features that would come to characterize the emerging koine emerged slowly. Jeff Siegel articulates a four-stage progression in the formation of Koine Greek.[15]

Siegel refers to "the unstabilized stage at the beginning of koineization" as pre-koine.[16] In the process of forming a compromise dialect, many of the distinctive or irregular Attic and Ionic features continue to be used "concurrently and inconsistently." This best typifies the form of the Greek language from the sixth to the beginning of the fourth century BCE.

With the greater political cohesion among the Greek city-states as a result of the First Maritime League (477 BCE onwards, from which Athens secured itself as the ascendant

[12] Porter, "Septuagint," 428.

[13] Bubenik, "Dialect Contact," 13.

[14] Porter, "Septuagint," 439–40.

[15] The four stages were popularized by Vit Bubenik ("Dialect Contact" and "The Rise of Koine") and are often traced back to him (e.g. Moţ, *Morphological and Syntactical Irregularities*, 231), although they were actually formulated by Siegel ("Koines and Koineization," 373–5), who in turn based his periodization on Mühlhäusler's ("Development of the Category of Number," 37) developmental phases of pidginization: (1) jargon, (2) stable pidgin, (3) expanded pidgin, (4) creole.

[16] Siegel, "Koines and Koineization," 373.

hegemony), the Attic-Ionic koine began to stabilize.[17] This is the result of lexical, phonological, and morphological levelling, in which a distinct dialect can clearly be discerned.

While the stabilizing process of koineization produces a reduced, simpler dialect, a koine can undergo a further stage of development as its influence is extended geographically (becoming the primary language of intergroup communication) and socially (becoming a literary language). This can actually result in linguistic expansion, for example, "in greater morphological complexity and stylistic options."[18] Koine Greek underwent this stage of geographical expansion during the conquests of Alexander (which ended in the 320s BCE), though the development of Koine Greek as a literary language in these extended regions was naturally a slower process.

Eventually the new dialect may become the first language (L1) of a linguistic community. Again, "this stage may also be characterized by further linguistic expansion (or elaboration), but here some of it may be the result of innovations which cannot be traced back to the original koineized varieties."[19] Vit Bubenik points to both nativized speakers, like Polybius, and Hellenized Semites, like Josephus or the New Testament authors, who embody a nativized koine.[20]

2.1 *The Rise of Atticism*

In seeking to collate Greek texts from the Greek and Roman periods, one key factor that must be considered is the rise of the Second Sophistic. The Second Sophistic was a literary movement that sought to revive those features of the Greek language that characterized the Attic dialect of Classical Athens.[21] The movement reached its height in the second century CE, and examples of Atticism can be found throughout a variety

[17] Bubenik, "Dialect Contact," 11.

[18] Siegel, "Koines and Koineization," 374.

[19] Siegel, "Koines and koineization," 374.

[20] Bubenik, "The Rise of Koine," 345.

[21] Cf. Kim, "The Literary Heritage," 468–82.

of genres.[22] Horrocks details a number of Atticistic features distinct from general Koine, including among others:[23]

- A return in part to patterns of Atticistic orthography (e.g. θάλαττα instead of θάλασσα)
- Use of the dual number form
- Use of preposition ξύν instead of σύν
- Retention of the longer form γίγνομαι instead of the Koine γίνομαι
- Use of the optative
- Extensive use of middle forms of verbs

In many ways, the presence of Atticistic writers within the synchronic corpus for Koine Greek should not be a cause for concern. Robert Browning notes that the early Atticists were concerned “primarily with style, and only secondarily with language,” wishing to avoid the “rambling” paratactic form of Hellenistic prose on the one hand, and the “over-ornate” style of the Asianic school on the other.[24] Atticism does not constitute its own dialect distinct from Koine—it is simply a higher register among other registers of the Koine period. Thus, authors that tend towards a more Atticistic style have not been excluded from this corpus (following the corpus design decisions of Todd Price and Matthew Brook O’Donnell).[25]

[22] Horrocks (*Greek*, 136) highlights authors composing works of oratory (Aelius Aristides, 117–180 CE and Herodes Atticus, 101–177 CE), philosophy (Claudius Aelianus, 172–c. 235 CE), historiography (Flavius Arrianus, c. 95–175 CE), biography (Philostratus, second century), geography (Pausanias, second century), and romantic fiction (Achilles Tatius, c. second century, and Longus, second to third century).

[23] Horrocks, “Greek in the Hellenistic World,” 82.

[24] Browning, “The Language of Byzantine Literature,” 106. Porter (“Septuagint,” 429) writes, “in the third century B.C., there was the rise of what is called Asianism. This was a reaction against the balanced and measured style of the literary form of Hellenistic Greek, such as found in Polybius, and so these writers indulged in a more exuberant and ornate style.”

[25] Price, *Structural Lexicology*; O’Donnell, *Corpus Linguistics*.

2.2 *Dating Koine Greek*

It is difficult to say with precision when the Koine period came to an end. Historically speaking, the Byzantine Period existed between 330 CE (the founding of Constantinople) and 1453 CE (the fall of the Byzantine Empire). Most scholars tend to adopt the former historical moment in the first half of the fourth century as a heuristic transition marking the terminus of the Koine period.[26] Chrys Caragounis traces Byzantine-Mediaeval Greek to 600–1500 CE, in which the Greek language begins to resemble a proto-Modern Greek form.[27] Browning places early Byzantine Greek in the fifth to seventh century CE.[28] Colvin notes that

> it is difficult to assign a convenient end date to the koine . . . [F]orms of koine remained the spoken language of the Greek world until the end of Late Antiquity, and also remained the written language for the purposes of epigraphy and other day-to-day documents.[29]

Colvin continues by observing that "texts later than Justinian (d. 565 CE) are rarely quoted to illustrate koine (as opposed to Byzantine) Greek," although "many later Byzantine literary texts are written in an archaizing idiom."[30]

In terms of demarcating the Koine period, some adopt a maximalist approach. Caragounis places Koine Greek between 300 BCE–300 CE, which is a subset of a wider transitional period he calls "post-classical" (300 BCE–600 CE).[31] Porter and O'Donnell also adopt fairly broad parameters (fourth century BCE to fourth century CE).[32] For the task of designing a synchronic corpus, however, these temporal designations are a little too broad, since they encompass the unstable stages of

[26] Jannaris, *An Historical Greek Grammar*, 5; Price, *Structural Lexicology*, 35; Robins, *The Byzantine Grammarians*, 1; Wallace, *Greek Grammar*, 12–30.

