A Distributional Perspective on Primary Sources from Ancient Greece

Laurent Gauthier

A Distributional Perspective on Primary Sources from Ancient Greece*

Laurent Gauthier†

August 05, 2021

Abstract

Ancient Greek history is sometimes perceived as offering a dearth of material for quantitative approaches, and it has not benefited from much interest from either quantitative historians or cliometricians. In this paper, we show that there is a strong potential for the quantitative exploitation of ancient Greek data, but new methods may be required. Indeed, focusing on primary sources, we build several large-scale and centralized datasets by exploiting electronically available, albeit sometimes atomic, literary and epigraphic texts, such as the Diorisis corpus, inscriptions from PHI, onomastic sequences from the BDEG and MAP, and individual names from the LGPN. Stepping aside from the methods that are common in quantitative history, we then consider the information in a distributional perspective inspired from the methods of complex systems analysis. Relying on this renewed combination of clio and metrics, we point out a series of regularity patterns in these historical records. Considering ancient Greek data in a non-atomic fashion, in the light of broad distributional patterns, raises new questions for historians.

Keywords: Ancient Greece, digital humanities, econometrics, power laws

*__Draft working paper__ - comments welcome!

†Email: laurent.o.gauthier@gmail.com. Laboratoire d'Économie Dionysien, Université Paris 8 Saint-Denis Vincennes, EA 3391 - Bâtiment D, 2, rue de la Liberté, 93526 Saint-Denis, France.

The etymological components of "cliometrics" refer to history on one hand, and more precisely to the Muse Κλειώ, who inspired poets when they sung the glory (κλέος) of past heroes, and to the taking of a measure on the other hand. Although there is no relation with economics in this etymology, cliometrics have by and large become a synonym for "quantitative historical economics", as a quick look at the field's handbook (Diebolt and Haupert (2019)) can attest. Some cliometricians have come to realize that cliometrics had drifted away from history, and in one such critic Fenoaltea (2019) regretted the increasing distance between these historical economists and their sources. In their *riposte* to Fenoaltea, Diebolt and Haupert (2020) argued, among other things, that modern cliometrics had helped further both economics and history by providing carefully grafted datasets, thanks to the importance of data for the cliometricians' uses, and thanks to their advanced econometric techniques. In the realm of ancient Greek history, however, the most significant contributions in the creation of large-scale primary sources have been due to specialized historians, classicists, philologists and computational linguists. Since the notion of an ancient economy is still hotly debated[1], and since there is little economics-related data pertaining to the period, the advances in cliometrics have not offered much help to ancient Greek history: the above-mentioned handbook does not contain a single reference to the Antiquity.

Considering that advanced econometric methods do not have to be restricted to the economy, but that large amounts of data can always benefit from quantitative scrutiny, we will review several important datasets pertaining to ancient Greece, directly derived from the Ancients' literary or epigraphic production. In our approach here, we choose indeed to remain close to the etymological view of cliometrics, as we focus on large-scale measures of ancient Greek primary sources. The question of the contribution of the methods from economics or econometrics to historiography has been examined in Gauthier (2021), and one of the salient conclusions was that if economics are to contribute results that can be exploited by historians to further their understanding of the past, it is important to focus on primary rather than secondary sources.

The list of sources of data pertaining to ancient Greece we discuss here is not intended to be fully comprehensive, but covers a fairly wide range of domains. We have grouped these data sources into three categories. First, we will examine core textual references, which include all literary tradition and all inscriptions; next, we will focus on instances of specialized textual sources, which are fundamentally derived from the former but are angled towards particular fields: votive inscriptions and 5th century theater. Finally, we will discuss relationship sources, whose representation can be construed as a network, also fundamentally derived from inscriptions

[1] See for example, in the case of the Roman economy, more often studied, Hobson (2014).

and literary sources: personal names and family relationships, and links between multiple gods in religious dedications.

Once equipped with the clio, our metrics will follow a different angle from the typical ways in which the data has been examined so far: we concentrate on its distributional features. While every observation remains important, we ask what kind of insights may be gained by considering the data in bulk, and in particular by looking at the probabilistic laws that may drive its occurrence. We can largely benefit from methods that have been used in statistical physics and in the study of complex systems. Before discussing the various sources we have listed, we will therefore briefly review some methodological aspects of distribution visualization and fitting.

All the sources mentioned above are either primary, directly produced by the Ancients, or are an organized subset of such primary sources. Nevertheless, it is sometimes helpful to leverage historical work in order to categorize such data from primary sources. One important resource for this purpose is the voluminous compilation by Hansen and Nielsen (2004). These authors gathered information on a large variety of categorizations and metrics applicable to each *polis*, for instance such as its size, political regime, or affiliation. The POLIS database[2] is the computerized and augmented version of this inventory, covering over 1,000 *poleis* of the Greek world. The majority of *poleis* listed in the POLIS database have geographic coordinates and in many cases the primary sources can be mapped to the polity-level information from POLIS[3].

For the most part, the primary sources of interest to us, as well as the POLIS data, are not designed to be electronically available in bulk. As a result, it is necessary to program specialized software to exploit the websites where the information is available as very large numbers of narrow subsets. For this purpose, we have mostly used the language `R`[4], in order to extract, process and analyze the data. Some specific packages or modules of that language are particularly useful for the treatment of the electronic data in the form in which it is made available[5]. Simply extracting and arranging these datasets requires programming expertise and powerful systems; this may explain why, so far, they have not often been exploited in bulk by historians.

[2]See Johnson and Ober (2014).

[3]The POLIS data itself, although many layers removed from primary sources, has been independently used in economics research: for example Fleck and Hanssen (2018) focused on political transition using this data exclusively.

[4]See R Core Team (2015).

[5]For the processing of textual data, Queiroz et al. (2020) is a standard. When the data needs to be accessed through a website, the browser emulation tools made available by Harrison and Kim (2020) are very handy.

1 Identifying Distributional Patterns

Most of the data sources we will discuss here contain large amounts of data of a categorical nature. Before we begin to drill into them, and try to understand various aspects of ancient Greek history under this light, it makes sense to establish what kind of patterns one may naturally expect from such data. Indeed, one could observe apparent regularities; what is the most direct manner in which to visualize them?

1.1 Some Useful Distribution Classes

Given the nature of the data at hand, we will typically seek to analyze the relationship between the size of some category (whether it be the number of times a word appears in a text corpus, or the number of votive acts a given god received, for example) and its rank. This can also equivalently be understood as considering the counter cumulative empirical probability: for an item $i \in [1..I]$ (a word, a god) with size s_i, its rank r_i is the number of items with a size greater than s_i, so that $r_i = |\{j \in [1..I] : s_j \geq s_i\}|$. Hence the empirical cumulative probability for the random variable representing the size S is $\mathbb{P}[S \geq s_i] = \frac{r_i}{I}$.

The kind of metrics we observe tend to be positive, and for the most part, their frequencies are strictly decreasing as a function of the observed values. While there is a very large number of random distributions that may account for these patterns, such positive observations are often compared with power laws, exponential distributions or truncated power laws, because these distributions are found in a large number of empirical phenomena and have strictly decreasing densities. These distributions are defined as follows:

- If X follows a power law of parameter[6] α and minimal value $x_{\min}$, we have

$$\mathbb{P}[X \in dx] = \mathbb{I}_{x>x_{\min}} \frac{\alpha}{x_{\min}} \left(\frac{x}{x_{\min}}\right)^{-1-\alpha} dx.$$

  A particular case of power law is Zipf's law, where $\alpha = 1$ and in which case the cumulative probability $\mathbb{P}[X > x]$ has the form $\frac{1}{x}$;

- If X follows an exponential distribution of parameter λ, shifted to take values above $x_{\min}$, we have

$$\mathbb{P}[X \in dx] = \mathbb{I}_{x>x_{\min}} \lambda e^{-\lambda(x-x_{\min})} dx;$$

[6] We follow the convention, more common in economics than in physics, of specifying the parameter as the exponent of the cumulative distribution, rather than that of the density. We hence have the same convention as Gabaix (1999), but that is different from that of Clauset, Shalizi, and Newman (2009), for example.

- If X follows a truncated power law of parameters α and λ, with minimal value $x_{\min}$, we have

$$\mathbb{P}[X \in dx] = \mathbb{I}_{x>x_{\min}} \frac{\lambda^{-\alpha}}{\Gamma(-\alpha, \lambda x_{\min})} e^{-\lambda x} x^{-1-\alpha} dx.$$

When considering distributions for which the frequency may not be decreasing, the lognormal distribution, as the exponential of a Gaussian, is a good candidate:

$$\mathbb{P}[X \in dx] = \mathbb{I}_{x>0} \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}} dx.$$

The generating mechanisms for power laws are generally associated with preferential attachment[7] or random growth processes[8] but can also be related to extreme value theory[9]. The more complex truncated power law can be associated with random group formation[10]. Exponential distributions are generally associated with random times, due to their "lack of memory" through conditioning. Connections have been found between exponential and power law distributions: the emergence of power laws can also be related to exponential processes[11]. In our discussion here we will not concern ourselves with the mechanisms that could have generated the distributions, only with establishing whether some empirical data is likely generated by such distributions.

1.2 Empirical Distribution Comparisons

Figure 1 plots some simulations of the three types of distributions mentioned above, for a large number of random draws. The parameters are chosen so that the curves are relatively close to each other on the logarithmic scale. Their shapes are clearly distinguishable: the straight line of the power law and the curvature of the exponential stand out. The truncated power law's distribution plot shows a straighter part on the left side (for high frequency, low value cases), and a curved part on the right side, where x is large and the exponential's behavior dominates the density. The exponential's effect hence affects the tail, the largest observations, while the lower values retain a power law-like behavior.

In this example, one may not need advanced statistical methods to establish the nature of these distributions: we can clearly see that the log/log plot, while heavily compressing the scale of

[7] See Barabási and Albert (1999), and more complex network formation mechanisms can still lead to some power law-like features, see Jackson and Rogers (2007).

[8] See Mitzenmacher (2004) and Gabaix (2016).

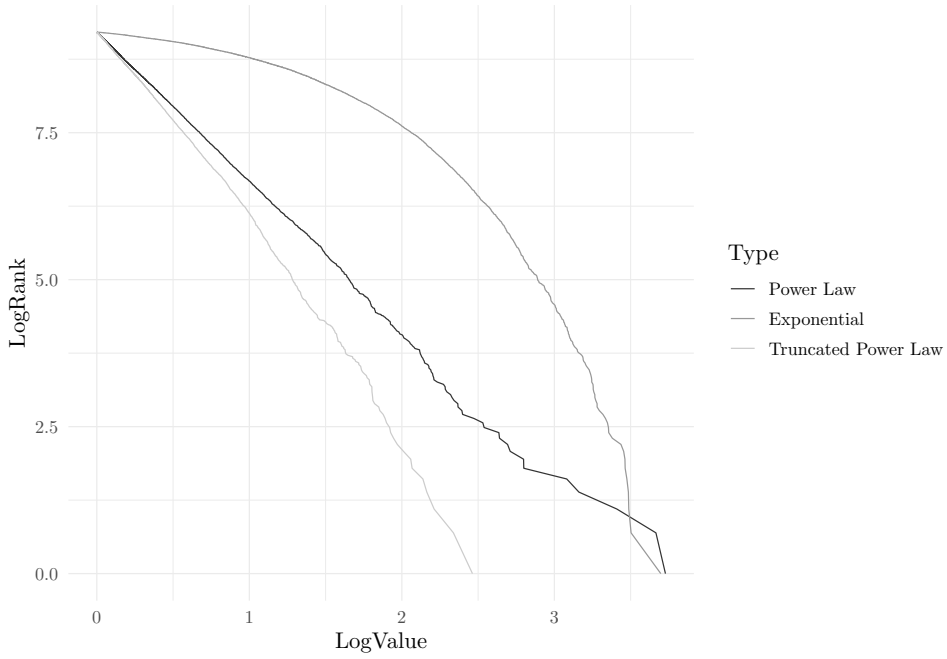
[9] See Alfarano and Lux (2010).

[10] See Baek, Bernhardsson, and Minnhagen (2011).

[11] See Reed and Hughes (2003).

the ranks and outcomes, does not betray the underlying density. It is not just any randomly generated numbers that, once ranked and rescaled, will exhibit the patterns visible in Figure 1.

Figure 1: *Comparison of Power Law, Exponential and Lognormal Distributions on a Large Sample*



Note: The parameters for the distributions are $\alpha = 2.5$ and $\lambda = 0.25$. The samples contain 10000 draws.

Clauset, Shalizi, and Newman (2009) offered a detailed treatment of the statistical methodology that can help empirically determine the type of probability law followed by some data, specifically in the case of power laws. They have shown that it is paramount to carry out proper distribution fit comparisons using a maximum likelihood approach, rather than simply carrying out a linear regression on the data on a logarithmic scale. Alstott, Bullmore, and Plenz (2014) proposed an implementation of their model in `Python`. Using the methodology from this implementation, we can compare the fits of the three distributions, as shown in Table 1.

We can see that the parameters estimates are very close to their true values, if we know what distribution to look up in each case. The goodness-of-fit ratios are clearly able to disambiguate the choice between power laws or truncated laws and the exponential. When we simulate a truncated power law the goodness-of-fit also indicated a preference for that form over a simple power law. However, when we simulate a power law, the tests cannot cleanly distinguish it from a truncated power law. In this case, however, the estimated parameter in either case is close to that of the underlying simulation, and the decay rate is very small. Given that the truncated

distribution with a small λ can come arbitrarily close to the straight power law, this is not surprising.

Table 1: *Summary Statistics on Distribution Fits*

Statistic	Power Law	Exponential	Truncated Power Law
Lambda Exp	1.481	0.253	2.028
Alpha Pow	2.482	0.751	2.955
Alpha Trunc	2.482	0.000	2.488
Lambda Trunc	0.000	0.100	0.222
Trunc vs Pow R	-0.013	76.626	3.464
Trunc vs Pow p	0.982	0.000	0.000
Trunc vs Exp R	10.644	-16.973	10.667
Trunc vs Exp p	0.000	0.000	0.000
Pow vs Exp R	10.643	-43.873	9.133
Pow vs Exp p	0.000	0.000	0.000

Note: The distribution names in the tests are abbreviated as follows: *Exp* = exponential, *Pow* = (pure) power, *Trunc* = power law with exponential decay. *R*: ratio of goodness-of-fit; a positive number means that the first law of the two is preferred. *p*: significance level; the probability that the preference would be due to randomness. The same abbreviations are used in other comparable tables.

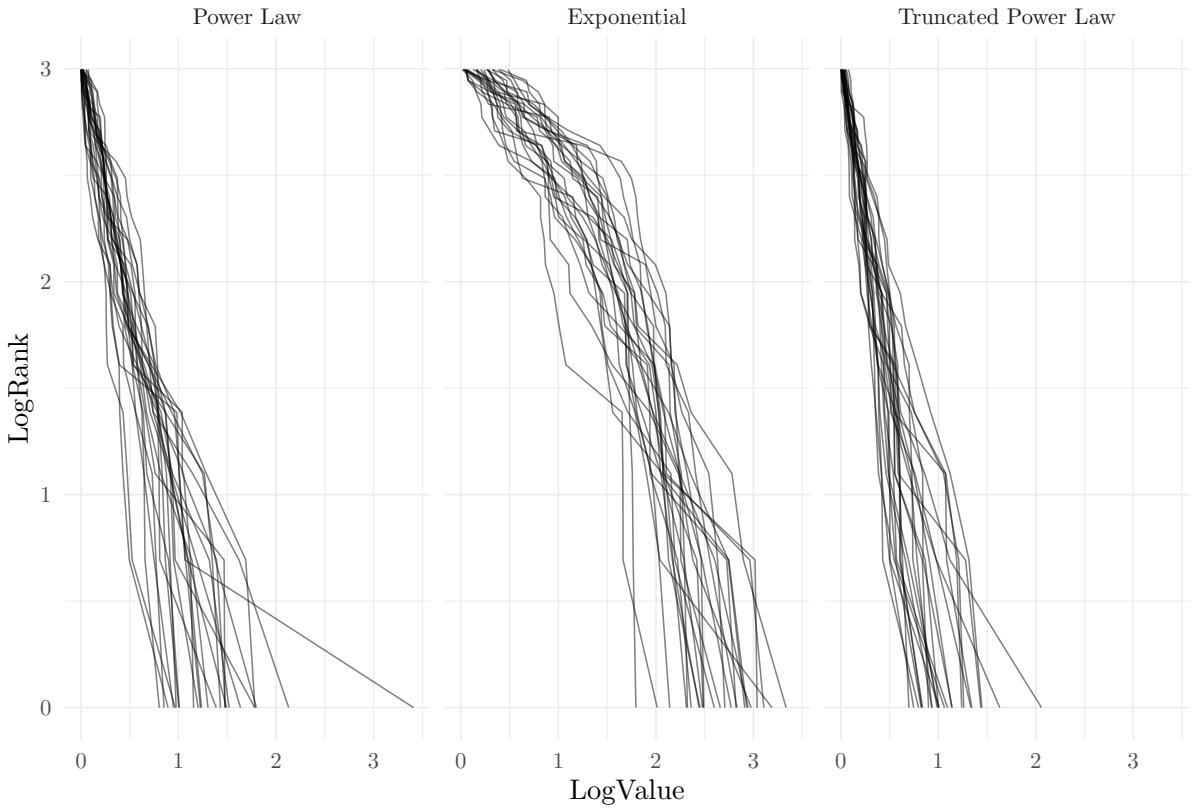
1.3 Dealing with Sample Size

The data that gives rise to distributional analyzes in the context of ancient Greek sources is in the form of categories, for which we can count occurrences. If the categories are words in a large text corpus, or individual persons in some historical record, or individual physical inscriptions across a large geographic space, then the number of these categories is very large, and sample sizes as well. However, in some cases one may not find so many different categories, when considering individual gods, or distinct characters in myth, for example.