[27] Caragounis, *The Development of Greek*, 45.

[28] Browning, "The Language of Byzantine Literature," 109.

[29] Colvin, "Koine Dialect," 3798.

[30] Colvin, "Koine Dialect," 3798.

[31] Caragounis, *The Development of Greek*, 22 and 39.

[32] Porter, "The Greek Language," 99; O'Donnell, "Designing and Compiling," 263.

koineization of the fourth century BCE, as well as early stages of Byzantine Greek. To avoid these transitional phases, the Koine Greek corpus will focus on Greek texts that could generally be considered as belonging to Siegel and Bubenik's "nativized koine."[33] While neither Siegel nor Bubenik states a firm date of this developmental stage of Koine Greek, the latter does point to the Greek historian Polybius (c. 200-118 BCE) as a typical example of nativized koine.[34] Designating a terminus for Koine Greek is slightly more arbitrary, but again, the aim is to avoid moving too far into the unstable phase of pre-Byzantine Greek development. Price limits the temporal coverage of his corpus to 200 BCE–200 CE.[35] Taking a similar minimalist approach, this corpus will cover a period from 250 BCE–250 CE. While this may exclude some texts that would otherwise be considered standard Koine, the timeframe is long enough to adequately capture the majority of what can be considered nativized koine, while remaining short enough to be considered synchronic (see 4.2).

3. *Issues in Ancient Language Corpus Design*

Having defined the temporal parameters of the prospective corpus, we can now turn to consider the primary issues in corpus design. This section discusses representativeness and corpus size in relation to the study of Ancient Greek.

3.1 *Representativeness in Corpus Design*

We begin with the issue of representativeness or balance. The value of any corpus is in the generalizability of the data—the ability to extrapolate from the corpus to the language in itself. Claims about a language or language domain are only as good as the language sample upon which they are based. A number of studies have sought to establish criteria by which the

[33] Siegel, "Koines and Koineization," 374; Bubenik, "The Rise of Koine," 345.

[34] Bubenik, "Dialect Contact," 21.

[35] Price, *Structural Lexicology*, 35.

representativeness of a corpus could be measured. Corpus design must begin with preliminary considerations of what kind of corpus is needed for the given research purposes. Stefan Th. Gries delineates four basic aspects of a corpus:[36]

General or specific. A general corpus attempts to create a sample representative of the language in general, while a specific corpus is restricted to a particular variety, register, genre, or other domain.

Diachronic or synchronic. A diachronic corpus will have broad enough temporal parameters to capture language change over time. A synchronic corpus aims at representing a snapshot of language at a particular moment or within a delimited period of time.

Monolingual or parallel corpora. While the former is only interested in one language, the latter attempts to compile the same or similar information across a number of languages for purposes of comparison (see, for example, the Canadian Hansard corpus, which consists of parallel texts in English and Canadian French).

Static corpora or dynamic/monitor corpora. Static corpora are of a fixed size. The Brown Corpus is a classic example of a static corpus; the corpus contains 500 language samples across 15 genres, each sample containing 2000 words.[37] More frequently, this kind of corpus is known as a sample corpus.[38] A monitor corpus by comparison aims at capturing as much of a language as possible, continually growing with the collation of new texts. Sample corpora can be created from a monitor corpus for specific research needs.[39]

One further qualification relevant to ancient language corpus design is provided by Sue Atkins, Jeremy Clear, and Nicholas Ostler:[40]

[36] Gries, “What is Corpus Linguistics?,” 1232–3.

[37] Kučera and Francis, *Frequency Analysis of English Usage*.

[38] Sinclair, *Corpus*, 23.

[39] Sinclair, *Corpus*, 26.

[40] Atkins et al., “Corpus Design Criteria,” 5.

Reception or production. A corpus can sample language that people read and hear (reception) or language that people speak and write (production). While the latter is often advocated for modern corpora, only the former is an available option for epigraphic languages.[41] Porter and O'Donnell suggest that there may be some benefit to a reception-focused corpus, in that "what we know of ancient scribal practice indicates that those texts that we still have today are those that were thought by the ancients to have particular religious, literary, cultural, historical or pedagogical value,"[42] and thus be representative of a particular stratum of language within that society. Since ancient Greek linguists are primarily interested in better understanding the language we do have (extant written texts), corpus linguistics can still yield legitimate results for epigraphic languages by focusing on reception.

Douglas Biber has advocated that a corpus should contain a diversified range of different genres and text types.[43] Given that there are "marked linguistic differences across registers,"[44] a representative sample must account for this range of variability within a population. This is sound advice for constructing a general corpus. However, the exact opposite seems to be the case for specific corpora. Todd Price and Francis Pang have both followed O'Donnell in employing a delimited, register-specific corpus in order to study the Greek of the New Testament (a subset of Koine Greek).[45] These studies still make use of a range of registers, but the selection of these registers is based on their attestation within the New Testament documents.[46] The key issue

[41] O'Donnell, "Designing and Compiling," 256.

[42] Porter and O'Donnell, "Theoretical Issues for Corpus Linguistics," 124.

[43] Cf. Biber, "Representativeness"; Biber, "Using Register-Diversified Corpora."

[44] Biber, "Using Register-Diversified Corpora," 220.

[45] Price, *Structural Lexicology*; and Pang, "Annotated Representative Corpus"; following O'Donnell, *Corpus Linguistics*.

[46] The key initial difference between a general versus specific corpus is slightly obscured by O'Donnell, whose study is entitled "Designing and Compiling a Register-Balanced Corpus of Hellenistic Greek for the Purpose of

for representativeness is defining the population that the sample intends to represent.[47]

A number of studies have sought to formulate criteria for representativeness in corpus design.[48] Marc Kupietz lists a number of common criteria:[49]

- Mode (spoken or written text)
- Place (of a text's publication or author's residence)
- Genre
- Topic domain (religious, political, historiographical, etc.)
- Audience (the primary recipients of the text)
- Time (when the text was originally produced)
- Register (literary, non-literary, etc.)

Some of these criteria are of greater or lesser importance to the study of Ancient Greek, and O'Donnell singles out genre and register as the most relevant.[50]

Yet the concept of representation is far more elusive than many are seemingly willing to admit. Despite the proliferation of the criteria that are meant to ensure representativeness,[51] there is

Linguistic Description and Investigation." This presents O'Donnell's target as a description of Hellenistic Greek, and one might assume this would require a general corpus. However, O'Donnell ("Designing and Compiling," 286) states that "Appendix A contains an outline of the texts that make up a small (596,049 words) corpus of Hellenistic Greek. It is intended initially to be a resource for the study of the Greek of the New Testament and then to serve as the basis for a larger—and more representative—corpus of Hellenistic Greek." This is confusing. Why create "a small corpus of Hellenistic Greek" when he admits that a much "larger" and "more representative" corpus is in fact needed? And in order to create such a corpus, the corpus that has been created must fundamentally change from specific to general, so why even use the general synchronic category "Hellenistic Greek" at all?

[47] Biber, "Representativeness," 243.

[48] Cf., inter alia, Atkins et al., "Corpus Design Criteria"; Biber, "Representativeness."

[49] Kupietz, "Constructing a Corpus," 65.

[50] O'Donnell, "Designing and Compiling," 274–80.

[51] Cf., inter alia, Atkins et al., "Corpus Design Criteria"; Biber, "Methodological Issues"; Biber, "Representativeness"; Biber "Using Register-diversified Corpora"; O'Donnell, "Designing and Compiling"; Gries,

an inherent circularity to the endeavour, in that "the object that we need to define is exactly the unknown object that we would like to investigate."[52] To know what constitutes a representative sample requires an objective knowledge of the language population itself, and since such knowledge is predicated upon generalizations made from the corpus sample, it cannot serve as the basis for selecting the sample. We should then leave behind the language of "criteria" when speaking of constructing a representative corpus. Rather, any decision to shape the corpus sample should be understood only as an approximation of representativeness. Importantly, these approximations are revisable. Corpus creation should be an iterative endeavour, as Kupietz writes: "start with a rough approximation of representativeness, act as if we had a perfect relation between sample and population, see what we can find out, check if the findings are justified, and, if possible, close the circle by using the findings to improve the corpus with respect to its representativeness."[53]

Despite the theoretical issues bound up with representativeness, one must still decide (initially) what will or will not make it into the corpus. In terms of methodology, Charles Meyer delineates two distinct types of sampling: probability sampling and non-probability sampling.[54] In probability sampling, the sample population is carefully selected by means of statistical formulas and demographic information, as a control for representativeness. Non-probability sampling is a form of "haphazard, convenience, or accidental sampling,"[55] in which the

"Dispersions and Adjusted Frequencies"; Gut and Voormann, *Corpus Design*.

[52] Kupietz, "Constructing a Corpus," 64.

[53] Kupietz, "Constructing a Corpus," 64. This is a further critique of O'Donnell ("Designing and Compiling") and those who have based their own corpus investigations on his work. There is no system of iterative revision in O'Donnell—once the criteria of genre and register are met, the work of corpus design and compilation are seemingly finished. That said, since this iterative process is dependent on specific tasks or research questions, this corpus presents a starting point for further research.

[54] Meyer, *English Corpus Linguistics*, 43–5.

[55] Kalton, *Introduction to Survey Sampling*, 90.

availability of the material is the main requisite for its inclusion in the corpus. Due to the limitations of working with Koine Greek, non-probability sampling is a more viable option. First of all, constructing a corpus for an epigraphic language is by nature accidental; the only texts that are eligible for sampling are those which remain extant. This obviously limits the language population to only written modes of communication, though practically, the language population is even more restricted, since texts generally need to be morphologically annotated, electronically stored, and open access in order to be easily integrated into a corpus. While this haphazard reception of epigraphic texts excludes the luxuries of "rigorously defined sample procedures,"[56] there are a number of benefits to non-probability sampling. Corpora that employ probability sampling procedures must use words or sentences rather than whole texts as the unit of measurement for creating evenly distributed samples. The British National Corpus created 2000-word samples from various predefined categories (modality, domain, medium). As Kupietz rightly notes,

> A downside of a strictly balanced design approach is that texts must be discarded if their inclusion in the corpus would have the consequence that one of the fixed quotas is exceeded—even if the texts come for free. If you were compiling a corpus and the Guardian newspaper offered you its whole archive for free, would it be not a pity if you had to answer "No thanks, we already have enough newspaper texts, but do you happen to have personal letters?"[57]

There is also no clear benefit to using delimited samples over full texts. "The use of samples of constant size gains only a spurious air of scientific method, since it confers no benefit on the corpus, and is as practical as Genghis Khan's fabled policy of having all his soldiers the same height."[58] In corpus linguistics, there has been a general move away from words and sentences to gathering whole documents. Linguistic features are not evenly distributed throughout any given text, and thus corpora that

[56] O'Donnell, "Designing and Compiling," 262.

[57] Kupietz, "Constructing a Corpus," 66.

[58] Sinclair, "Corpus Typology," 28.

contain full-length texts are more linguistically diverse and hence more useful for greater range of linguistic enquiry.[59] Furthermore, depending on the specifics of one's research question, smaller corpora can be drawn from the larger corpus as the need arises. This is pragmatically effective, since the effort involved in creating a smaller corpus from a larger one is far less than creating a specialized corpus *de novo*.

A clear example of this kind of corpus is the German Reference Corpus *DeReKo*.[60] The corpus contains 3.4 billion words from various text types, and continues to grow at a rate of around 300 million words a year. While some would refer to such a collection as an archive rather than a corpus,[61] the *DeReKo* creators refer to it as a "primordial sample" (*Ur-Stichprobe*), from which smaller, specialized samples—or "virtual corpora" (*virtuelle Korpora*)—can be prepared.[62] This is essentially the same as John Sinclair's "monitor" or Gries's "dynamic" corpus.[63] While any epigraphic corpus cannot reach the kind of coverage found in monitor corpora for modern languages, the benefits of these kinds of non-probability samples remain consistent across varying corpus sizes.[64] Kupietz summarizes the benefits well:

> In general, and from an economic point of view, such an approach allows for a better exploitation of the available corpus data, as no texts need to be discarded and the corpus data are reusable for a range of different purposes that would otherwise require the creation of new corpora from scratch.[65]

[59] Sinclair, *Corpus*, 18.