Figure 2 shows the empirical cumulative distributions for a series of small samples from the same distributions as those represented in figure 1. Even with few observations in each case, and the variability across samples, we can see that the exponential distribution and power laws are markedly different. The behavior of the exponential for low ranks and low values gives it away. Since the effect of truncation on the power law is only apparent on the much scarcer high value and high rank outcomes, the distinction between the truncated and pure power law is naturally more difficult.

Distribution fits can be applied to each small sample, and the resulting statistics once aggregated are displayed in Table 2. The median significance levels for the goodness-of-fit are not very high across the board (high probability of the difference in distributions to be fortuitous). For power

Figure 2: *Comparison of Power Law, Exponential and Lognormal Distributions on Small Samples*



Note: The parameters for the distributions are $\alpha = 2.5$ and $\lambda = 0.25$. The 25 samples for each law contain 20 draws.

law and exponential distribution draws, the median parameter estimate comes out close to the real value, but that is not the case for the truncated power law; here again a consequence of the additional degree of freedom in fitting to a small number of observations. These results illustrate that with a reduced sample, it may still be possible to distinguish between exponential and power laws (just as one could see it on the chart, in Figure 2), but the more flexible nature of truncated power laws makes them more difficult to determine.

Now that we have defined a few practical statistical tools, we can turn to the sources of interest.

Table 2: *Summary Statistics Across Small Sample Distribution Fits*

Statistic	Power Law (Med.)	Power Law (St. Dv.)	Exponential (Med.)	Exponential (St. Dv.)	Truncated Power Law (Med.)	Truncated Power Law (St. Dv.)
Lambda Exp	1.817	0.769	0.260	0.060	2.055	0.581
Alpha Pow	2.807	0.742	0.742	0.097	2.914	0.622
Alpha Trunc	2.122	1.374	0.000	0.000	1.537	1.356
Lambda Trunc	0.341	0.557	0.119	0.037	0.732	0.688
Trunc vs Pow R	0.471	0.803	4.206	2.173	0.955	1.144
Trunc vs Pow p	0.745	0.342	0.060	0.195	0.545	0.348
Trunc vs Exp R	0.895	1.071	-1.644	1.125	0.167	1.168
Trunc vs Exp p	0.212	0.236	0.129	0.206	0.354	0.237
Pow vs Exp R	0.669	1.359	-3.005	1.383	-0.223	1.497
Pow vs Exp p	0.286	0.250	0.058	0.127	0.371	0.304

2 Core Textual Sources

Texts, more than any other artifact from the past, are the prime raw material of historiography. The most natural primary source for the study of ancient Greece, logically, is the corpus of all recorded literature, covering the range from the Archaic period to the late Antiquity. These texts have reached us, for the most part, through manuscripts which were copied through time; there are extremely few cases where we possess original literary writings. There are also numerous inscriptions, many with lacuna or missing characters, but they nevertheless constitute quite a large sample. Hence, the corpus of ancient Greek texts available today is the result of a combination of chance findings as well as specific choices that were made over more than 2,500 years about which works deserved conserving, and which ones did not. In spite of these layers of selection and filtering, the texts we have today belong to a fairly diverse set of genres.

Referring to ancient texts or written material immediately raises the issue of how they were established. Deciphering multiple manuscripts and compiling them into a cleanly set Greek text is an ever evolving process, even without the potential (re-)discovery of paleographic or epigraphic

sources. The Greek texts that are typically freely available for the kind of processing we are discussing here require to be free of copyright restrictions, and therefore usually are somewhat old. As a result, the underlying editions reflected in the electronic resources do not always reflect the latest advances in philological research. This issue is worth keeping in mind when carrying out a particular analysis, the conclusions of which may hinge upon very few individual words. To some extent, this issue can also be transposed to electronically available inscriptions.

In this section, we will concentrate in turn on two large ancient Greek text corpora: the Diorisis, which gathers literary works, and the PHI, which gathers inscriptions.

2.1 Large Scale Annotated Corpus: Diorisis

There are multiple electronically available resources compiling ancient Greek texts, such as the Abridged Thesaurus Linguae Graeca (TLG)[12], or the large Perseus website hosted by Tufts University[13]. These resources are nevertheless generally not absolutely comprehensive, do not allow users to download the data as one unique set, and only contain the raw text. For inflected languages such as Greek, morphological inflection makes the identification of variations of a same word difficult. In order to associate any instance of a word to its lemma, the noninflected root, and hence to identify each word's role in a sentence, the words need to be categorized, with a so-called POS tagger[14]. This can be done using specialized tools, such as the Classical Language Toolkit (CLTK)[15], but the systematic tagging of a large-scale text corpus demands significant computing resources and always suffers from identification errors[16]. Hence, there is significant value added in being able to use a centralized and comprehensive corpus of ancient Greek text that is already tagged, and contains more than raw text. This is the case of the Diorisis, whose tagging further benefits from on-going manual revisions. According to the corpus's presentation in Vatri and McGillivray (2018b):

> The corpus, aimed at classics and historical linguistics scholars, is the largest of its kind and can be used as an evidence basis for a wide range of studies on the Ancient Greek language.

This corpus indeed includes all of Perseus, which will be discussed in more detail in Section 3.2,

[12]See Pantelia (2020).

[13]See Crane (2012).

[14]Part-of-speech tagger, that can automatically determine the word's grammatical position in a sentence. See Ide (2004).

[15]See Johnson et al. (2019).

[16]Celano, Crane, and Majidi (2016) give a comparison of various POS taggers.

as well as the Little Sailing digital library[17], and the Bibliotheca Augustana digital library[18]. In total, the Diorisis gathers 820 different works.

The Diorisis Ancient Greek Corpus was originally composed in order to analyze semantic change in ancient Greek over time, with computational linguistics methods. Although some other corpora with similar information already existed, such as the Ancient Greek Dependency Treebank[19] or the PROIEL treebank[20], they were smaller by an order of magnitude (with less than a million words). However, Diorisis does not contain a full treebank: the text was automatically enriched with morphological information for each word, using a part-of-speech tagger to disambiguate certain words[21].

The Diorisis data contains one row per word or punctuation sign, for a total of over 14m rows. A lemma is mapped to each word, representing its root: the nominative form for a noun, or the singular first person of the present indicative for a verb in most cases. The particular declension or conjugation of the word is also specified. Additional information within the text, such as, for example, the speaker in a theater play, is not provided in the data. The canonical line reference is not either available, and the sentence identification that is given is only specific to Diorisis. Diorisis is indeed originally intended as a source for linguistics analysis, not specifically for literary or historical research.

The raw underlying data is available online[22]. The data can be accessed as a large list of files, one for each different ancient work, in `XML` format. Processing all these files in order to create a unique centralized dataset and adding some categorization information requires a good amount of compute power. Table 3 shows an excerpt of the first few entries in the resulting dataset.

Since the Diorisis contains so many different works, it is useful to be able to categorize these works. A very practical catalog has been compiled for that purpose, which can be found in Tauber (2020a). For each one of the works, it gives a year estimate, and a two-level literary genre category. Combining the catalog data with the Diorisis corpus, we can carry out a wide range of analyzes. Figure 3 shows the distribution of the number of words in Diorisis as a function of the

[17]See Perdikouris (2007).

[18]See Harsch (n.d.).

[19]See Celano, Crane, and Majidi (2016) and Celano, Crane, and Almas (2020).

[20]See Haug et al. (2009) and Eckhoff et al. (2018).

[21]Celano, Crane, and Majidi (2016) find that the best taggers reach an accuracy of about 90% on ancient Greek texts.

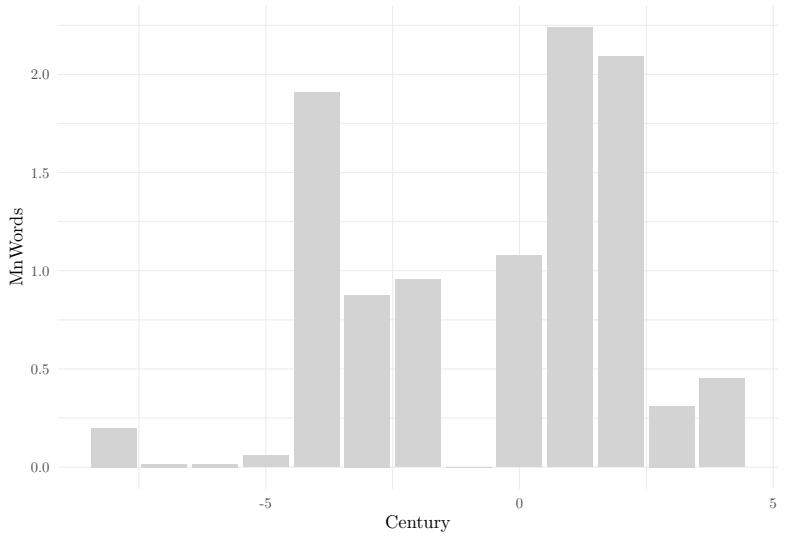
[22]See Vatri and McGillivray (2018a). Tauber (2020b) offers a critique of the Diorisis data structure as well as some advice on processing it.

Table 3: *Excerpt from Diorisis Corpus*

Title	Author	SentenceID	Word	WordID	Lemma	LemmaPOS	Punct.	AllRoles
Leucippe and Clitophon	Achilles Tatius	1	Σιδὼν	1	Σιδών	proper		fem nom/voc sg
Leucippe and Clitophon	Achilles Tatius	1	ἐπὶ	2	ἐπί	preposition		indeclform (prep)
Leucippe and Clitophon	Achilles Tatius	1	θαλάττῃ	3	θάλασσα	noun		fem dat sg (attic epic ionic)
Leucippe and Clitophon	Achilles Tatius	1	πόλις	4	πόλις	adjective		fem acc pl (epic doric ionic aeolic) | fem nom sg
Leucippe and Clitophon	Achilles Tatius	1					,	
Leucippe and Clitophon	Achilles Tatius	1	Ἀσσυρίων	5	Ἀσσύριοι	proper		masc gen pl
Leucippe and Clitophon	Achilles Tatius	1	ἡ	6	ὁ	article		fem nom/voc sg
Leucippe and Clitophon	Achilles Tatius	1	θάλασσα	7	θάλασσα	noun		fem nom/voc sg
Leucippe and Clitophon	Achilles Tatius	1					,	
Leucippe and Clitophon	Achilles Tatius	1	μήτηρ	8	μήτηρ	noun		fem nom sg

year when the work is estimated to have been composed. We can see the large concentrations in the traditional classical period (5th and 4th centuries), in Hellenistic Greece afterwards, and during the Roman Empire in the early Christian period.

Figure 3: *Distribution of the Number of Words by Work Period*



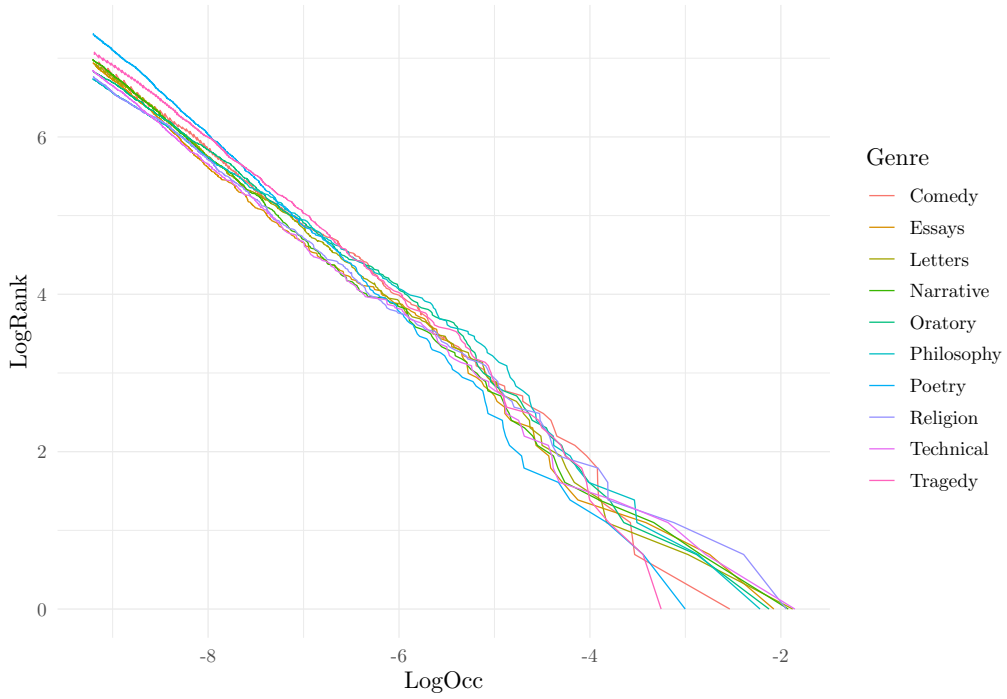
Note: The number of words is given in millions.

Typical linguistic patterns can also be observed on the data. Zipf's law has been famously observed and studied as it applies to the relationship between the frequency of occurrence of words in natural language and their frequency rank[23]. Figure 4 illustrates this relationship in the Diorisis corpus, grouped according to literary genre according to Tauber's catalog. The left hand side of the curves, representing the tail of the distribution, are indeed quite straight and with a slope close to 1, and it appears reasonable to call these distributions Zipf laws. We can

[23]See the overview in Piantadosi (2014).

note that poetry, comedy and tragedy all seem to exhibit thinner tails than the other genres, with their most frequent words being less common than for other genres; we can presumably relate this pattern to the fact that these bodies of literature typically resort to a broader range of vocabulary than the others. Studies on contemporary languages have exhibited similar differences by genre, as well as across languages: see Grabska-Gradzińska et al. (2012) for example, who focused on English and Polish.

Figure 4: *Log/Log Cumulative Distribution on Diorisis Corpus for Various Literary Genres*



Note: The data includes the lemmas for each genre that have more than 0.01% of occurrences. The horizontal axis is the logarithm of normalized frequency of each lemma, and the vertical axis is the logarithm of the lemma's rank.

Table 4 shows the application of the distribution statistics discussed in Section 1.2 to the data displayed in Figure 4. We can see that in all five cases, the distribution is identified as a power law or truncated power law, without a strong distinction between these two. The parameter α is close to 1: lower than 1 for prose (fatter tail), and higher for poetry (thinner tail), which is consistent with our earlier observation.

Table 4: *Summary Statistics on Distribution Fits Across Literary Genres*

Statistic	Narrative	Oratory	Philosophy	Poetry	Tragedy
Lambda Exp	0.001	0.001	0.001	0.003	0.003
Alpha Pow	0.828	0.850	0.815	1.139	1.021
Alpha Trunc	0.792	0.805	0.770	1.106	0.870
Lambda Trunc	0.000	0.000	0.000	0.000	0.000
Trunc vs Pow R	1.438	1.576	1.869	0.990	1.829
Trunc vs Pow p	0.000	0.011	0.009	0.214	0.029
Trunc vs Exp R	4.924	5.178	6.077	5.019	4.644
Trunc vs Exp p	0.000	0.000	0.000	0.000	0.000
Pow vs Exp R	4.866	5.098	5.984	4.969	4.313
Pow vs Exp p	0.000	0.000	0.000	0.000	0.000

Note: The data excludes words with less than 100 occurrences in text.

2.2 Inscriptions: PHI

The ancient Greeks used to inscribe a wealth of information on stone and on various artifacts, and these inscriptions are found on monuments, on steles, on vases and in many other places. These writings have reached us in their original form, albeit with some lacuna in many cases. This makes for a different situation from literary texts, which, with the exception of papyri, were copied many times over as manuscripts, and very often curated. The literary tradition has kept works that were deemed of value, but many written documents such as accounts or contracts have been lost[24]. Inscriptions therefore provide unique evidence for ancient Greece, since they constitute true original documents.

Reading and analyzing these documents is the role of the discipline of epigraphy, and transcribing, editing and contextualizing the texts from archaeological material is a complex process[25]. Epigraphic sources have been collated into large volumes since the Renaissance, and there are few centralized editions for Greek or Roman Antiquity. Online, the largest and most comprehensive repository of Greek inscriptions is the Packard Humanities Institute (PHI)'s Searchable Greek Inscriptions website[26].