[60] Cf. Kupietz and Keibel, "Mannheim German Reference Corpus"; Kupietz et al., "The German Reference Corpus."

[61] Atkins et al., "Corpus Design Criteria," 1.

[62] Kupietz, "Constructing a Corpus," 66.

[63] Sinclair, *Corpus*, 23–4; Gries, "What is Corpus Linguistics," 1233.

[64] That is to say, the benefits of a non-probability monitor corpus extend to the epigraphic Diorisis Ancient Greek Corpus (10.2 million words, on which see below) as much as to the *DeReKo* for modern German (3.4 billion words). It is not to say that size is unimportant, as I will make clear in the next section.

[65] Kupietz, "Constructing a Corpus," 66.

3.2 *Corpus Size*

Another key aspect that must be considered when constructing a corpus is size. As I have noted above, the constraints of working with an epigraphic language are most apparent here. Theoretically, a corpus needs a certain critical mass in order to be useful in terms of the generalizability of its findings. However, it is not possible to specify how big a corpus must be for it to produce useful results. Certain kinds of research questions will necessitate more or less data for effective analysis.[66] Most languages contain no large corpora, yet corpus analysis must carry on nonetheless, making the best with the data at hand.[67] There is a general impression among corpus linguists, however, that "more data is better data,"[68] and that "while balance is desirable, size is even more desirable."[69] Larger corpora are more likely to contain evidence for word usages or syntactic structures that may not appear in a more limited sample size. Or to put it in opposite terms, if rigorous methodological balancing ultimately empties our corpus of eligible texts, then the sample will ultimately fail to achieve the level of representativeness that such methodology sought to ensure in the first place. This serves as another reason to opt for a non-probability sampling method, since limiting one's text sample to a 2000-word (or a similarly arbitrary) selection places severe limitations on corpus size. This is not necessarily a problem for data-rich languages like modern English or German, but for epigraphic languages like Ancient Greek, placing our own artificial limits on a language population already diminished by

[66] For a Natural Language Processing example, Clérice and Munson ("Qualitative Analysis," 89 and 101) compare semantic similarity scores produced from Bullinaria and Levy's ("Extracting Semantic Representations") log-likelihood model to their Word2Vec model. They note the log-likelihood model produced more reliable results, yet the primary reason for this was their corpus size, which was limited to the Greek New Testament. The vector space model Word2Vec requires a much larger data input than Clérice and Munson provide, so it is unsurprising their model produced unreliable results.

[67] Gries, "What is Corpus Linguistics?," 1237–8.

[68] Church and Mercer, "Introduction," 18.

[69] Church, "Speech and Language Processing," 2.

the "vagaries of climate" and the "whims of ancient librarians"[70] in the preservation and transmission of texts seems an unusual way to maximize the range of variability within a corpus.

Sarah Hunston notes that most modern corpora are at least one million words in size.[71] This is certainly an achievable goal for the study of Ancient Greek. However, most work operating within the scope of the Hellenistic and Roman periods has worked well under this threshold. A number of studies solely make use of the Greek New Testament as their corpus,[72] which contains only 27 texts made up of just over 138,000 tokens (total number of words in a corpus) and 5437 word-types (unique tokens).[73] O'Donnell's proposed corpus for Koine Greek contains only 48 texts of 596,049 tokens,[74] while Pang's corpus (which is modelled on O'Donnell's) is 15 percent shorter with 521,226 tokens from just over 50 documents.[75] O'Donnell and Porter have annotated 45 papyri texts (3341 tokens), though none of this work seems to be open access.[76] Price employs a primary and secondary corpus in his study, the former consisting of 177 texts containing 1,740,830 tokens, while the latter has 161 texts with 2,298,673 tokens.[77] Ryder Wishart has utilized a corpus of approximately 7.6 million Greek words (31,000 unique tokens) for training the vector space model Word2Vec.[78] Price and

[70] Porter and O'Donnell, "Theoretical Issues for Corpus Linguistics," 201.

[71] Hunston, "Corpus Linguistics," 234.

[72] E.g. Clérice and Munson, "Qualitative Analysis."

[73] Based on Mounce, *Greek Grammar*, 31.

[74] O'Donnell, "Designing and Compiling," 294–5.

[75] Pang, "Annotated Representative Corpus," 286–7 and 166–71.

[76] Porter and O'Donnell, "Building and Examining Linguistic Phenomena."

[77] Price, *Structural Lexicology*, 44–9.

[78] Wishart, "Lexical Field Theory," 407. Wishart's corpus covers a period from 300 BCE to 300 CE. The texts used in his corpus are available at: https://github.com/gcelano/LemmatizedAncientGreekXML. While Wishart's lemmatized corpus is suited to the needs of corpus linguists, there are a few reasons to prefer the adoption of the Diorisis corpus introduced at the end of this section. The Diorisis project has an international platform, being developed at the Alan Turing Institute at the University of Cambridge, and has the

Wishart's corpora are certainly the largest to date for corpus linguistic work with a special interest in the New Testament, and the only ones to surpass Hunston's minimal million-word suggestion. Only Wishart's corpus can be deemed of sufficient size for most Natural Language Processing (NLP) tasks, though admittedly most of these other corpora were designed only for the particular research purposes of their original compilers (which did not include NLP). For many of these corpora it is also not clear how tokens have been counted (punctuation can count as a token), and many words are simply unimportant for the study of lexical semantics. Lexical items such as articles, pronouns, and prepositions (often referred to as stopwords) amount to little more than noise in the system and should be filtered out during the preprocessing stage of corpus compilation.[79] None of the above studies gives any indication that stopwords have been removed from their corpora, meaning a significant proportion of each corpus makes no contribution, semantically speaking, to lexicographical questions the corpus was created to investigate. The most common word in any Greek corpus is normally the definite article; following Zipf's Law, which states that a lexeme's frequency in a corpus is inversely proportional to its rank, the article will be attested twice as frequently as the next most common word.[80] If a stoplist were applied to the New Testament corpus, the token count would drop from 138,925 to 64,018, a 46 percent decrease.[81]

Efforts outside of the particular interests of biblical linguistics have not necessarily fared much better. The Ancient Greek Dependency Treebank by the Perseus Project currently has only 33 texts with 557,922 tokens (punctuation marks are included as tokens here),[82] while the PROIEL Treebank only has annotations

potential to be a standard monitor corpus for Koine Greek. At a practical level, the XML formatting for Diorisis is clearer than that of Wishart's corpus, which may prove to be a barrier for other scholars (especially within the field of biblical studies) who are interested in corpus-based research.

[79] On stoplists, see Burns, "Constructing Stoplists."

[80] Rydberg-Cox, "Co-occurrence Patterns," 124.

[81] On the stoplist used for this corpus, see footnote 101 below.

[82] "The Ancient Greek and Latin Dependency Treebank," [n.d.]

for Herodotus's *Histories* and the New Testament.[83] One Greek database that does not suffer deficiencies in size is the *Thesaurus Linguae Graecae* (TLG), established by the University of California Irvine in 1972.[84] The collection boasts nearly 10,000 texts, consisting of over 75 million words, from Homer through to the fall of Byzantium in 1453.[85] TLG has proved to be an important tool for certain tasks performed by lexicographers, classicists, and biblical scholars, such as specialized word studies or locating initial attestations of rare words and forms.[86] Unfortunately, TLG has severe limitations in its functionality that have deemed it inadequate for a number of studies that employ the most rudimentary corpus methods.[87] Search functions only produce strings of sentences, not centered, sortable key-word-in-context (KWIC) lines, and the encoding conventions developed in the 1970s are outdated. Maria Pantelia recognizes that adopting new encoding standards (TEI-XML and Unicode) "will significantly improve content-based search queries and will provide a well defined [sic], stable standard for the long-term maintenance of the digital collection," allowing for the possibility of "compatibility with other digital library projects, and allow[ing] the use of cross-platform, cross-library software."[88] For now, the limited functionality of the TLG interface, as well as issues surrounding copyright, prevents this present study from making use of the large TLG database in the construction of a Koine Greek corpus.[89]

Thankfully, the recent collation of the Diorisis Ancient Greek Corpus by The Alan Turing Institute at the University of

[83] The treebank's Latin library is slightly better represented. See http://dev.syntacticus.org/proiel.html.

[84] "Thesaurus Linguae Graece," [n.d.]

[85] Pantelia, "Noûs, into Chaos," 1.

[86] Adrados and Somolinos, "The 'TLG' Data Bank," 246.

[87] Pang, "Annotated Representative Corpus," 168; Porter, *Linguistic Analysis*, 29–46; Price, *Structural Lexicology*, 44.

[88] Pantelia, "Noûs, into Chaos," 7.

[89] Unlike the Perseus Project, the TLG does not operate under a creative commons. For a discussion on the issue of copyright and Ancient Greek texts, see Porter, *Linguistic Analysis*, 17–28.

Cambridge has made significant forward progress in Greek corpus design.[90] Diorisis is a diachronic corpus of 820 texts, containing 10.2 million lemmatized and part-of-speech-tagged words, spanning a time period from Homeric Greek through to the fifth century CE.[91] The corpus was designed specifically for machine learning models that track semantic change in Ancient Greek and therefore requires large quantities of data. Several studies have employed the Diorisis corpus for various NLP tasks,[92] made possible both by the size of the corpus and its TEI-compliant markup format.[93]

4. *Designing a Corpus for Koine Greek*

This section describes the design of the corpus created for this study. The corpus is intended to represent a non-probability general corpus of the reception of Koine Greek.

4.1 *A Sub-sample (virtuelle Korpus) of the Diorisis Corpus*

It is always more practical to re-use or improve upon an existing corpus than create a new one entirely from scratch.[94] Of course, practicality should not come at the cost of sacrificing the metadata, annotations, and corpus size necessary to properly investigate the research questions at hand. While the majority of Ancient Greek corpora fail to meet the minimum standards of size (Perseus, PROIEL treebank) and functionality (TLG), the Diorisis corpus has made NLP tasks achievable for Ancient Greek.

The corpus prepared for this study can be described in Kupietz's terms as a *virtuelle Korpus*, drawn from the Diorisis

[90] "Figshare," [n.d.]

[91] Vatri and McGillivray, "The Diorisis Ancient Greek Corpus."

[92] Cf. Rodda et al., "Vector Space Models"; Perrone et al., "GASC"; McGillivray et al., "A Computational Approach."

[93] Text Coding Initiative (TEI) is a consortium that develops and maintains standards for machine-readable texts, mainly for use within the humanities and social sciences (http://tei.org).

[94] Kupietz, "Constructing a Corpus," 63.

Ur-Stichprobe.[95] It would be slightly misleading to use the labels of "static/dynamic"[96] or "sample/monitor"[97] in this case, since neither this Koine corpus nor the Diorisis corpus is actively expanding, although there are certainly no methodological constraints that inhibit expansion. Since this Koine Greek corpus is merely a subset of the larger Diorisis corpus, it bears a greater affinity to the distinctions proposed by Kupietz.

4.2 *A Synchronic Corpus*

The size and temporal coverage of Diorisis allows the corpus to function like a primordial or monitor corpus, from which a more specific, synchronic corpus can be constructed. Synchrony is not in itself a better approach to lexical semantics than diachrony, as Saussure is sometimes thought to have believed.[98] Both are of interest to the lexicographer and linguist in determining the semantic change of a lexeme over a period of time (diachrony), and in determining the lexeme's range of meaning at a particular evolutionary moment in the history of the language (synchrony). There is a sense in which the synchronic is simply a product of diachronic study. Katerina Stathi writes,

> A corpus is thus understood as a collection of uses that have become established as the norm through constant use over a longer period of time. The "present state" of language is seen as the product of a development that is reflected in the established norms, viz. manifested in statistically significant phenomena.[99]

Another issue with synchrony concerns what actually constitutes "the past" and "the present." In the study of language use, can "the present" extend to cover a period of time (e.g. the Hellenistic or Roman periods)? Is not such an approach simply

[95] Kupietz, "Constructing a Corpus," 66.

[96] Gries, "What is Corpus Linguistics?," 1233.

[97] Sinclair, *Corpus*, 23–4.