Many inscriptions are in a damaged or fragmentary state, and epigraphy is also concerned with forming the best hypotheses in order to fill in the missing information, based on some commonly observed recurrences. This has traditionally been done manually by epigraphy specialists: the publications of the epigraphic sources contain a variety of codes that make explicit the editors'

[24]Many perishable documents have nevertheless been found in Egypt thanks to the particularly dry weather, dating back to the Hellenistic period or to the Roman empire; they are the subject of papyrology.

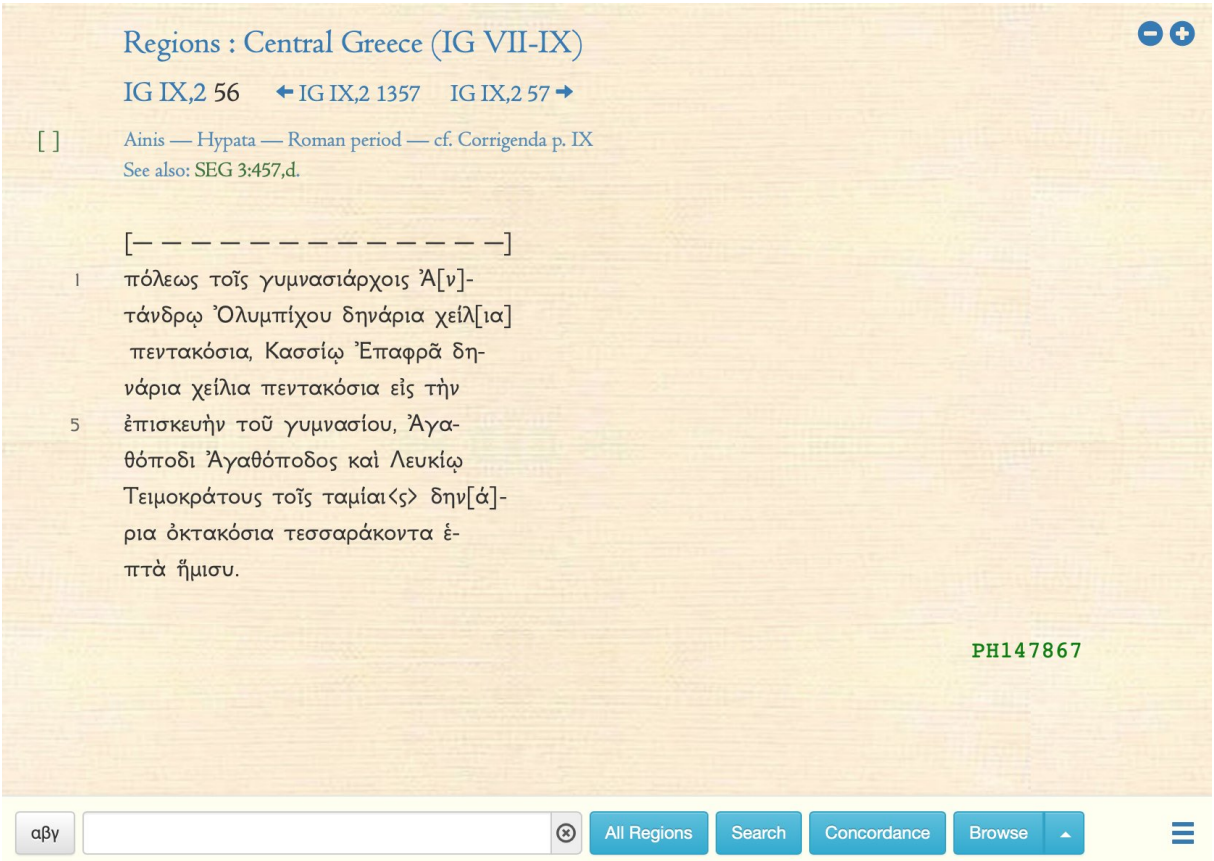
[25]in his handbook, Schaps (2011) gives an overview in the chapter dedicated to the field.

[26]See ("PHI Greek Inscriptions" n.d.).

choices and analyzes when there is material missing, or the inscriptions have been deteriorated[27]. Various alternative methods to restore missing texts have been explored, most notably artificial intelligence based on the surrounding words that are known and on large volumes of sufficiently legible sources[28].

The PHI data available online is represented as one webpage for each inscription, that contains the inscription's text as well as some additional geographic information. Figure 5 shows an example. We can see in this example some of the particular signs that are used to denote missing letters that have been hypothesized by the editor, or suggestions for letters that the engraver omitted.

Figure 5: *Query for a Particular Entry in the PHI*



Since there is no centralized database available containing the entirety of the PHI data, it must be processed page after page using an automatized web browser. While such large-scale data extractions have been carried out for the purpose of linguistics processing, it does not appear that truly cross-sectional historical analyses of ancient Greek inscriptions have been previously published. One can build a general extraction method by following the logic of Assael,

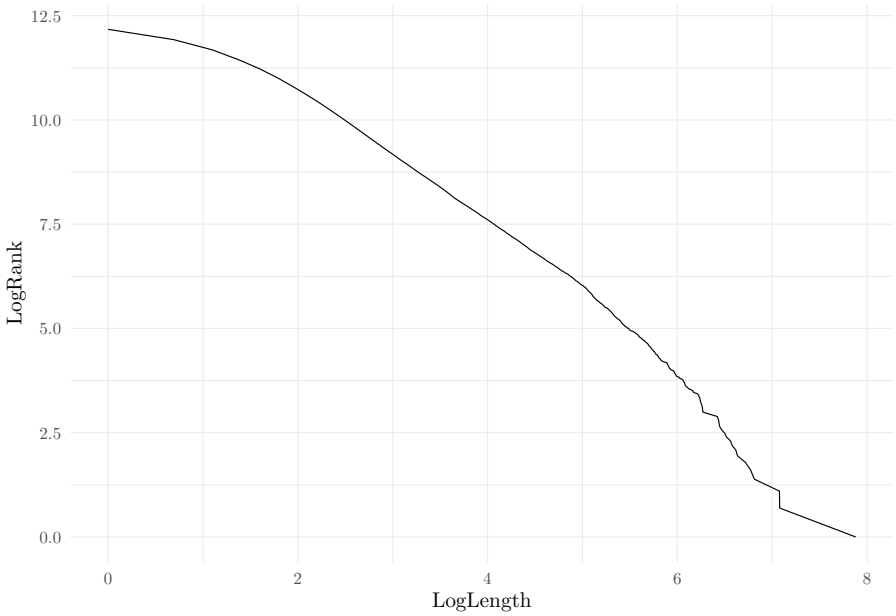
[27]For a description of the various codes in question, see Schaps (2011), p. 218-222.

[28]See Assael, Sommerschield, and Prag (2019) for example.

Sommerschield, and Prag (2020), who provide some `Python` code. Once all the inscriptions are electronically available in one place, it is possible to process them in order to map each recognizable word to a lemma, using the CLTK. Further, location information can be joined with the *polis*-level categorization data from the POLIS database.

With the centralized PHI data, one can carry out cross-sectional analysis which would be otherwise impossible. Figure 6 shows the distribution of the number of lines in the PHI inscriptions on a logarithmic scale. The shape of that curve seems consistent with an exponential distribution. There are several factors that could affect the length of inscriptions: first, the fact that the material may have been randomly broken naturally has an effect. However it may also be related to the cost of putting messages to stone, and to the nature of the messages themselves.

Figure 6: *Log/Log Cumulative Distribution of the Number of Lines in PHI Inscriptions*



As the summary of the data across regions illustrates, in Table 5, the typical inscription that has been recovered is quite short, although there are substantial geographical differences. The median quality of the text, as measured by the share of voids, is not constant across geographies either.

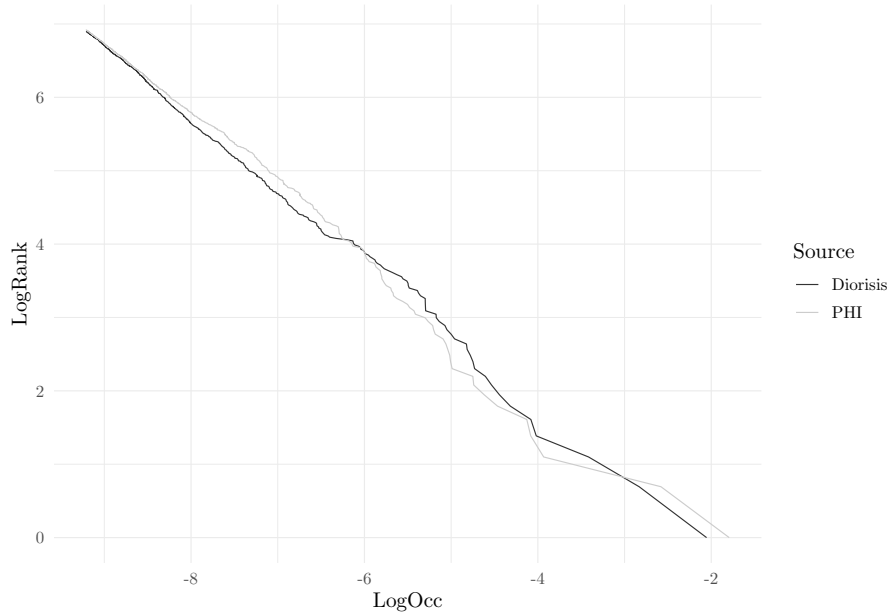
Table 5: *Aggregate Characteristics of PHI Inscriptions by Region*

Region	Nb Inscr.	Nb Lines	Nb Chars	Char per Line	Pct Voids
Attica	29264	3	37	26	13
Egypt and Nubia	13755	2	27	23	4
Ionia	11635	4	52	28	5
Megaris, Oropia, and Boiotia (IG VII)	8310	1	18	27	7
Italy, incl. Magna Graecia	7288	3	33	19	5
Thrace and Moesia Inferior	6951	3	45	21	8
Caria	6922	5	69	29	3
Macedonia	6834	4	54	20	8
Sicily, Sardinia, and neighboring Islands	6119	2	18	18	7
Rhodes and S. Dodecanese (IG XII,1)	5888	2	26	19	6
Cos and Calymna (IG XII,4)	4856	4	61	31	14
Thessaly (IG IX,2)	4761	3	36	26	8

Note: Voids are the characters that are unknown in the inscriptions. The aggregate characteristics are reported as medians.

Interestingly, although inscriptions were generally short, the distribution of word lemma occurrences broadly matches that of a literary corpus. Figure 7 plots a comparison of the distribution of lemma occurrences in the Diorisis literary corpus and in the PHI corpus. They look very similar, which indicates that the language on inscriptions has comparable features to the standard literary Greek language.

Figure 7: *Log/Log Cumulative Distribution on Diorisis and PHI Corpora*



Note: The data includes the most common lemmas for each corpus, which account for more than 0.01% of occurrences. The horizontal axis is the logarithm of normalized frequency of each lemma, and the vertical axis is the logarithm of the lemma's rank.

Nevertheless, a more precise statistical fit, displayed in Table 6, shows that the use of words in inscriptions is less fat-tailed than in literary works: the inscriptions use rare words more often. This is presumably not related to a greater presence of poetry in these inscriptions, but rather to the more common appearance of random proper names, which would be much less likely in literature.

Table 6: *Summary Statistics on Distribution Fits*

Statistic	PHI	Diorisis
Lambda Exp	0.001	0.001
Alpha Pow	0.852	0.774
Alpha Trunc	0.822	0.746
Lambda Trunc	0.000	0.000
Trunc vs Pow R	1.161	1.962
Trunc vs Pow p	0.006	0.000
Trunc vs Exp R	4.386	7.321
Trunc vs Exp p	0.000	0.000
Pow vs Exp R	4.346	7.256
Pow vs Exp p	0.000	0.000

Note: The data excludes words with less than 100 occurrences in text.

3 Sources Composed from Textual Corpora

While the textual sources we have discussed above contain a majority of the texts from Ancient Greece[29], every possible analysis stemming from the underlying texts is not directly doable through automatic processing. In fact, it takes a lot of human effort in some cases in order to create particular corpora from these sources, adapted to the questions they seek to address. Through selection of subsets of all literature or inscriptions, and the categorization of atomic elements, historians or classicists have created other, more complex corpora that are electronically available.

We will focus on two particular examples: the BDEG database of divine epithets, and Perseus's Greek theater texts.

3.1 Divine Epithets: BDEG

The study of ancient Greek religion and ritual is a significant part of historical research on the period. Inscriptions, when available, give us a glimpse of actual votive acts, and have been closely examined for this purpose. Epigraphic sources related to religious offerings or dedications typically contained the names of one or several gods along with some qualifications and reasons for the inscription. The centralization and analysis of such inscriptions have been carried out for over a century, starting with thousands of paper files in the 19th century. Bonnet and Lebreton (2019) discuss the historiography of the study of divine invocations, and explain how methods have evolved over time, converging towards the use of databases.

One instance of such a database is the Base de Donnée des Epiclèses Grecques (BDEG)[30], a project initiated in the early 2000s. It contains the information from thousands of epigraphic inscriptions or literary references to votive acts, including location, the god or gods names and the nature of the inscription[31]. The BDEG's goal was to allow researchers to study Greek polytheism and Greek religion in a more systematic and quantitative manner than what had been possible up to that point. More specifically, researchers gathered and made available data about Greek divine epithets from all sources, periods and region they may come from. The general notion of god name was understood as a system combining the actual name of the god, typically in the first place, with some additional qualification (commonly an adjective). The additional qualification,

[29] Excluding papyri.

[30] See Lebreton et al. (2014), and Brulé and Lebreton (2007).

[31] The data also contains divine names associated to epithets that could have appeared in literary or epigraphic sources about a sanctuary or a celebration and hence do not match a precise votive act.

named cult-epithet, describes a particular feature of the god. The nominal group thus stored into the database each time refers to a precise divine entity worshipped by ancient Greeks.

The data has been entered in over 11,000 forms, with in principle one for each observation of a divinity and epithets[32]. The BDEG presents its data as separate webpages, as illustrated in Figure 8. For each inscription or group of inscriptions, unstructured data generally captures some details on the source and context:

- The source itself may be literary, epigraphic, papyrological or numismatic.
- Based on the available information, such as the archaeological context, a dating range has sometimes been made available as a period or a century.
- The presence of a cult is an important qualification; some mentions in the inscription of a sacrifice, a priest, a sanctuary or just a dedication confirmed that there was a cult to the deity.

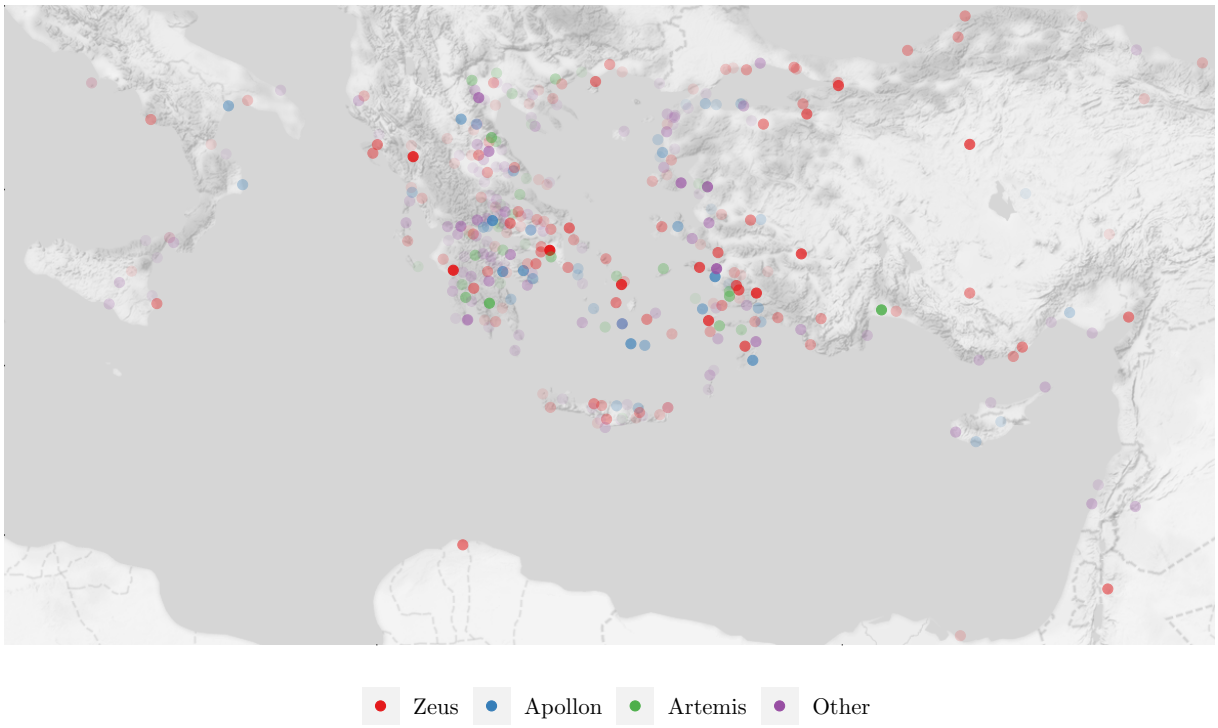
Figure 8: *Query for a Particular Entry in the BDEG*

The BDEG data is designed to allow researchers to visualize a form at a time, and in order to analyze the data in bulk, it is necessary to programmatically collate, parse and compile all the

[32]One significant issue has affected the BDEG database: the data was initially entered as one form for all the observations of a particular epithet in a location, but later evolved so that each entry represented a separate observation. However, the forms also contain categorical data specifying a range for the number of observations of the specific divinity and epithets, and that information can be used to carry out adjustments.

data from all these webpages. Location information is provided in some cases with coordinates, and in other cases only with the gallicized Greek name. Once the names are manually mapped to the corresponding anglicized Greek names, location information can be joined with the data from the POLIS database. Unfortunately, in many cases neither geographic coordinates nor mappable names are provided, and we cannot properly associate these observations to a *polis* or a location. Figure 9 shows the geographic coverage in the BDEG data, based on this location information, as well as the prevalence of some particular gods in the observations. The figure shows that epigraphic sources pertaining to rituals are quite extent geographically, and that there appears to be a definite concentration on some particular gods.