[98] Thiselton, "Semantics and New Testament," 80. Although note Saussure's (*Course in General Linguistics*, 40) words, "The synchronic point of view predominates, for it is the true and only reality to the community of speakers . . . [If a linguist] takes the diachronic perspective, he no longer observes language [*langue*] but rather a series of events that modify it."

[99] Stathi, "Korpusbasierte Analyse," n.p. (translation mine).

diachrony on a smaller scale? But if there are theoretical issues with defining synchrony, this undoubtedly affects our understanding of diachrony also, since diachrony is nothing more than the comparison of two or more synchronic moments. In advocating for a synchronic corpus, then, two truths must be kept in tension, as Price articulates: "the slice of language under consideration must be long enough to provide an adequate number of examples of the lexical item in usage . . . but short enough that it does not contain examples which are no longer applicable due to meaning change over time."[100] The decision to opt for more limited temporal parameters (250 BCE–250 CE) and avoid the unstable periods of koineization (see 2) are an attempt to achieve this balance.

4.3 *Corpus Content*

We can now briefly summarize the contents of the Koine Greek corpus compiled for this study. The Koine Greek corpus is a synchronic *virtuelle* corpus, in that it was created by collating texts from the Diorisis corpus that fall within the temporal parameters designated in section 2.2 (250 BCE–250 CE). It can be further described as a general corpus (its contents are not restricted to a particular register or language domain), employing non-probability sampling (the collection of all available data). The key specifications are as follows:

Corpus	Synchronic Corpus of Koine Greek
Corpus Type	General, non-probability, sample (*virtuelle*) corpus
Text Count	449
Text Source	The Diorisis Corpus of Ancient Greek
Temporal Coverage	c. 250 BCE–250 CE
Token Count	c. 6,580,081 tokens

[100] Price, *Structural Lexicology*, 33.

Token Count (stopwords removed)[101]	c. 3,199,633 tokens
Word-type Count[102]	c. 36,362 tokens
Markup	TEI-XML
Metadata	Author, date, genre, sub-genre
Annotation	POS-tagging, lexical form

The contents of the Koine Greek corpus are tabulated in the Appendix below (6).[103]

5. *Conclusion*

Drawing together a synchronic corpus for Koine Greek from the larger Diorisis corpus is a relatively straightforward task. However, the task of compilation is preceded by a number of important theoretical issues concerning representativeness, size, and temporal coverage. I have argued that a synchronic corpus for ancient languages should be general, non-probability sampling, and reception-focused. This allows us to make use of all the available texts at our disposal, maximizing corpus size without having to extend the temporal coverage of the corpus beyond the limits of synchrony. In the case of Koine Greek, this opens up new avenues for corpus linguists and biblical scholars working with Greek of the Hellenistic and Roman periods, especially for those using NLP for corpus-based research.[104]

[101] The stoplist employed for this project is a modified and expanded version of the list constructed by Aurélien Berra: https://github.com/aurelberra/stopwords/tree/master/stopwords_for_quanteda.

[102] It is possible that some non-alphanumeric symbols remain after preprocessing, and thus the token/word-type counts are only approximate.

[103] Tauber ("Working with the Diorisis") has produced a more comprehensive list of the Diorisis metadata, including details of genre and sub-genre (https://github.com/jtauber/diorisis/blob/master/catalog.tsv). The word-counts in section 6 were calculated independently.

[104] For an example of NLP research employing this Koine Greek corpus, see List, "Exploring Kinship."

6. *Appendix*

	Author/ Collection	*Date Range*[105]	*Number of Texts*	*Word Count*	*Approx. Word Count (No Stopwords)*
1	Achilles Tatius	120 CE	1	41515	20334
2	Aelian	200–230 CE	3	143910	69417
3	Agathemerus	250 CE	1	1961	1127
4	Apollonius Rhodius	245 BCE	1	38808	25593
5	Appian	165 CE	14	222820	110141
6	Aretaeus	100 CE	4	50654	27054
7	Aristides, Aelius	142 CE	55	298438	132454
8	Arrian	120 CE	6	112802	53079
9	Asclepiodotus	35 BCE	1	6546	3123
10	Athenaeus	228 CE	1	267810	140711
11	Barnabas	130 CE	1	6713	3231
12	Bion of Phlossa	100 BCE	2	942	602
13	Cassius Dio	229 CE	1	189024	78472
14	Chariton	100 CE	1	34718	18420
15	Claudius Ptolemy	160 CE	1	37935	17064
16	Clement of Alexandria	195 CE	3	33145	17287
17	Demetrius	200 BCE	1	15409	7197
18	Dio Chrysostom	90 CE	1	173642	78600
19	Diodorus Siculus	35 BCE	3	377892	192000

[105] All date ranges are approximate, and were sourced from the metadata of the Diorisis Ancient Greek Corpus, although changes have been made to these designations for the New Testament and Septuagint collections.

20	Diogenes Laertius	230 CE	1	109099	53707
21	Dionysius of Halicarnassus	10 BCE	13	378008	184021
22	Epictetus	108 CE	3	83617	37660
23	Flavius Josephus	78 CE	4	466854	231121
24	Galen	170 CE	1	31808	14245
25	Harpocration	175 CE	1	37022	18402
26	Longinus	0 CE	1	12535	6131
27	Longus	150 CE	1	19679	10400
28	Lucian	145–200 CE	71	275582	127589
29	Marcus Aurelius	180 CE	1	29229	13481
30	Moschus	150 BCE	4	3158	1991
31	New Testament	50–170 CE	27	138925	64018
32	Onasander	50 CE	1	11521	5874
33	Oppian	171 CE	1	22752	15287
34	Oppian of Apamaea	211 CE	1	13482	9605
35	Parthenius of Nicaea	20 BCE	1	6399	3280
36	Pausanias	176 CE	1	217284	110293
37	Philostratus of Lemnos	230 CE	1	22885	10965
38	Philostratus the Athenian	213–238 CE	6	150854	70749
39	Philostratus the Younger	250 CE	1	7147	3508
40	Plotinus	270 CE	1	213493	80282
41	Plutarch	95 CE	145	996944	509210
42	Polybius	250 BCE	1	311307	145224

43	Pseudo Apollodorus	100 CE	1	26999	15438
44	Pseudo-Aristides	200 CE	1	22689	9584
45	Pseudo-Plutarch	200 CE	2	23352	11398
46	Septuagint (with Apocrypha)	250–150 BCE	53	587630	292852
47	Strabo	7 BCE	1	284516	136339
48	Triphiodorus	250 CE	1	4232	3037
49	Xenophon of Ephesus	110 CE	1	16395	8036
Total			449	6,580,081	3,199,633

Bibliography

Adrados, F. R. *A History of the Greek Language: From Its Origins to the Present*. Leiden: Brill, 2005.