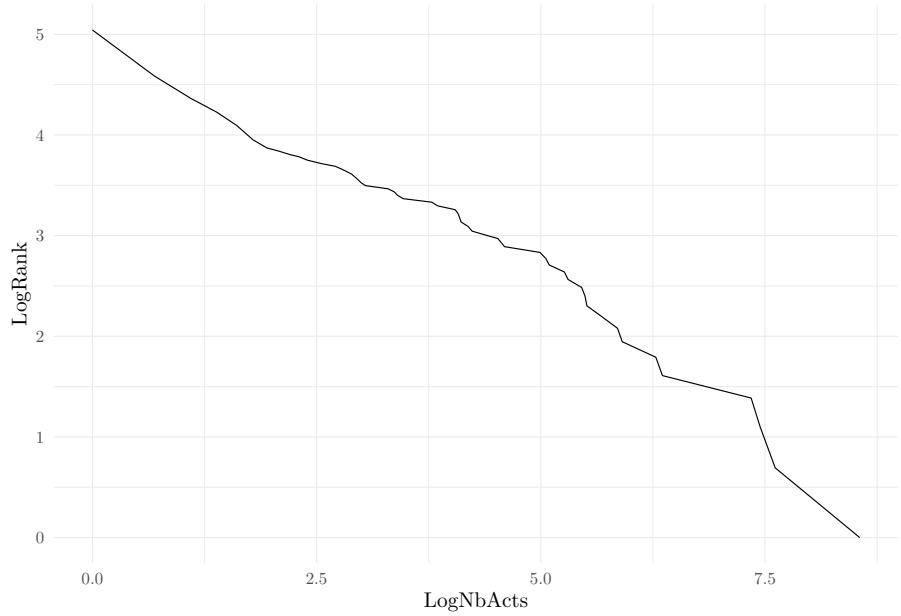
Figure 9: *Geographic Distribution of BDEG and POLIS Data*



Note: The map background shows modern political boundaries. The dots show the *poleis* on which the BDEG database provided some information on worship acts, with opaqueness a function of the logarithm of the number of observations.

The divinities in the Greek pantheon were indeed far from equally treated by humans. Taken across the entire dataset, the distribution of votive acts across individual gods is shown in Figure 10.

Figure 10: *Log/Log Cumulative Distribution of the Distribution of Votive Acts Among Gods*



The approximate straight line is indicative of a power law, as confirmed by the statistical tests from Table 7. In addition, the estimate gives us approximately $\alpha = \frac{1}{2}$, so that the distribution has a fatter tail than Zipf's law: there are a few gods that attract large numbers of dedications.

Table 7: *Summary Statistics on Distribution Fits for Votive Acts*

Statistic	BDEG
Lambda Exp	0.013
Alpha Pow	0.482
Alpha Trunc	0.510
Lambda Trunc	0.000
Trunc vs Pow R	1.710
Trunc vs Pow p	0.060
Trunc vs Exp R	7.175
Trunc vs Exp p	0.000
Pow vs Exp R	7.106
Pow vs Exp p	0.000

3.2 Fifth Century Theater from Perseus

Out of the hundreds of literary works that compose the ancient Greek corpus, theater plays a particular role, and analyzing this specific stream of literature requires tracking verse-level categorization information, such as who the speaker is, information on the characters, and the phase of the play being represented[33].

The very word of theater is linked to the notion of dramatic performance since it comes from

[33]Such as the *parodos*, when the chorus enters the stage, or the *exodos* when it leaves it.

the Greek θεάομαι, which means to watch attentively, but it takes on a meaning that goes well beyond a simple representation. The ancient Greek theater, even if it is related to what we would identify today as a spectacle, remains fundamentally embedded in religion, politics and the civic activity of the *polis*: it is a popular national theater closely linked to the city and to the gods in the case of Athens[34]. In fact, it makes up an important part of the community practices which have played a decisive role in the emergence of the *polis*, according to historical anthropology[35]. For Vernant, tragedy is "a social institution that [...] the city sets up alongside its political and judicial bodies[36]".

The plays were performed only once, during the Great Dionysies. These works combined text, acting, song, dance, stage acting, audience members (especially in comedy), and the context of an important religious celebration. The material elements available to us today, in the form of texts, various inscriptions, or monuments, are only fragmentary projections of this complex reality. For comedy as well as for tragedy, there is in this theater a specific formal structure as well as a metric framework which influences the nature of the language[37]. Despite these limitations, theater texts can open a window on psychology in ancient Greece. The dialogues in a scene must indeed show everything that happens over a given period of time with great consistency, and not a restricted selection: all the characters and all their interactions are represented, and there is no equivalent in other contemporary literary sources. This is an important characteristics for studying ancient Greek human interactions, since it is then possible to identify all of their occurrences in a particular situation.

In order to obtain theater texts in a structured fashion, reflecting speaker information and respecting verses and metric, one cannot use Diorisis. Indeed, it is necessary to go to one of its sources, Perseus[38]. Since Perseus is designed to allow users to read the ancient texts online, it contains all the required information, as can be seen in Figure 11. However, the texts are not available in a unique dataset, but either through the webpages, or as separate `XML` files, and they do not follow a unique TEI format[39]. Based on the data from Perseus, we have constituted a

[34]Saïd, Trédé, and Le Boulluec (2017), p. 117.

[35]See Azoulay (2014).

[36]Vernant and Vidal-Naquet (2001), p. 24.

[37]In tragedy, in particular, the main characters (actors who often play heroes) express themselves in everyday language with a metric structure, the choir (of ordinary citizens who often play a group of ordinary people) speaks in the language of lyrical poetry for the sung parts. See Vernant and Vidal-Naquet (2001), p. 27.

[38]See Crane (2012).

[39]For several plays, this work can be made easier by using the data from DraCor, who have sought to unify the formats from the Perseus plays, see Fischer, Frank, Börner, Ingo, and Göbel, Mathias (2019).

unified corpus of all the 44 non-fragmentary plays from the 5th century that are available today. A good amount of data processing and cleaning is required in order to address typographic noise, such as slight differences in character names, or differences in the encoding of the plays' metadata. We obtained the texts of all 7 of Aeschylus's plays, all 7 of Sophocles, and all 19 of Euripides, for a total of 33 tragedies, and 11 comedies from Aristophanes.

Figure 11: *Query for a Particular Play in Perseus*

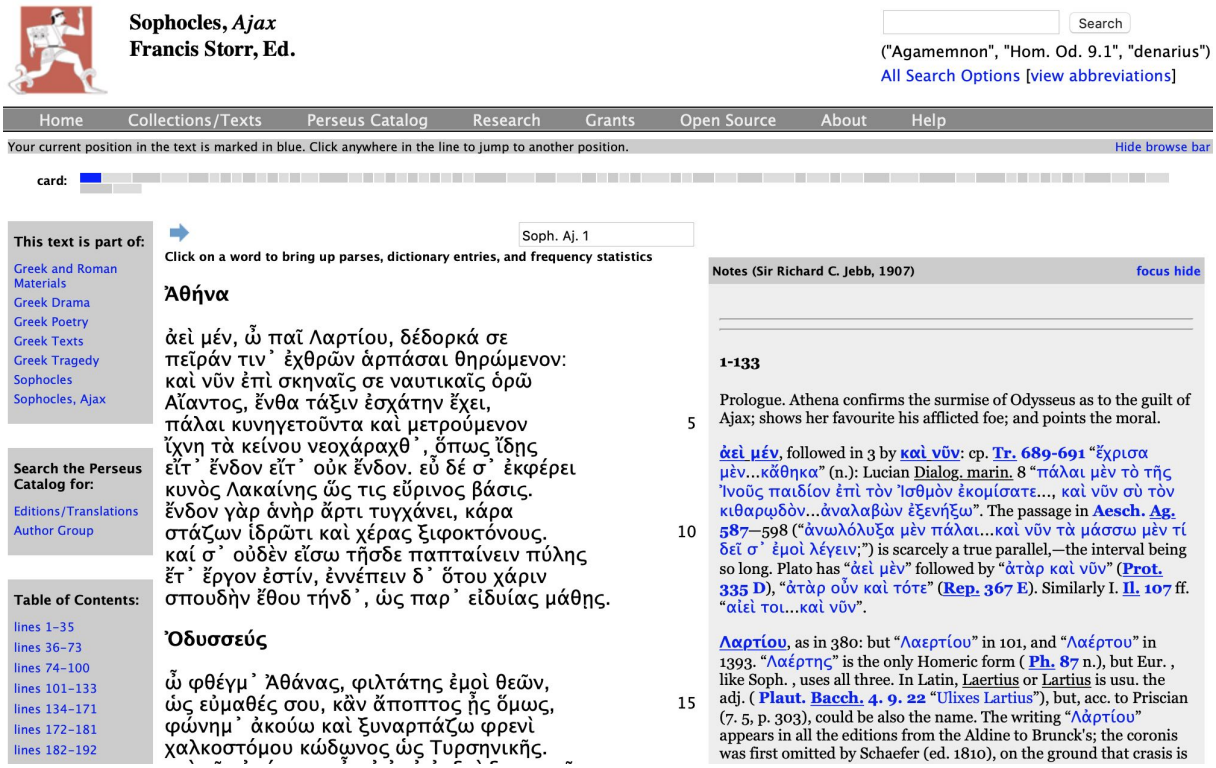


Table 8 is an extract from the principal data table we generated from Perseus. For practical reasons, it is useful to have a latinized representation of the Greek text and character names. Representing the entire 5th century theater corpus as a data table allows for the easy identification of characters, verses, and interactions.

In addition to the table containing the text, it is advisable to systematically categorize all the characters. Table 9 shows an extract of such a categorization. The dimensions along which the characters are defined capture the main drivers of social distinction in the archaic and classical *polis*, and in Athens in particular:

- Slavery is a critical social distinction in Greek Antiquity, and Athens had by most estimates tens of thousands of slaves in the Classical period (about a third of the total population[40]), who were socially integrated in the *polis*, but politically excluded. The opposition between

[40]Ober (2015), in table p. 92, as well as Pébarthe (2008), p. 163, arrive to this same order of magnitude.

Table 8: *Extract from the Main Theater Text Table (Sophocle's Ajax)*

Author	Title	Speaker	LatinizedSpeaker	Line	Text
Sophocles	Ajax	Ἀθήνα	Athena	1	ἀεὶ μέν, ὦ παῖ Λαρτίου, δέδορκά σε
Sophocles	Ajax	Ἀθήνα	Athena	2	πεῖράν τιν᾽ ἐχθρῶν ἁρπάσαι θηρώμενον·
Sophocles	Ajax	Ἀθήνα	Athena	3	καὶ νῦν ἐπὶ σκηναῖς σε ναυτικαῖς ὁρῶ
Sophocles	Ajax	Ἀθήνα	Athena	4	Αἴαντος, ἔνθα τάξιν ἐσχάτην ἔχει,
Sophocles	Ajax	Ἀθήνα	Athena	5	πάλαι κυνηγετοῦντα καὶ μετρούμενον
Sophocles	Ajax	Ἀθήνα	Athena	6	ἴχνη τὰ κείνου νεοχάραχθ᾽, ὅπως ἴδῃς
Sophocles	Ajax	Ἀθήνα	Athena	7	εἴτ᾽ ἔνδον εἴτ᾽ οὐκ ἔνδον. εὖ δέ σ᾽ ἐκφέρει
Sophocles	Ajax	Ἀθήνα	Athena	8	κυνὸς Λακαίνης ὥς τις εὔρινος βάσις.
Sophocles	Ajax	Ἀθήνα	Athena	9	ἔνδον γὰρ ἀνὴρ ἄρτι τυγχάνει, κάρα
Sophocles	Ajax	Ἀθήνα	Athena	10	στάζων ἱδρῶτι καὶ χέρας ξιφοκτόνους.
Sophocles	Ajax	Ἀθήνα	Athena	11	καί σ᾽ οὐδὲν εἴσω τῆσδε παπταίνειν πύλης
Sophocles	Ajax	Ἀθήνα	Athena	12	ἔτ᾽ ἔργον ἐστίν, ἐννέπειν δ᾽ ὅτου χάριν
Sophocles	Ajax	Ἀθήνα	Athena	13	σπουδὴν ἔθου τήνδ᾽, ὡς παρ᾽ εἰδυίας μάθῃς.
Sophocles	Ajax	Ὀδυσσεύς	Odysseus	14	ὦ φθέγμ᾽ Ἀθάνας, φιλτάτης ἐμοὶ θεῶν,
Sophocles	Ajax	Ὀδυσσεύς	Odysseus	15	ὡς εὐμαθές σου, κἂν ἄποπτος ᾖς ὅμως,

free man and slave has long been seen as essential in defining the notion of citizen[41]. In order to categorize each character's status with respect to slavery, we considered that war captives were slaves on the one hand, and that nannies, servants or maids, and pedagogues were also slaves on the other. These functions were in fact generally assigned to slaves[42].

- At the other end of the spectrum, Ober estimates the share of the elite at just over 1% of the population, based on an econometric model[43].
- Even if life expectancy at birth in ancient times was short, this was due to the very high infant mortality rate and does not mean that there were no old people; in fact, conditioned upon reaching the age of 10, the probability of reaching 70 was about 15%[44]. In aggregate, it is estimated that people over 60 represented about 5-10% of the total population[45].
- Among free men, foreigners, whether passing through or domiciled in the city, represented an important part of the population in Athens. Roubineau retains the figure of 15 to 20% of free adults[46]. The unique notion of *xenos*, the foreigner, in fact covers a whole gradation, and different terms and realities depending on each particular *polis*, but in general, foreigners had specific duties vis-à-vis the host city, and the right they had to live there was a privilege.

[41]See Mossé (2011).

[42]See Bonnard, Dasen, and Wilgaux (2017), p. 217-220.

[43]See Ober (2015), p. 92-97; this is the proportion relative to the population including all social classes.

[44]See Bonnard, Dasen, and Wilgaux (2017), p. 147.

[45]See Bonnard, Dasen, and Wilgaux (2017), p. 147 and Corvisier (1986), p. 57.

[46]See Roubineau (2015), p. 39; these figures are comparable to those given by Mansouri (2011), p. 27.

Table 9: *Excerpt from the Character Categorization Table (Sophocle's Ajax and Euripides' Medea)*

Author	LatinizedTitle	LatinizedSpeaker	PlayType	Chorus	Abstract	Generic	Religious	Slave	Xenos	Divine	Royal	Warrior	Age	Gender
Euripides	Medea	Trophos	Tragedy	No	No	No	No	Yes	Yes	No	No	No	Mid	Female
Euripides	Medea	Paidagogos	Tragedy	No	No	No	No	Yes	Yes	No	No	No	Old	Male
Euripides	Medea	Medeia	Tragedy	No	No	No	No	No	Yes	No	Yes	No	Mid	Female
Euripides	Medea	Choros	Tragedy	Yes	No	Yes	No	No	No	No	No	No	Mid	Female
Euripides	Medea	Kreon	Tragedy	No	No	No	No	No	No	No	Yes	No	Old	Male
Euripides	Medea	Iason	Tragedy	No	No	No	No	No	Yes	No	Yes	Yes	Mid	Male
Euripides	Medea	Aigeus	Tragedy	No	No	No	No	No	Yes	No	Yes	No	Old	Male
Euripides	Medea	Angelos	Tragedy	No	No	Yes	No	No	No	No	No	No	Mid	Male
Euripides	Medea	Pais	Tragedy	No	No	No	No	No	Yes	No	Yes	No	Child	Male
Sophocles	Ajax	Agamemnon	Tragedy	No	No	No	No	No	No	No	Yes	Yes	Mid	Male
Sophocles	Ajax	Aias	Tragedy	No	No	No	No	No	No	No	Yes	Yes	Mid	Male
Sophocles	Ajax	Angelos	Tragedy	No	No	Yes	No	No	No	No	No	No	Mid	Male
Sophocles	Ajax	Athena	Tragedy	No	No	No	No	No	No	Yes	No	No	Young	Female
Sophocles	Ajax	Choros	Tragedy	Yes	No	Yes	No	No	No	No	No	No	Mid	Male
Sophocles	Ajax	Hemichorion 1	Tragedy	Yes	No	Yes	No	No	No	No	No	No	Mid	Male
Sophocles	Ajax	Hemichorion 2	Tragedy	Yes	No	Yes	No	No	No	No	No	No	Mid	Male
Sophocles	Ajax	Menelaos	Tragedy	No	No	No	No	No	No	No	Yes	Yes	Mid	Male
Sophocles	Ajax	Odysseus	Tragedy	No	No	No	No	No	No	No	Yes	Yes	Mid	Male
Sophocles	Ajax	Tekmessa	Tragedy	No	No	No	No	Yes	Yes	No	No	No	Mid	Female
Sophocles	Ajax	Teukros	Tragedy	No	No	No	No	No	Yes	No	No	No	Mid	Male

Combining character categorization with the theater data allows for large-scale analysis, across tens of thousands of verses. Table 10 displays the volume of verses allocated to various character categories across our entire theater corpus. We can observe that men occupy two thirds of the speaking time, which constitutes an imbalance in relation to the natural sex ratio, presumably around 50%. Women, in relation to their share of the population, are under-represented in theater. In relation with the estimated age pyramid we briefly discussed earlier, we can also see that older characters are over-represented in these texts.