Adrados, F. R., and J. R. Somolinos. “The ‘TLG’ Data Bank, the ‘DGE’ and Greek Lexicography.” *Emerita* 62 (1994) 241–52.

“The Ancient Greek and Latin Dependency Treebank,” [n.d.], https://perseusdl.github.io/treebank_data/.

Atkins, S, et al. “Corpus Design Criteria.” *Literary and Linguistic Computing* 7 (1992) 1–16.

Biber, D. “Methodological Issues Regarding Corpus-Based Analyses of Linguistic Variation.” *Literary and Linguistic Computing* 5 (1990) 257–69.

———. “Representativeness in Corpus Design.” *Literary and Linguistic Computing* 8 (1993) 243–57.

———. “Using Register-Diversified Corpora for General Language Studies.” *Computational Linguistics* 19 (1993) 219–41.

Brixhe, C. “Linguistic Diversity in Asia Minor during the Empire: Koine and Non-Greek Languages.” In *A Companion to the Ancient Greek Language*, edited by Egbert J. Bakker, 228–52. Chichester: Wiley-Blackwell, 2010.

Browning, R. “The Language of Byzantine Literature.” In *Greek Literature: Greek Literature in the Byzantine Period*, edited by G. Nagy, 103–33. New York: Routledge, 2001.

Bubenik, V. “Dialect Contact and Koineization: The Case of Hellenistic Greek.” *International Journal of the Sociology of Language* 99 (1993) 9–23.

———. “Koine, Origins of.” In *Encyclopedia of Ancient Greek Language and Linguistics*, edited by G. K. Giannakis, 217–85. Leiden: Brill, 2014.

———. “The Rise of Koine.” In *A History of Ancient Greek: From the Beginnings to Late Antiquity*, edited by A.-P. Christidēs et al., 342–45. Cambridge: Cambridge University Press, 2006.

Bullinaria, J. A., and J. P. Levy. “Extracting Semantic Representations from Word Co-occurrence Statistics: Stop-lists, Stemming, and SVD.” *Behavior Research Methods* 44 (2012) 890–907.

Burns, P. J. “Constructing Stoplists for Historical Languages.” *Digital Classics Online* 4 (2018) 4–20.

Caragounis, C. C. *The Development of Greek and the New Testament: Morphology, Syntax, Phonology, and Textual Transmission*. Grand Rapids: Baker Academic, 2006.

Church, K. W. “Speech and Language Processing: Where Have We Been and Where Are We Going?” Paper presented at the Eurospeech Conference: 8th European Conference on Speech Communication and Technology, Geneva, Switzerland, September 1–4, 2003.

Church, K. W., and R. L. Mercer. “Introduction to the Special Issue on Computational Linguistics Using Large Corpora.” *Computational Linguistics* 19 (1993) 1–24.

Clérice, T., and M. Munson. “Qualitative Analysis of Semantic Language Models.” In *Ancient Manuscripts in Digital Culture*, edited by D. Hamidović et al., 87–114. Leiden: Brill, 2019.

Colvin, S. *A Brief History of Ancient Greek*. Oxford: John Wiley & Sons, 2013.

———. “Koine Dialect.” *The Encyclopedia of Ancient History*, edited by Andrew Erskine et al. No pages. DOI: 10.1002/9781444338386.

Colwell, E. C. “The Greek Language.” In *The Interpreter’s Dictionary of the Bible*. Vol. 2, edited by G. A. Buttrick, 479–87. New York: Abingdon Press, 1962.

Gries, S. Th. “Dispersions and Adjusted Frequencies in Corpora.” *International Journal of Corpus Linguistics* 13 (2008) 403–37.

———. “What is Corpus Linguistics?” *Language and Linguistics Compass* 3 (2009) 1225–41.

Gut, U., and H. Voormann. *Corpus Design*. Oxford: Oxford University Press, 2014.

Horrocks, G. C. *Greek: A History of the Language and Its Speakers*. 2nd ed. Oxford: Wiley-Blackwell, 2010.

Hunston, S. “Corpus Linguistics.” In *Encyclopedia of Language and Linguistics*, edited by K. Brown, 234–48. 2nd ed. Oxford: Elsevier, 2006.

Jannaris, A. N. *An Historical Greek Grammar: Chiefly of the Attic Dialect as Written and Spoken from Classical Antiquity Down to the Present Time, Founded Upon the Ancient Texts, Inscriptions, Papyri and Present Popular Greek*. London: Macmillan, 1897.

———. “The True Meaning of the Κοινή.” *The Classical Review* 17 (1903) 93–96.

Kalton, G. *Introduction to Survey Sampling*. Beverly Hills: Sage, 1983.

Kim, L. “The Literary Heritage as Language: Atticism and the Second Sophistic.” In *A Companion to the Ancient Greek Language*, edited by Egbert J. Bakker, 468–82. Chichester: Wiley-Blackwell, 2010.

Köstenberger, Andreas J., et al. *Going Deeper with New Testament Greek: An Intermediate Study of the Grammar and Syntax of the New Testament*. Nashville: B&H Academic, 2017.

Kučera, H., and W. Francis. *Frequency Analysis of English Usage*. Boston: Houghton Mifflin, 1982.

Kupietz, M. “Constructing a Corpus.” In *The Oxford Handbook of Lexicography*, edited by P. Durkin, 62–75. Oxford: Oxford University Press, 2015.

Kupietz, M., and H. Keibel. “The Mannheim German Reference Corpus (DEREKO) as a Basis for Empirical Linguistic Research.” In *Working Papers in Corpus-Based Linguistics and Language Education*, edited by Makoto Minegishi et al., 53–59. Tokyo: Tokyo University of Foreign Studies, 2009.