Although only a fraction of the Athenian population of the 5th century could be categorized as belonging to the aristocracy, they took over the stage. We can see in the table that aristocrat characters pronounced close to 50% of the verses in the corpus. In contrast, while slaves represented a large part of the population, they have almost no existence on stage compared to the free. *Xenoi*, on the other hand, seem to benefit from a share in the dialogues equal to what we can estimate their share in the general population to be, although they tended not to be in a position of authority. We can associate this more intense presence with the special attention that the tragedy pays to the characters put in this situation[47].

If we consider the different categories and social statuses mentioned so far, it appears that, in general and with a few exceptions, the bearers of more authority have a greater presence on stage. We can extend this logic to the case of the gods: indeed, even if they are little represented

[47] "In tragedy it is not uncommon to find, expressed perhaps in different words, the situation of the man living in a foreign land, deprived of his civil rights, who must accept rigorous limitations on his ability to act and assert himself in society." See Citti (1988), p. 456.

on stage, they are naturally rare in the *polis*[48].

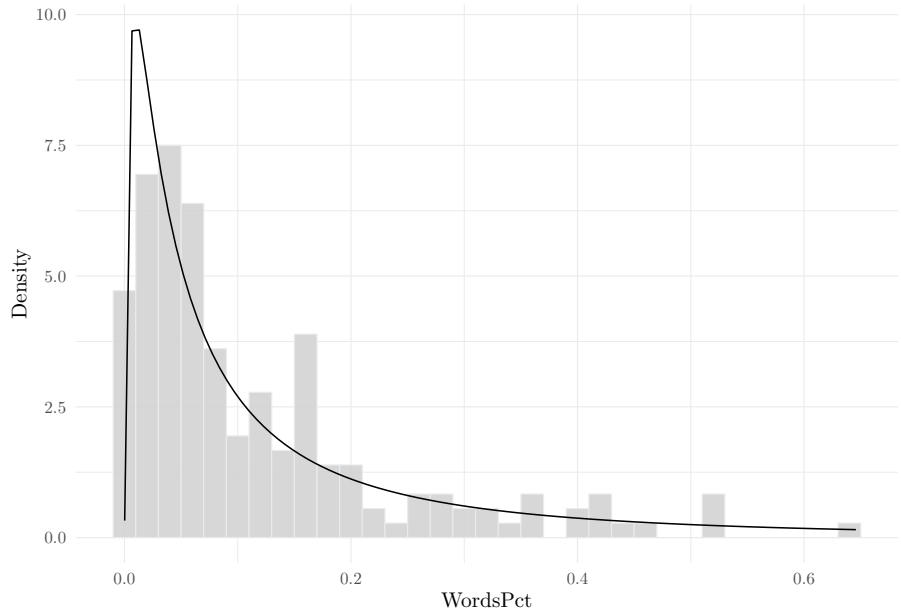
Table 10: *Relationships Between Character Categorization and Importance in Text*

Category Type	Category	Nb Characters	Pct of Words
Gender	Female	57	32.6
	Male	123	67.4
Age	Child	3	0.1
	Young	38	21.8
	Mid	102	46.6
	Old	37	31.5
Xenos	No	149	81.3
	Yes	31	18.7
Royal (Elite)	No	110	52.7
	Yes	70	47.3
Slave	No	158	91.4
	Yes	22	8.6
Divine	No	161	91.4
	Yes	19	8.6

In the way we have analyzed the importance given to characters in theater texts across the 5th century as a function of their social status, we cannot exclude the fact that certain types of characters would be mostly found in some plays, and the numbers in Table 10 would reflect the breakdown of these plays which concentrated on certain types of characters. At the scale of a play, the importance of the characters may be measured as the proportion of words they have, relative to all the words in that play. Figure 12 shows the distribution, across all characters, of their share of words overlaid with a fitted lognormal distribution. We can see that it is broadly distributed, and hence, at the level of each play, there is substantial mixing of characters of varied importance.

[48]We cannot however simply assert that the gods did not exist: for example, Herodotus (1.61.4) relates the story of Pisistratus arriving in Athens on a chariot with a young girl who introduced herself as Athena, and according to him many people thought they had seen the goddess.

Figure 12: *Distribution of the Share of Words Pronounced by Each Character in a Play*



Note: The horizontal axis is the number of words pronounced by each character relative to all words in the play. The curve is a moments-fitted lognormal distribution.

4 Sources on Relationships Represented as Networks

The use of network theory by historians has expanded over the years[49], and has given rise to renewed epistemological debates around which kinds of historical data could be conceived of as networks. Network analysis in historiography is usually carried out in a descriptive way, to account for a set of relationships extracted from a given document corpus.

The application of network analyzes to literary texts, and to classical texts in particular, has also given rise to several publications, but they tended to remain the domain of physicists or computer scientists. For example, Kydros, Notopoulos, and Exarchos (2015) looked at the networks between characters in mythology, and in their book, Kenna, MacCarron, and MacCarron (2017) discuss the analysis of networks stemming from various mythological and historical texts. With a less quantitative but more illustrative logic, Rydberg-Cox (2011) was interested in networks in the specific context of the Greek tragedy as well as their visualization. Finally, the works of Waumans, Nicodème, and Bersini (2015), Rochat (2014) and Elson, Dames, and McKeown (2010) provide a general perspective on the use of networks in literature. More specifically in theater, Fischer, Frank et al. (2019) undertook the analysis of plays across languages using network

[49]See the discussions of this evolution in Lemercier (2005) and Lemercier (2012), and more recently in Karila-Cohen et al. (2018).

techniques, with the DraCor project[50]. In all these approaches, however, the precise nature of the links that put individuals (characters in a play, or historical characters in a narrative) in a relationship is not well determined.

In this section, we concern ourselves with some ancient Greek data which more naturally represents relations, and have therefore been tackled by historians as networks, in a few instances at least. These particular data are mostly compiled from inscriptions: they are the Lexikon of Greek Personal Names, and the Mapping Ancient Polytheisms database. Both rely on arduous work carried out by specialists, having sifted through hundreds of thousands of inscriptions in order to produce the resulting electronic data.

4.1 Onomastics and Prosopography: LGPN

In ancient Greece, something as fundamental as naming a baby worked in a very different fashion from what we are familiar with today: one would only get a single anthroponym, a unique name. To this unique name could be associated a patronym. In some cases, a person could acquire a nickname through their actions or their physical aspect, but the single name remained a core principle. The choice of a name by the parents carried meaning, and could reflect through etymological links the child's belonging to the broad family. Giving to a child the name of a grand-parent, papponymy, was quite common. Names were often formed through derivational morphology: using one or two nouns or adjectives to make a name. In the study of ancient history, the analysis of names, onomastics, often goes hand in hand with prosopography, that is the gathering of all available historical information on particular individuals. For persons who have been considered as important in historiography, one can usually get sizable volumes of information. For less well-known people, it is a complex historical enquiry[51].

The available sources, typically epigraphic and at times partially damaged, give anthroponyms with sometimes a patronym. In order to go from lists of names to the notion of individuals, it is necessary to relate these inscriptions to each other, knowing that chronological information is imprecise most of the time. Prosopographers rely on instances of identical, or closely related names in given geographic area so that they can transform these names into references to particular individuals. A systematic perspective on how to carry out this research was proposed in Bresson (1981), who suggested the use of certain family relationships in a network context, combined with the naming conventions that appeared to be most common, in order to build a

[50]See Fischer, Frank, Börner, Ingo, and Göbel, Mathias (2019).

[51]See for example Karila-Cohen (2016), who describes the manner in which names and persons can be mapped.

family tree. This logic initially applied to Rhodes has been generalized[52]. In his his seminal study, he made it a central hypothesis that the practices observed in Rhodes in the modern period and up to contemporary times[53] obeyed similar rules to those practiced by the Ancients. He showed among other things the way in which papponymy was applied, with the names from the father and mother's sides alternating. Using this historical information, one can therefore create onomastic networks: someone with such name had a child with such name. Then, with prosopographic work, one can create prosopographic networks: such person had such person as a child. There is more onomastic data available than prosopographic, and the prosopographic reconstruction relies on many assumptions.

Large volumes of onomastic and prosopographic data have been made electronically available through the Lexicon of Greek Personal Names (LGPN)[54] housed by Oxford University. The project started in the early 1970s, and has led to the publication of references to hundreds of thousands of names. The statement of purpose from this effort is clear:

> To collect and publish with documentation all known ancient Greek personal names (including non-Greek names recorded in Greek, and Greek names in Latin), drawn from all available sources (literature, inscriptions, graffiti, papyri, coins, vases and other artifacts), within the period from the earliest Greek written records down to, approximately, the sixth century A.D.

Certain names have been excluded by the editors, as they presumably did not exist in reality, did not fit within the chronological bounds, or did not fit in the definition of a personal name[55].

The electronic interface to the LGPN is not designed for its data to be processed and analyzed in bulk. It is designed as a tool to query a name or a name root, and observe its occurrences. Figure 13 shows the results of query for a name. This feminine name, Αβα, appears 20 times and for each one of those there is a separate identification (the ID), because each entry may be considered a separate individual. Information is provided on the volume and publication from which this observation is pulled. Some chronological information is given[56], itself typically derived from the epigraphic analysis of the inscription where the name was observed. The reference gives an

[52] See for example Karila-Cohen (2018), who provides detailed examples applied to the aristocracy in an Athenian deme, and addresses the construction of the onomastic and prosopographic networks.

[53] And well documented by anthropologists and sociologists; see Vernier (1980) for example.

[54] See Parker, Yon, and Depauw (1996).

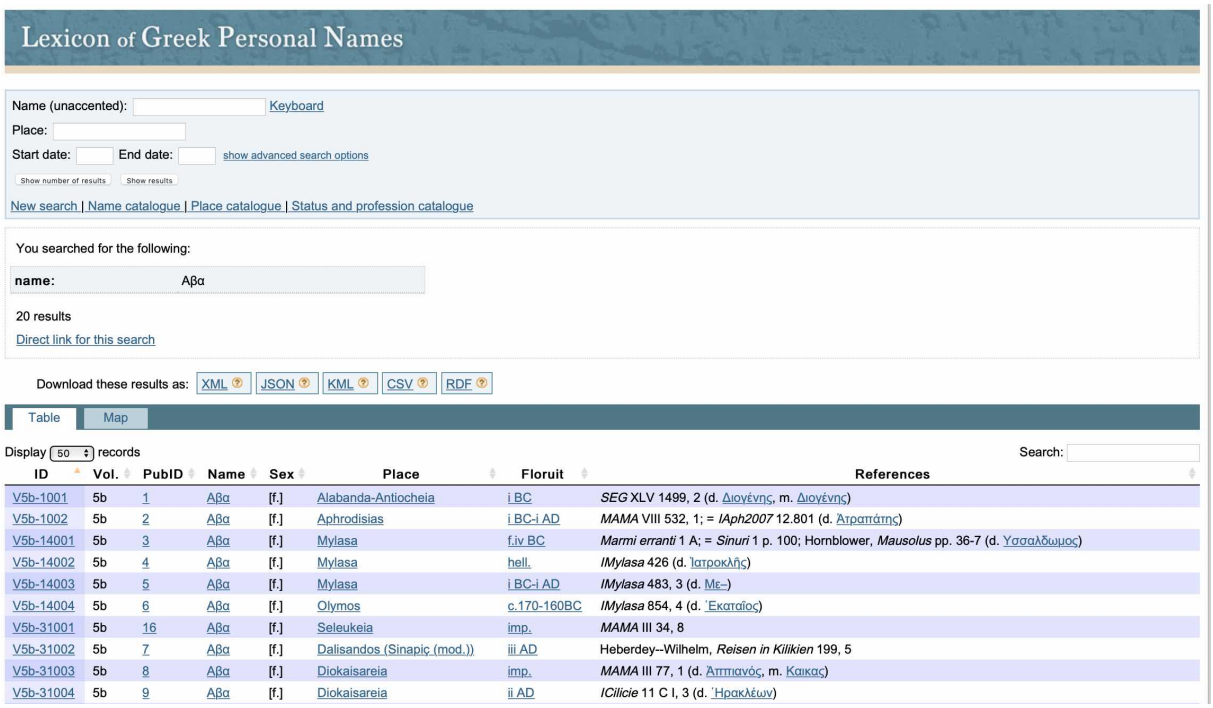
[55] Such as mythological and certain heroic names, Mycenaean names, later Byzantine names and geographical names.

[56] For example on the first line "i BC-i AD" meaning within a century before or after CE.

inscription and a line number[57]. Finally, the reference field shows the name of the mother and the daughter of this person[58].

One can also see from Figure 13 that although it is possible to obtain the data in various formats, this data only pertains to the multiple occurrences of a single name. It is hence impossible to simply download the entire data in a structured form in one batch; it has to be reconstructed from the data pertaining to each possible name. In addition, the various file formats that are illustrated in the figure do not all contain the same information. In particular, the file in `CSV` format contains the core individual and name data, while the `XML` file contains data on relationships, locations, and bibliographical references. Both data files hence need to be extracted each time.

Figure 13: *Query for a Name in the LGPN*



Using the package `RSelenium`[59], and after substantial parallel processing, we created a centralized dataset containing all these elements, in a structured fashion in the sense that all the data is stacked together in unique data frames, and so that it may be joined with additional data sources, such as the POLIS database through geographic information. We found close to 40,000

[57]For example, on the first line, "SEG XLV 1499, 2" refers to the *Supplementum Epigraphicum Graecum* volume 45:1499, a white marble fragment broken at right; built into the wall of a private house in Doğanyurt Köyü according to Chaniotis et al. (1995).

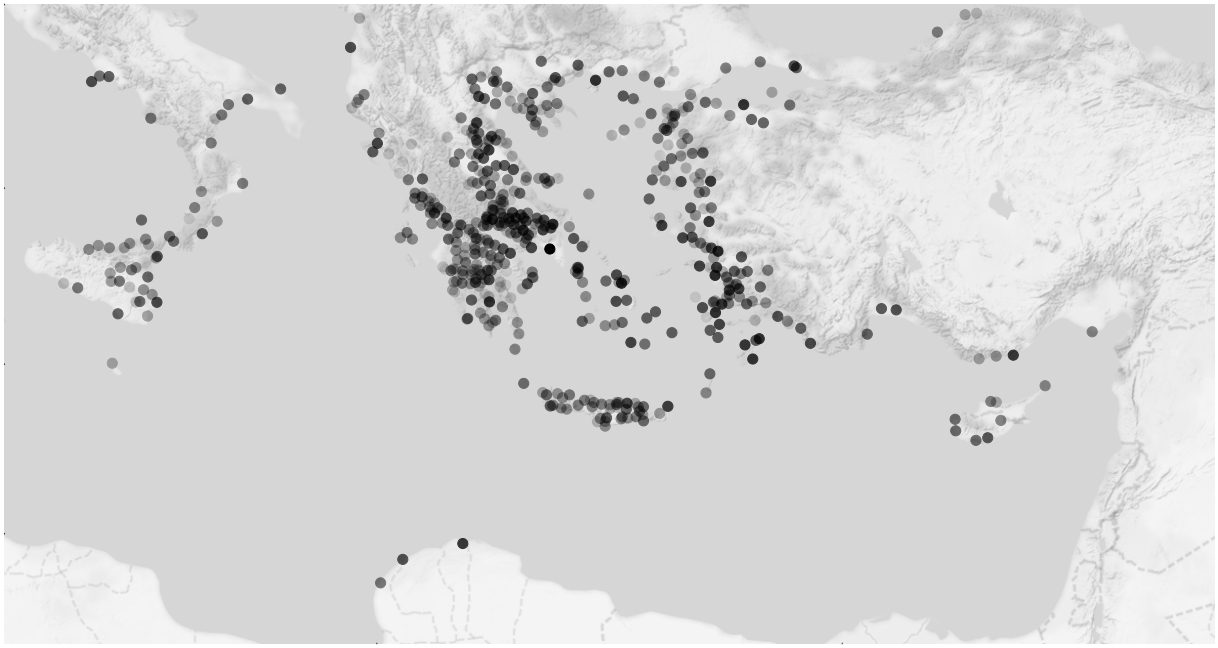
[58]In the case at hand, the inscription reported in Chaniotis et al. (1995) does point to the name of the person's mother, but this may not always be the case, and the family relationship may have been extrapolated through a prosopographic analysis.

[59]See Harrison and Kim (2020).

unique names spread across about 350,000 individual entries. The relationship table includes approximately 250,000 links.

Ancient location information in the case of inscriptions does raise some issues: there is a distinction between where an inscription is found and where the people that it refers to used to live, a piece of information that may not be explicit in the source, and when ancient locations are referred to in a source, it is not necessarily clear where exactly that location is[60]. Keeping these issues in mind, we mapped the location information from the LGPN to that of the POLIS catalog first based on names and then based on coordinates (using a distance threshold in terms of degrees of latitude and longitude). Figure 14 shows the geographic extent in the Greek world of the LGPN name data. The density presumably reflects the distribution of the population, but also the extent to which people were used to writing inscriptions.

Figure 14: *Geographic Distribution of LGPN Data*



Note: The map background shows modern political boundaries. The dots show the *poleis* with geographic coordinates for which the LGPN database provided some information on names, with opaqueness a function of the logarithm of the number of observations.

Across the Greek world, the distribution of names was also not strongly concentrated, in particular

[60]Consider, for example, the uncertainty surrounding the location of Troy.

for women. Table 11 displays some aggregate information on the name data. Women were essentially excluded from political life, and inscriptions very often reported official acts, and it is hence expected that the number of observations of inscriptions would be heavily biased towards men. Only looking at these ten most common names, it is interesting that the distribution of feminine names is much flatter than that for men: the number of occurrences for the tenth rank is more than half that of the first rank for women, but about a third for men. Out of the ten names, none of those for men are straight substantives; however for women, three are simple nouns (for Hope, Victory, and Fate).

Table 11: *Ten Most Common Man and Woman Names in LGPN*

Men Names	Men Number	Women Names	Women Number
Διονύσιος	4762	Ζωσίμη	317
Ἀπολλώνιος	4027	Ἐλπίς	304
Δημήτριος	3325	Στρατονίκη	261
Ἀλέξανδρος	2359	Νίκη	245
Ἀρτεμίδωρος	1702	Ἀφροδισία	211
Ἀπολλόδωρος	1425	Δημητρία	207
Ἀσκληπιάδης	1380	Ἀρτεμισία	205
Ζώσιμος	1373	Κλεοπάτρα	200
Φίλων	1362	Διονυσία	197
Θεόδωρος	1356	Τύχη	186

Looking at the distribution of names more systematically in logarithmic scale in Figure 15, we can confirm the intuitions from Table 11: the curve for women appears steeper than that for men, indicative of a flatter distribution. The overall shape, however, is not particularly straight, and resembles more that of an exponential distribution than a pure power law.

The distribution fits in 12 tell us that the distribution is more likely to be a truncated power law than an exponential, with a very small exponential parameter λ. Baek, Kiet, and Kim (2007) have shown that family names distributions across the world tended to follow power laws, but not always, and that the parameters were not constant from one country to the other. Specifically in the US, and over time, Li (2012) showed that a simple power law could not account for the distribution of (given) first names. The distribution of ancient Greek names, being neither family names or first names as we understand them today, could not be expected to match either empirical model, but can offer an interesting test case of generating mechanisms.

Figure 15: *Log/log Plot of Name Distributions for Men and Women Names*

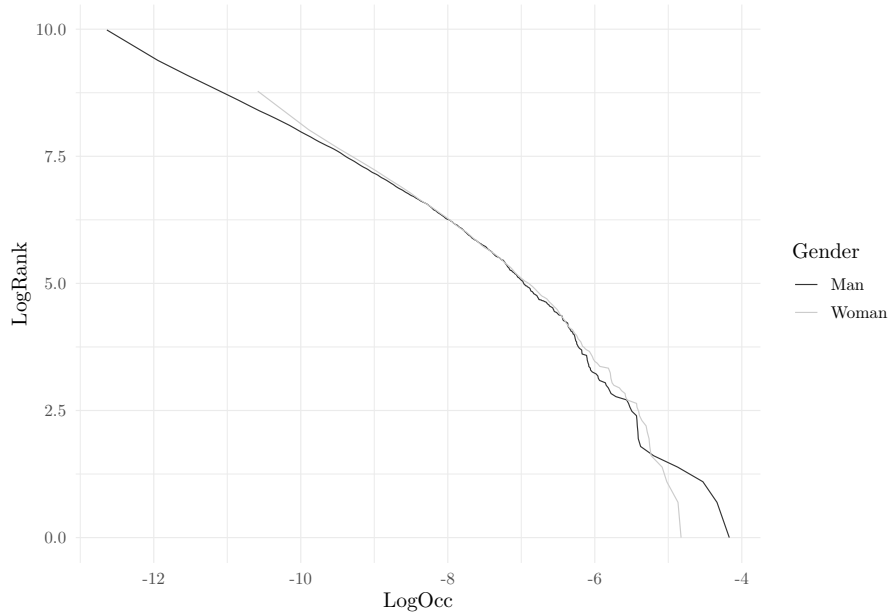


Table 12: *Summary Statistics For Name Distributions Fits by Gender*

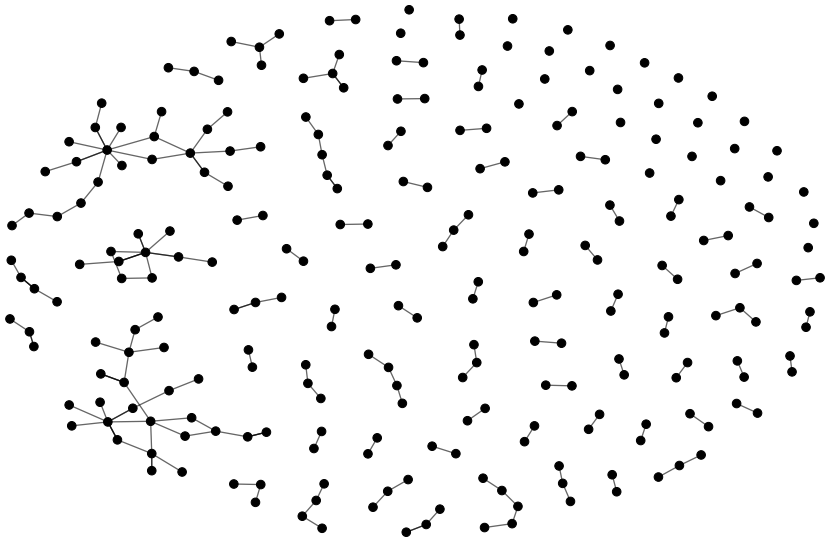
Statistic	Man	Woman
Lambda Exp	0.104	0.284
Alpha Pow	0.654	0.783
Alpha Trunc	0.747	0.946
Lambda Trunc	0.001	0.003
Trunc vs Pow R	18.134	12.879
Trunc vs Pow p	0.000	0.000
Trunc vs Exp R	33.355	22.523
Trunc vs Exp p	0.000	0.000
Pow vs Exp R	32.623	21.094
Pow vs Exp p	0.000	0.000

Using the parental relationships between names in a given location, we can build a network of these links. We implemented a comparable method as that laid out in Karila-Cohen (2018) (where she focused on particular demes in Athens), but applied it to the entire LGPN data. We constitute a network where the edges are unique names, and the vertices are parental relationship entries in the LGPN: the resulting network is hence onomastic from the perspective of the edges, and prosopographic from the perspective of the vertices. Restricting the LGPN entries to same demes, we find approximately the same nodes and edges as in Karila-Cohen (2018).

Figure 16 displays the network reduced to the demes Oion Kerameikon and Oion Dekeleikon in Athens. We can see that a majority of nodes are linked, and even for this small network there is wide range of components size.

Building all local networks at the *polis* level in the same manner, we can create a large network

Figure 16: *LGPN Network Map for Oion Demes*

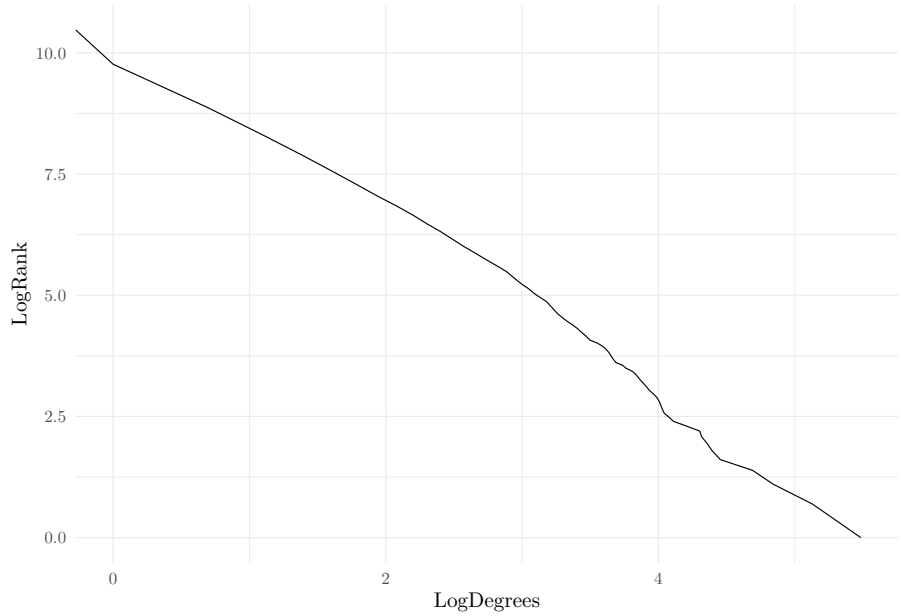


covering the entire dataset, for both men and women, and for the whole Greek world. Focusing on this large network, Figure 17 displays the degree distribution, that is, the distribution of the number of different names that have been chosen by a person with a given name, in a given location. The shape of the curve comes out very straight, which would clearly be indicative of a power law relationship.

The form of relationship is confirmed by the tests in Table 13: the number of distinct names chosen by each name is a fat-tailed power law. In a network growth process, a pure power law distribution for degrees can be related to preferential attachment[61]: new links are more likely to attach to the nodes as a function of their existing number of links. This would hence imply that, on a large scale, the more ancient Greeks had diversified the names of their progeny, the more they would continue to do so.

[61]See Barabási and Albert (1999).

Figure 17: *Degree Distribution of Name Networks across Poleis*



Note: Local parental networks are built only for locations with more than 1000 observations.

Table 13: *Summary Statistics on Distribution Fits for the Number of Degrees in the LGPN*

Statistic	LGPN
Lambda Exp	0.002
Alpha Pow	0.340
Alpha Trunc	0.322
Lambda Trunc	0.000
Trunc vs Pow R	3.431
Trunc vs Pow p	0.130
Trunc vs Exp R	7.472
Trunc vs Exp p	0.000
Pow vs Exp R	7.352
Pow vs Exp p	0.000

4.2 Polytheist God and Epiclese Network: MAP

The history of ancient religions has traditionally been written starting from the gods, generally considered as persons or personifications. This simplistic and static representation has failed to meet the challenges of the structural and dynamic complexity of religious systems in Antiquity. Understanding the gods as powers with multiple facets, as captured in the way they are addressed, it becomes possible to analyze the networks they engender and the way in which their environment shapes them.

The BDEG data we discussed earlier gathered data in a rather unstructured fashion on the way

ancient invocations addressed the gods, in addition to the god's name: the forms in that database contained one or two epithets in each case. On a large scale this data is useful in order to track the number of votive acts, but it is difficult to exploit if one wants to precisely understand the way in which the gods were invoked. The more recent large-scale MAP project[62] offers a much more precise and entirely revamped representation of the underlying data (mostly inscriptions), covering the Greek and Semitic ancient worlds, and allows for the close study of the formulas the Ancients used to address divine beings. To understand the relational logic which structures these divine powers, MAP takes into account a wide range of divine onomastic sequences, combinations of divine names or elements (names, epithets, titles, propositions), some shared by several gods, others specific.

Indeed, the simple act of performing a rite in ancient Greece involved addressing one or more gods with a complex series of qualifiers, the epicleses, ordered and expressed in a deliberate manner. These qualifiers, which were sometimes common between certain gods, effectively created a network between all the gods; and the gods also created a system of relations between the qualifiers. These onomastic sequences taken as a whole form a representation of the way in which the Greeks conceived of polytheism.

The detailed study of onomastic sequences has been going on for many years, as we alluded to earlier, and recent research in this area includes Brulé (1998), Brulé (2005), Bonnet and Belayche (2017), who establish various assumptions implicit in how onomastic sequences are constructed. The idea of putting these sequences within the framework of network analysis is at the basis of the MAP project, as Lebreton (2019) explains. Another corner stone is the extraction of onomastic formulas from the sources which, thanks to a particular syntax, can account for the great complexity of these inscriptions; this notion of formula is detailed in Bonnet and Lebreton (2019). Other research has addressed the names of gods more specifically in literature, such as Elwert, Gerhards, and Sellmer (2017) for example, but their literary sources do not reflect actual votive acts, only mentions of gods in mythological texts. The large scale use of networks to study religious invocations has also been experimented in other historical fields, such as Oriental research for example[63].

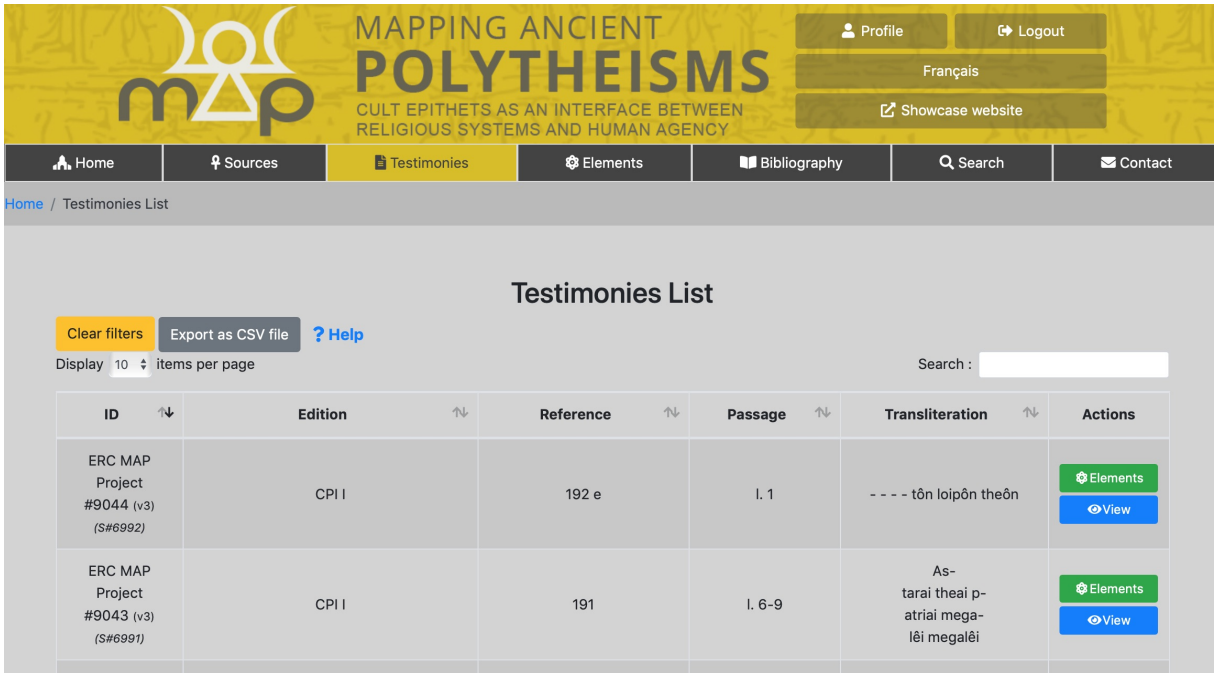
The data that constitutes MAP, established from the inscriptions by a team of specialists, is highly structured and accessible in bulk. Each material item, such as a stele or part of a monument for example, is an entry in a source data table, with detailed information such as its publication

[62]Mapping Ancient Polytheisms, see Bonnet (2017).

[63]See Alstola et al. (2019).

or location. A source contains one or several testimonies, that is addresses to the gods, each entered in a testimony table. Figure 18 shows the testimony query interface on the MAP website. Each testimony, usually in the form of a sentence[64] referring to divinities and qualifiers. These sentences are converted to formulaic expressions akin to mathematical formulas expressing the links between the various elements in the testimony. These elements, mostly adjectives, nouns or god names, are also centralized in a specific element table.