Kupietz, M., et al. “The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research.” *Proceedings of the 7th Conference on International Language Resources and Evaluation* (2010) 1848–54.

List, N. “How Can We Investigate Ancient Greek Categories without the Influence of Our Own? Exploring Kinship Terminology using Word2Vec.” *International Journal of Lexicography* (forthcoming).

Mambrini, F. “Nominal vs Copular Clauses in a Diachronic Corpus of Ancient Greek Historians.” *Journal of Greek Linguistics* 19 (2019) 90–113.

McGillivray, B. S., et al. “A Computational Approach to Lexical Polysemy in Ancient Greek.” *Digital Scholarship in the Humanities* 34 (2019) 893–907.

Meyer, C. F. *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press, 2002.

Missiou, A. “The Hellenistic Period.” In *A History of Ancient Greek: From the Beginnings to Late Antiquity*, edited by A.-P. Christidēs et al., 325–41. Cambridge: Cambridge University Press, 2006.

Moț, L. *Morphological and Syntactical Irregularities in the Book of Revelation: A Greek Hypothesis*. Leiden: Brill, 2015.

Mühlhäusler, P. “The Development of the Category of Number in Tok Pisin.” In *Generative Studies on Creole Languages*, edited by P. Muysken, 35–84. Berlin: De Gruyter Mouton, 1981.

O’Donnell, M. B. *Corpus Linguistics and the Greek New Testament*. Sheffield: Sheffield Phoenix, 2005.

———. “Designing and Compiling a Register-Balanced Corpus of Hellenistic Greek for the Purpose of Linguistic Description and Investigation.” In *Diglossia and Other Topics in New Testament Linguistics*, edited by S. E. Porter, 255–97. JSNTsup 193. Sheffield: Sheffield Academic, 2000.

Pang, F. “Why We Need an Annotated Representative Corpus of Hellenistic Greek: The Compositionality of Greek *Aktionsart* for Movement Verbs as an Example.” In *In Mari Via Tua: Philological Studies in Honour of Antonio Piñero*, edited by Israel Muñoz Gallarte and Jesús Peláez del Rosal, 157–82. Córdoba: Ediciones El Almendro, 2016.

Pantelia, M. “‘Noûs, into Chaos’: The Creation of the Thesaurus of the Greek Language.” *International Journal of Lexicography* 13 (2000) 1–11.

Perrone, V., et al. “GASC: Genre-Aware Semantic Change for Ancient Greek.” In *The 1st International Workshop on Computational Approaches to Historical Language Change: Proceedings of the Workshop*, 56–66. Stroudsburg: ACL, 2019.

Porter, S. E. “The Greek Language of the New Testament.” In *A Handbook to the Exegesis of the New Testament*, edited by S. E. Porter, 99–130. Leiden: Brill, 1997.

———. *Linguistic Analysis of the Greek New Testament: Studies in Tools, Methods, and Practice*. Grand Rapids: Baker Academic, 2015.

Porter, S. E., and M. B. O'Donnell. "Building and Examining Linguistic Phenomena in a Corpus of Representative Papyri." In *The Language of the Papyri*, edited by T. Evans and D. Obbink, 287–311. Oxford: Oxford University Press, 2010.

———. "Theoretical Issues for Corpus Linguistics and the Study of Ancient Languages." In *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*, edited by A. Wilson et al., 119–37. New York: Peter Lang, 2003.

Price, T. L. *Structural Lexicology and the Greek New Testament: Applying Corpus Linguistics for Word Sense Possibility Delimitation Using Collocational Indicators*. Piscataway, NJ: Gorgias, 2015.

Robins, R. H. *The Byzantine Grammarians: Their Place in History*. Berlin: Walter de Gruyter, 1993.

Rodda, M. A., et al. "Panta rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek." In *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-It 2016*, edited by A. Corazza et al., 258–62. Torno: Accademia University Press, 2016.

———. "Vector Space Models of Ancient Greek Word Meaning, and a Case Study on Homer." *Traitement Automatique des Langues* 60 (2019) 63–87.

Rydberg-Cox, J. "Co-occurrence Patterns and Lexical Acquisition in Ancient Greek Texts." *Literary and Linguistic Computing* 15 (2000) 121–30.

Saussure, F. de. *Course in General Linguistics*. London: Peter Owen, 1959.

Siegel, J. "Koines and Koineization." *Language in Society* 14 (1985) 357–78.

Silk, M. *Standard Languages and Language Standards — Greek, Past and Present*. London: Routledge, 2016.

Sinclair, J. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.

———. "Corpus Typology — A Framework for Classification." In *Studies in Anglistics*, edited by G. Melchers and B. Warren, 17–33. Stockholm: Almqvist & Wiksell, 1995.

Stathi, K. "Korpusbasierte Analyse der Semantik von Idiomen." *Linguistik Online* 27 (2006). No pages. Online: http://www.linguistik-online.net/27_06/stathi.html.

Tauber, J. "Working with the Diorisis Ancient Greek Corpus." *J. K. Tauber* (January 2020). No pages. Online: https://jktauber.com/2020/01/20/working-with-the-diorisis-ancient-greek-corpus/.

"Thesaurus Linguae Graece," [n.d.], http://stephanus.tlg.uci.edu/.

Thiselton, A. C. “Semantics and New Testament Interpretation.” In *New Testament Interpretation: Essays on Principles and Methods*, edited by I. H. Marshall, 75–104. Carlisle: Paternoster, 1977.

Toufexis, N. “One Era’s Nonsense, Another’s Norm: Diachronic Study of Greek and the Computer.” In *Digital Research in the Study of Classical Antiquity*, edited by G. Bodard and S. Mahony, 105–20. Abingdon: Routledge, 2010.

Vatri, A., and B. McGillivray. “The Diorisis Ancient Greek Corpus.” *Research Data Journal for the Humanities and Social Sciences* 3 (2018) 55–65.

Wallace, D. *Greek Grammar Beyond the Basics: An Exegetical Syntax of the New Testament*. Grand Rapids: Zondervan, 1993.

Wishart, R. A. “Hierarchical and Distributional Lexical Field Theory: A Critical and Empirical Development of Louw and Nida’s Semantic Domain Model.” *International Journal of Lexicography* 31 (2018) 394–419.