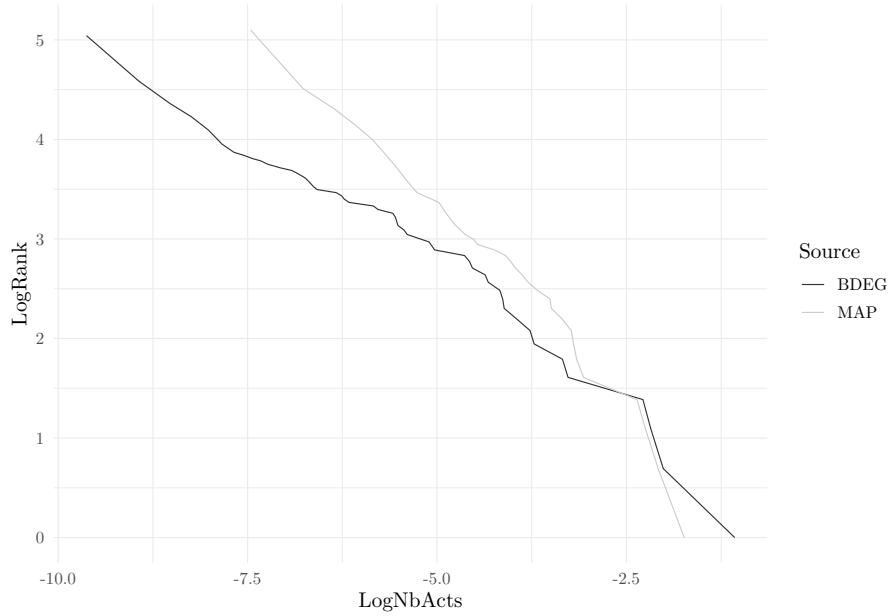
Figure 18: *MAP Testimony List Page*



The MAP data, currently with close to 4,000 testimonies, is still growing and does not yet cover the entire geographic space of the Greek world; in that respect it contains less information than the BDEG. Nevertheless, the distribution of mentions of each given god across the entire dataset (restricted to Greek) has a very similar shape to that we obtained from the BDEG, as Figure 19 shows. The difference in the slope of the curve, whereby the MAP data is more fat-tailed, is presumably due to the fact that there are for the time being less individual god entries in that dataset than in the BDEG.

[64]The sentence may be in ancient Greek, but also in many other languages such as Phoenician, Etruscan, Assyrian, etc. Although the extent of the MAP endeavor also includes the Semitic world, we will for our purposes only consider its entries as far as they are in Greek, and pertain to the Ancient Greek world.

Figure 19: *Log/Log Cumulative Distribution of the Distribution of Votive Acts Among Gods*



In order to understand how to construct a network that captures the semantic relationships between gods, it is necessary to examine the logic of onomastic formulas. One example of a testimony would be number 100, from Egypt, in the MAP database, from an inscription, stating:

> παρὰ
> τῶν κυρίων θεῶν Πριω τοῦ θεοῦ
> μεγίστου καὶ Ὡρεγέβθιος καὶ Ἴσιδος
> Ῥεσακέμεως καὶ οἱ (σις) σὺν αὐτοῖ (σις) θεῶν
> μεγίστων[65]

This dedication is encoded as a formula referencing elements like the gods (Prion or Isis) or qualifications (Great), as well as the connections between them (such as "and", for example). In this case, this is recorded as the following formula:

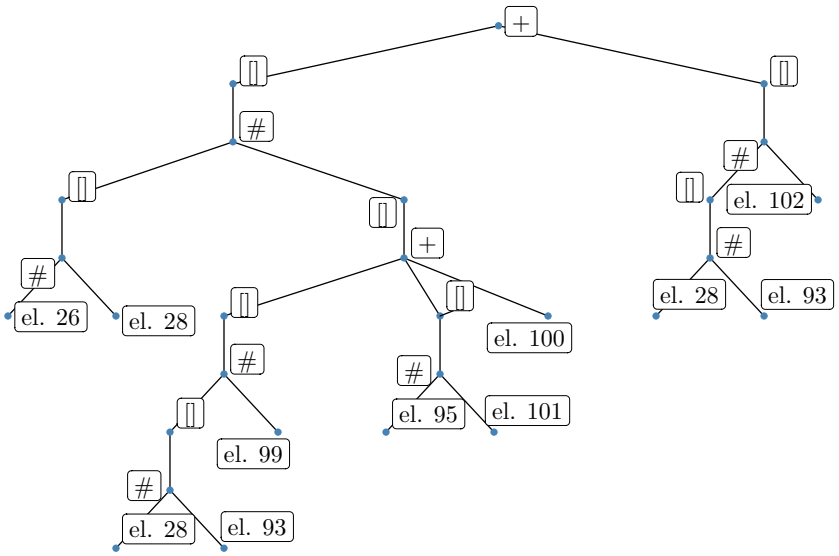
```
[[{26}#{28}]#[[{99}#[{28}#{93}]]+{100}+[{95}#{101}]]]+[{102}#[{28}#{93}]]
```

where the numbers are the identifications of particular elements, and the operators are the MAP researchers' interpretation of the logic of the text. All these formulae can be systematically converted into tree network representations, as shown in Figure 20 for the case at hand. Any operator-based expression could in fact be converted to a tree representation.

In order to better seize the nature of the links created between the elements and the gods, one

[65] "Beside the Lords Gods Priô the Greatest God and Horegebthis and Isis Rhesakemis and the Greatest Gods Who are with them" (MAP translation).

Figure 20: *Network Representation of Testimony #100*



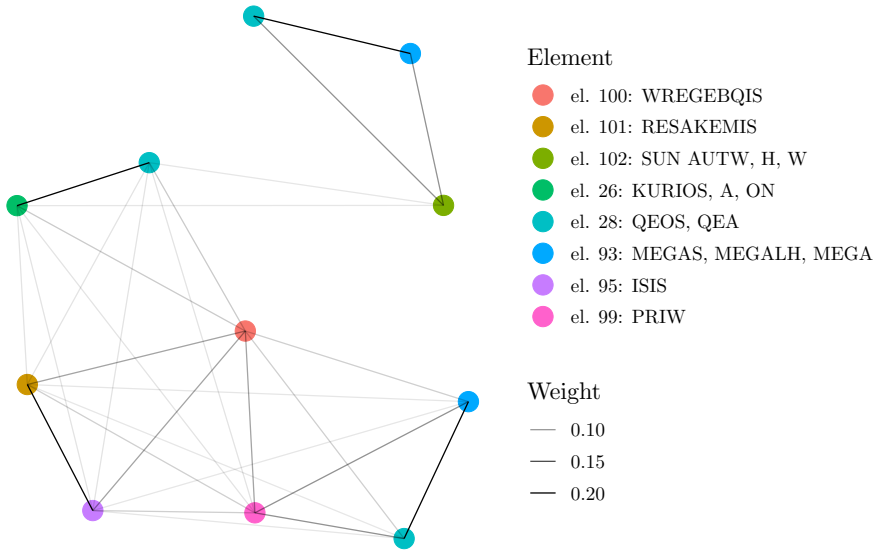
Note: Every node represents either an element or an operator and the links represent the application of an operator to one or several elements.

approach is to allocate a certain distance to all the links in Figure 20 depending on the operators, so that the edge connecting two entities related by "and", for example, would have a lesser distance than a simple apposition. Then, the total distance from any entity to any other entity in the formula can be computed, reflecting these differences among the links. Finally, the strength of the connection between two entities, which can be used as a weight in the graph connecting all the entities in a formula, can be chosen as the inverse of the distance. Figure 21 shows a representation of the resulting graph for Testimony 100, where only the links above a certain weight are shown. This way of approaching onomastic formulas through the networks they can generate is inspired from quantitative linguistics[66].

One can build the large disconnected network generated by an entity: the set of all the strength-weighted network representations of onomastic formulas where the entity appears. Picking a god who does not appear very often, Helios, we can plot all these components: see Figure 22. We can see that there is a good degree of dispersion in the number of elements in each component of this network, and in the number of other entities the god is strongly linked to.

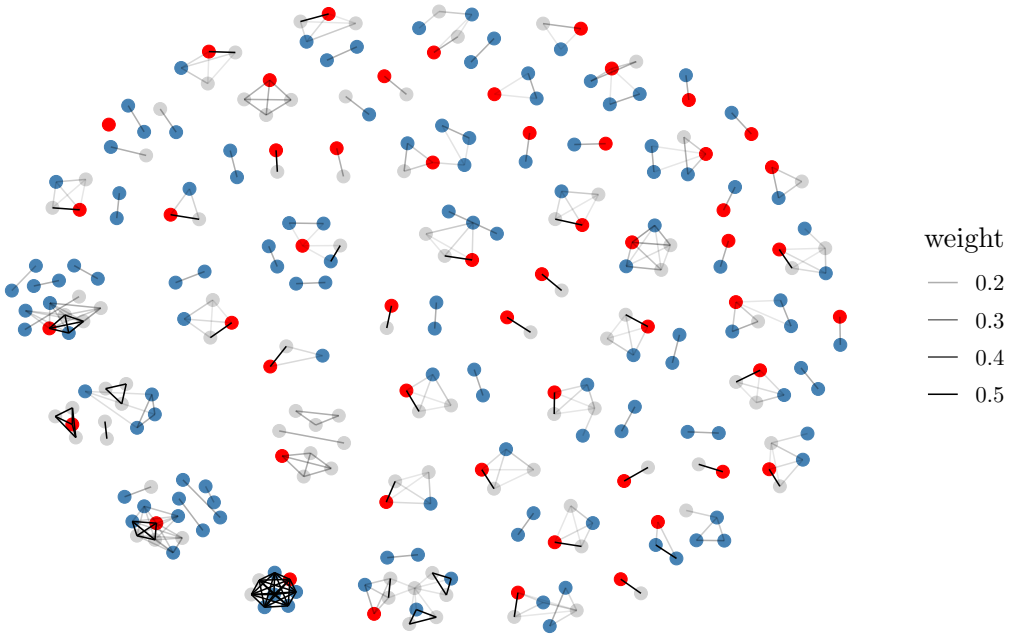
[66]See, for example, Mehler et al. (2016).

Figure 21: *Network Representation of Testimony #100 Using Distances*



Note: All the links between the nodes with a weight above 0.05 are represented, and they are weighed as a function of the formulaic distance: closer relationships are marked by denser lines. The nodes representing entities are in grey, and qualifications in blue.

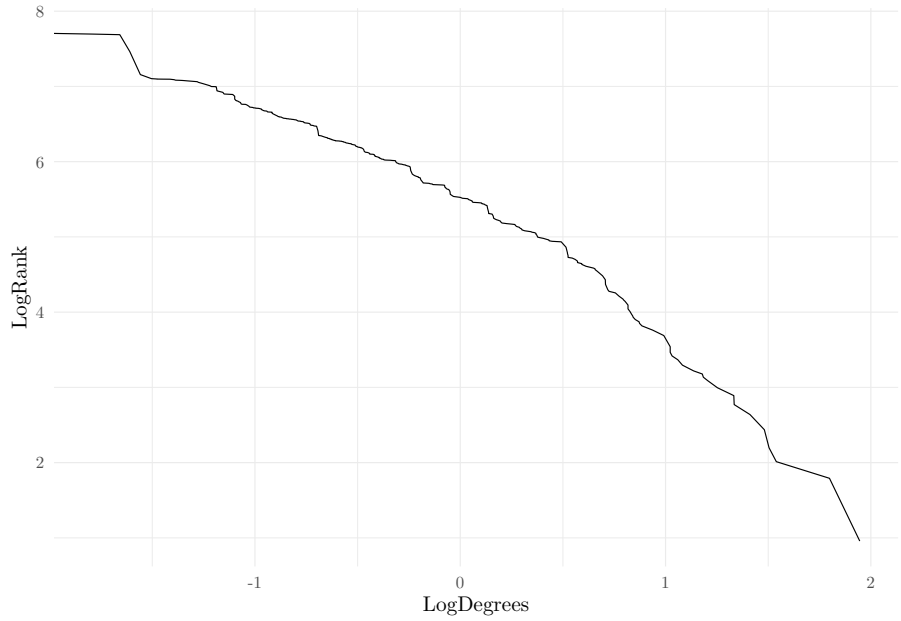
Figure 22: *Network Representation of All Dedications Involving Helios*



Note: All the links between the nodes with a weight above 0.1 are represented, and they are weighed as a function of the formulaic distance: closer relationships are marked by denser lines. The nodes representing Helios are in red, other entities in grey, and qualifications in blue.

Looking at all the gods in all testimonies, we can focus on the weighted degree distribution, as shown in 23, which represents the strength with which each occurrence of each god in invocations is related to other gods. This degree distribution does not appear as a straight line and it is therefore presumably not a simple power law, and not generated by a basic preferential attachment mechanism.

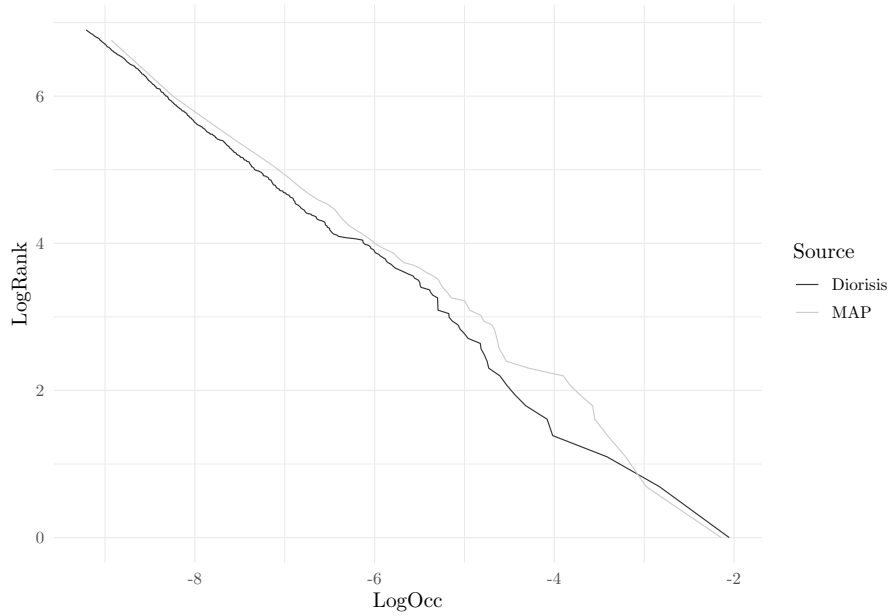
Figure 23: *Degree Distribution of Divinities in Distance-Weighted Graph*



One interesting question is whether the language used in invocations possesses characteristics of a natural language. Many detailed analyzes can be carried out, in particular when reflecting the network structure of syntactic relationships[67], but we restrict ourselves to a simple comparison, looking at the distribution of terms. Elements, which correspond to common names or proper names, are the natural equivalent of lemmas in POS tagging. Figure 24 plots the occurrence distribution for Diorisis and for the onomastic formulas in MAP, and they appear quite close to each other.

[67] See Čech, Mačutek, and Liu (2016) for example.

Figure 24: *Log/Log Cumulative Distribution of Diorisis and MAP Corpora*



Note: The data includes the most common lemmas for each corpus, which account for more than 0.01% of occurrences. The horizontal axis is the logarithm of normalized frequency of each lemma, and the vertical axis is the logarithm of the lemma's rank.

The statistical fits in Table 14 confirm this impression, with power law coefficients that are almost identical to each other. In spite of a very formulaic nature, the term distribution of divine invocations resembles that of natural language superficially; this may not hold, however, if one were to look more closely at the syntactic structure of both languages.

Table 14: *Summary Statistics on Distribution Fits*

Statistic	MAP	Diorisis
Lambda Exp	0.190	0.001
Alpha Pow	0.777	0.774
Alpha Trunc	0.968	0.746
Lambda Trunc	0.000	0.000
Trunc vs Pow R	3.935	1.962
Trunc vs Pow p	0.000	0.000
Trunc vs Exp R	6.690	7.321
Trunc vs Exp p	0.000	0.000
Pow vs Exp R	6.472	7.256
Pow vs Exp p	0.000	0.000

Note: The data excludes words with less than 100 occurrences in the Diorisis corpus

5 Conclusion: the Benefits of Distributional Approaches

Before taking a close look at some statistical methods and at a series of primary sources on ancient Greek society, we stressed that this endeavor attempted to combine the clio and the metrics. The historiographical aspect of our approach emanates to a large extent from the fact we remained focused on primary sources and avoided historical or literary information that would have been remote from these sources.

We tackled measurement in a different fashion from common practice. In quantitative history, analyzes tend to be descriptive or aggregative and question the content of the data more than its shape. In traditional cliometrics, econometric methods are more advanced, especially as they pertain to time series, but they are exclusively applied to the economy. Instead, leaving aside the chronological aspects of serial data, we adopted a complex systems perspective and concentrated on distributional aspects.

Having discussed how to create appropriate large-scale datasets for this approach, we pointed out certain patterns in the probabilistic outcomes visible in the data. We showed how the literary tradition, recovered inscriptions, and addresses to the divine all shared certain linguistic features. We also found that the distribution of names in the Greek world had specific features, potentially differentiating it from those of modern first or family names. In ancient Greek polytheism, we saw that many gods received attention, but that the extent of that attention also had a probabilistic structure to it. Theater texts from the 5th century showed us that the importance of characters appeared to exacerbate the presence of certain authoritative social groups.

All these points raise the natural follow-up question of why it is the case, and it is at this juncture that economics could enter back into cliometrics, when they may help the historians comprehend in the sense of Aron, as opposed to explain[68], the generating mechanisms behind all these striking patterns.

[68]See Aron (1981).

References

Alfarano, Simone, and Thomas Lux. 2010. "Extreme Value Theory as a Theoretical Background for Power Law Behavior." 1648. *Kiel Working Papers*. Kiel Working Papers. Kiel Institute for the World Economy. https://ideas.repec.org/p/kie/kieliw/1648.html.

Alstola, Tero, Shana Zaia, Aleksi Sahala, Heidi Jauhiainen, Saana Svärd, and Krister Lindén. 2019. "Aššur and His Friends: A Statistical Analysis of Neo-Assyrian Texts." *Journal of Cuneiform Studies* 71 (January): 159–180.

Alstott, Jeff, Ed Bullmore, and Dietmar Plenz. 2014. "Powerlaw: A Python Package for Analysis of Heavy-Tailed Distributions." *PLoS ONE* 9 (1): e85777.

Aron, Raymond. 1981. "Quelques remarques sur la compréhension et l'explication." *Revue européenne des sciences sociales* 19 (54/55): 71–82.

Assael, Yannis, Thea Sommerschield, and Jonathan Prag. 2019. "Restoring Ancient Text Using Deep Learning: A Case Study on Greek Epigraphy." In, 6369–6376. Hong Kong: EMNLP.

———. 2020. "Sommerschield/Ancient-Text-Restoration." June 6, 2020. https://github.com/sommerschield/ancient-text-restoration.

Azoulay, Vincent. 2014. "Repenser le politique en Grèce ancienne." *Annales HSS* 69 (3): 605–626.

Baek, Seung Ki, Sebastian Bernhardsson, and Petter Minnhagen. 2011. "Zipf's Law Unzipped." *New Journal of Physics* 13 (4): 043004.

Baek, Seung Ki, Hoang Anh Tuan Kiet, and Beom Jun Kim. 2007. "Family Name Distributions: Master Equation Approach." *Physical Review E* 76 (4): 046113.

Barabási, Albert-László, and Réka Albert. 1999. "Emergence of Scaling in Random Networks." *Science* 286 (5439): 509–512.

Bonnard, Jean-Baptiste, Véronique Dasen, and Jérôme Wilgaux. 2017. *Famille et société dans le monde grec et en Italie du V^{e} au II^{e} siècle av. J.-C.* Didact Histoire. Rennes: Presses Universitaires de Rennes.

Bonnet, Corinne. 2017. "Mapping Ancient Polytheisms." 2017. https://map-polytheisms.huma-num.fr/.

Bonnet, Corinne, and Nicole Belayche, eds. 2017. *Puissances divines à l'épreuve du comparatisme : constructions, variations et réseaux relationnels*. Bibliothèque de l'École des Hautes Études, sciences religieuses, volume 175. Turnhout: Brepols.

Bonnet, Corinne, and Sylvain Lebreton. 2019. "Mettre les polythéismes en formules ? À propos de la base de données Mapping Ancient Polytheisms." *Kernos* 32 (December): 267–296.

Bresson, Alain. 1981. "Règles de nomination dans la Rhodes antique." *Dialogues d'histoire ancienne* 7 (1): 345–362.

Brulé, Pierre. 1998. "Le langage des épiclèses dans le polythéisme hellénique (l'exemple de quelques divinités féminines). Quelques pistes de recherches." *Kernos* 11 (11): 13–34.

———. 2005. "Le polythéisme en transformation : les listes de dieux dans les serments internationaux en Grèce antique (V^{e}-II^{e} siècles)." In *Nommer les dieux : Théonymes, épithètes, épiclèses dans l'Antiquité*, edited by Nicole Belayche, Pierre Brulé, Gérard Freyburger, Yves Lehman, Laurent Pernot, and Antoine Prost, 143–173. Recherches sur les rhétoriques religieuses. Turnhout: Brepols Publishers.

Brulé, Pierre, and Sylvain Lebreton. 2007. "La Banque de données sur les épiclèses divines (BDDE) du Crescam : sa philosophie." *Kernos* 20: 217–228.

Celano, Giuseppe G. A., Gregory Crane, and Bridget Almas. 2020. "PerseusDL/Treebank_data." August 12, 2020. https://github.com/PerseusDL/treebank_data.

Celano, Giuseppe G. A., Gregory Crane, and Saeed Majidi. 2016. "Part of Speech Tagging for Ancient Greek." *Open Linguistics* 2 (1): 393–399.

Chaniotis, A., H. W. Pleket, R. S. Stroud, and J. H. M. Strubbe. 1995. "SEG 45-1499: Alabanda, Honorary Inscription for Aba, Late Hellenistic Period." *Supplementum Epigraphicum Graecum*, January.

Citti, Vittorio. 1988. "The Ideology of Metics in Attic Tragedy." In *Forms of Control and Subordination in Antiquity*, edited by Tōru Yuge and Masaoki Doi, 456–464. Tokyo ; Leiden: Society for Studies on Resistance Movements in Antiquity ; Brill.

Clauset, Aaron, Cosma Rohilla Shalizi, and M. E. J. Newman. 2009. "Power-Law Distributions in Empirical Data." *SIAM Review* 51 (4): 661–703.

Corvisier, Jean-Nicolas. 1986. "La vieillesse en Grèce ancienne d'Homère à l'époque hellénistique." *Annales de démographie historique* 1985 (1): 53–70.

Crane, Gregory R. 2012. "Perseus Digital Library." April 2012. http://www.perseus.tufts.edu.

Čech, Radek, Ján Mačutek, and Haitao Liu. 2016. "Syntactic Complex Networks and Their Applications." In *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, edited by Alexander Mehler, Andy Lücking, Sven Banisch, Philippe Blanchard, and Barbara Job, 167–186. Understanding Complex Systems. Berlin, Heidelberg: Springer.

Diebolt, Claude, and Michael Haupert, eds. 2019. *Handbook of Cliometrics*. 2nd ed. Springer Reference. Cham: Springer International Publishing.

———. 2020. "How Cliometrics Has Infiltrated Economics – and Helped to Improve the Discipline." *Annals of the Fondazione Luigi Einaudi* 54 (1): 219–230.

Eckhoff, Hanne, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Haug, Odd Einar Haugen, and Marius Jøhndal. 2018. "The PROIEL Treebank Family: A Standard for Early Attestations of Indo-European Languages." *Language Resources and Evaluation* 52 (1): 29–65.

Elson, David, Nicholas Dames, and Kathleen McKeown. 2010. "Extracting Social Networks from Literary Fiction." In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 138–147. Uppsala, Sweden: Association for Computational Linguistics.

Elwert, Frederik, Simone Gerhards, and Sven Sellmer. 2017. "Gods, Graves and Graphs – Social and Semantic Network Analysis Based on Ancient Egyptian and Indian Corpora." *Digital Classics Online* 3 (2): 124–137.

Fenoaltea, Stefano. 2019. "Spleen: The Failures of the Cliometric School." *Annals of the Fondazione Luigi Einaudi* 53 (2): 5–24.

Fischer, Frank, Börner, Ingo, and Göbel, Mathias. 2019. "DraCor – Drama Corpora Project." 2019. https://dracor.org.

Fischer, Frank, Börner, Ingo, Göbel, Mathias, Hechtl, Angelika, Kittel, Christopher, Milling, Carsten, and Trilcke, Peer. 2019. "Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama." In *Proceedings of DH2019: "Complexities"*, 1–6. Utrecht: Zenodo.

Fleck, Robert K., and F. Andrew Hanssen. 2018. "What Can Data Drawn from the Hansen-Nielsen Inventory Tell Us About Political Transitions in Ancient Greece?" In *Ancient Greek History and Contemporary Social Science*, 213–238. Edinburgh University Press.

Gabaix, Xavier. 1999. "Zipf's Law for Cities: An Explanation." *The Quarterly Journal of Economics* 114 (3): 739–767.

———. 2016. "Power Laws in Economics: An Introduction." *Journal of Economic Perspectives* 30 (1): 185–206.

Gauthier, Laurent. 2021. "Putting Clio Back in Cliometrics." Working Paper hal-03289608, to appear in *History & Theory* in June 2022. HAL. Paris: Université Paris 8. https://hal-univ-paris8.archives-ouvertes.fr/hal-03289608.

Grabska-Gradzińska, Iwona, Andrzej Kulig, Jarosław Kwapień, and Stanisław Drożdż. 2012. "Complex Network Analysis of Literary and Scientific Texts." *International Journal of Modern Physics C* 23 (07): 1250051/1–15.

Hansen, Mogens Herman, and Thomas Heine Nielsen. 2004. *An Inventory of Archaic and Classical Poleis*. Oxford ; New York: Oxford University Press.

Harrison, John, and Ju Yeong Kim. 2020. *RSelenium: R Bindings for 'Selenium WebDriver'* (version 1.7.7).

Harsch, Ulrich. n.d. "Bibliotheca Augustana." Accessed September 20, 2020. http://www.hs-augsburg.de/~harsch/augustana.html#gr.

Haug, Dag T T, Marius L Jøhndal, Hanne M Eckhoff, Mari J B Hertzenberg, and Angelika Müth. 2009. "Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages." *Traitement Automatique Des Langues* 50 (2): 17–45.

Hobson, Matthew S. 2014. "A Historiography of the Study of the Roman Economy: Economic Growth, Development, and Neoliberalism." *Theoretical Roman Archaeology Journal* 0 (2013): 11–26.

Ide, Nancy. 2004. "Preparation and Analysis of Linguistic Corpora." In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 289–305. Malden, MA, USA: Blackwell Publishing Ltd.

Jackson, Matthew O., and Brian W. Rogers. 2007. "Meeting Strangers and Friends of Friends: How Random Are Social Networks?" *American Economic Review* 97 (3): 890–915.

Johnson, Kyle P., Patrick J. Burns, John Stewart, and Todd Cook. 2019. "The Classical Language Toolkit." January 2019. http://cltk.org.

Johnson, Tim, and Josiah Ober. 2014. "POLIS." 2014. http://polis.stanford.edu.

Karila-Cohen, Karine. 2016. "Prosopographia Attica 2.0 : base de données et raisonnement prosopographique." *Revue historique* 680 (4): 869–904.

———. 2018. "Le graphe, la trace et les fragments : l'apport des méthodes quantitatives et des outils numériques à l'étude des élites civiques athéniennes." *Annales HSS* 73 (4): 785–815.

Karila-Cohen, Karine, Claire Lemercier, Isabelle Rosé, and Claire Zalc. 2018. "Nouvelles cuisines de l'histoire quantitative." *Annales HSS* 73 (4): 773–783.

Kenna, Ralph, Máirín MacCarron, and Pádraig MacCarron, eds. 2017. *Maths Meets Myths: Quantitative Approaches to Ancient Narratives*. Understanding Complex Systems. Cham: Springer International Publishing.

Kydros, Dimitrios, Panagiotis Notopoulos, and Georgios Exarchos. 2015. "Homer's Iliad – A Social Network Analytic Approach." *International Journal of Humanities and Arts Computing* 9 (1): 115–132.

Lebreton, Sylvain. 2019. "Des dieux et des personnes : une approche par les réseaux (mondes grecs et ouest-sémitiques, ca. 1000 av. n. è. – 400 d. n. è.)." In *La personne en question dans les réseaux*.

Lebreton, Sylvain, Jean-Baptiste Barreau, Karine Karila-Cohen, and Pierre Brulé. 2014. "Banque de Données des Épiclèses Grecques (BDEG)." 2014. https://epiclesesgrecques.univ-rennes1.fr/?lang=en.

Lemercier, Claire. 2005. "Analyse de réseaux et histoire." *Revue d'histoire moderne contemporaine* 52-2 (2): 88–112.

———. 2012. "Formale Methoden der Netzwerkanalyse in den Geschichtswissenschaften: Warum und Wie?" *Österreichische Zeitschrift für Geschichtswissenschaften* 23 (1): 16–41.

Li, Wentian. 2012. "Analyses of Baby Name Popularity Distribution in U.S. For the Last 131 Years." *Complexity* 18 (1): 44–50.

Mansouri, Saber. 2011. *Athènes vue par ses métèques (V^{e}-IV^{e} siècle av. J.-C.)*. Paris: Tallandier.

Mehler, Alexander, Andy Lücking, Sven Banisch, Philippe Blanchard, and Barbara Job, eds. 2016. *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*. Understanding Complex Systems. Berlin, Heidelberg: Springer Berlin Heidelberg.

Mitzenmacher, Michael. 2004. "A Brief History of Generative Models for Power Law and Lognormal Distributions." *Internet Mathematics* 1 (2): 226–251.

Mossé, Claude. 2011. "Peut-on Parler de "Classes" à Propos Des Sociétés Du Monde Grec Antique ?" In *L'histoire Comme Impératif Ou La "Volonté de Comprendre" : Hommage à Jean-Pierre Vernant et Pierre Vidal-Naquet*, 93–96. Naples: Publications du Centre Jean Bérard.

Ober, Josiah. 2015. *The Rise and Fall of Classical Greece*. Princeton: Princeton University Press.

Pantelia, Maria. 2020. "Thesaurus Linguae Graecae." 2020. http://stephanus-tlg.uci.edu.gorgone.univ-toulouse.fr/index.php.

Parker, Robert, Jean-Baptiste Yon, and Mark Depauw. 1996. "Lexikon of Greek Personal Names." 1996. http://www.lgpn.ox.ac.uk.

Perdikouris, Agelos. 2007. "The Little Sailing: Ancient Greek Texts." 2007. http://www.mikrosapoplous.gr/en/texts1en.html.

Pébarthe, Christophe. 2008. *Monnaie et marché à Athènes à l'époque classique*. Belin Sup Histoire. Paris: Belin.

"PHI Greek Inscriptions." n.d. Accessed August 31, 2020. https://inscriptions.packhum.org/.

Piantadosi, Steven T. 2014. "Zipf's Word Frequency Law in Natural Language: A Critical Review and Future Directions." *Psychonomic Bulletin & Review* 21 (5): 1112–1130.

Queiroz, Gabriela De, Colin Fay, Emil Hvitfeldt, Os Keyes, Kanishka Misra, Tim Mastny, Jeff Erickson, David Robinson, and Julia Silge. 2020. "Tidytext: Text Mining Using 'Dplyr', 'Ggplot2', and Other Tidy Tools." July 11, 2020. https://CRAN.R-project.org/package=tidytext.

R Core Team. 2015. "R: A Language and Environment for Statistical Computing." 2015. http://www.R-project.org.

Reed, William J., and Barry D. Hughes. 2003. "Power-Law Distributions from Exponential Processes: An Explanation for the Occurrence of Long-Tailed Distributions in Biology and Elsewhere." *Scientiae Mathematicae Japonicae* 8: 329–339.

Rochat, Yannick. 2014. "Character Networks and Centrality." PhD Thesis, Lausanne: Université de Lausanne. http://infoscience.epfl.ch/record/203889.

Roubineau, Jean-Manuel. 2015. *Les Cités Grecques (VI^e^-II^e^ Siècle Avant J.-C.) : Essai d'histoire Sociale*. Paris: Presses Universitaires de France.

Rydberg-Cox, Jeff. 2011. "Social Networks and the Language of Greek Tragedy." *Journal of the Chicago Colloquium on Digital Humanities and Computer Science* 1 (3): 1–11.

Saïd, Suzanne, Monique Trédé, and Alain Le Boulluec. 2017. *Histoire de la littérature grecque*. 3rd ed. Quadrige Manuels. Paris: PUF.

Schaps, David M. 2011. *Handbook for Classical Research*. London ; New York: Routledge.

Tauber, James. 2020a. "Jtauber/Diorisis." January 14, 2020. https://github.com/jtauber/diorisis.

———. 2020b. "Working with the Diorisis Ancient Greek Corpus." J. K. Tauber. January 20, 2020. https://jktauber.com/2020/01/20/working-with-the-diorisis-ancient-greek-corpus/.

Vatri, Alessandro, and Barbara McGillivray. 2018a. "The Diorisis Ancient Greek Corpus." May 2, 2018. https://figshare.com/articles/dataset/The_Diorisis_Ancient_Greek_Corpus/6187256.

———. 2018b. "The Diorisis Ancient Greek Corpus: Linguistics and Literature." *Research Data Journal for the Humanities and Social Sciences* 3 (1): 55–65.

Vernant, Jean-Pierre, and Pierre Vidal-Naquet. 2001. *Mythe et tragédie en Grèce ancienne*. Vol. 1. 2 vol. Paris: La Découverte.

Vernier, Bernard. 1980. "La circulation des biens, de la main-d'œuvre et des prénoms à Karpathos : du bon usage des parents et de la parenté." *Actes de la Recherche en Sciences Sociales* 31 (1): 63–92.

Waumans, Michaël C., Thibaut Nicodème, and Hugues Bersini. 2015. "Topology Analysis of Social Networks Extracted from Literature." *PLOS ONE* 10 (6): e0126470.