DHQ: Digital Humanities Quarterly
Preview
2023
Volume 17 Number 2

SEDES: Metrical Position in Greek Hexameter

Stephen A. Sansom <sasansom_at_fsu_dot_edu>, Florida State University https://orcid.org/0000-0003-0583-1906
David Fifield <david_at_bamsoftware_dot_com>, Independent Scholar

Abstract

This article outlines the processes of SEDES, a program that automatically identifies, quantifies, and visualizes the metrical position of lemmata in ancient Greek hexameter poetry; and gives examples of its application to investigate the effects of metrical position on poetic features such as formularity, expectancy, and intertextuality.

Introduction

1 Metrical position has long interested scholars of Greek hexameter poetry. In his *Encheiridion dia Metrôn* (“Handbook on Meter”), Hephaestion of Alexandria (2nd c. CE) referred to positions within meter as *χῶραι* (“places; positions,” cf. *LSJ* s.v. *χώρα* 3) and used the term to describe places in the line that contain various types of metrical “feet” (*πόδες*, cf. [LSJ] s.v. *πούς* 4), such as the dactyl (– ⏑ ⏑) and spondee (– –), as well as the “final” (*τῆς τελευταίας*, 7.1) position that may contain syllabic diversity, such as catalexis.[1] Aristides Quintilianus (circa 4th c. CE) likewise used *χώρα* to describe metrical position in his *Peri Musikês* (“On Music”), including the “first places” (*ταῖς πρώταις χώραις*, 1.24.27) of feet in the line (cf. [van Ophuijsen 1987, 82] [Winnington-Ingram 1963]). But it was not until the early 20th century that scholars interested in the statistical analysis of meter systematized notation for all possible positions of linguistic phenomena in the hexameter line; of these efforts, [O’Neill 1942] has gained the most currency.[2] Later, [Giangrande 1959] and others Latinized the *terminus technicus* from *χώρα* to *sedes* (“seat; place”), and now the term is ubiquitous in philological studies of metrical position.

2 With this greater precision, scholars have examined various forms of repetition in hexameter according to their metrical position. These studies largely focus on articulating metrical tendencies, such as colometric patterns [Porter 1951], [Barnes 1986] and the primacy of *caesurae* (*τομαί*) at certain *sedes* [Fränkel 1926], [Beekes 1972]. [3] In particular, [O’Neill 1942] discovered what he called the “localization” of “word-types,” that is, the “concentration of occurrences (of specific metrical shapes) in but a few of the possible positions” (114). Following O’Neill, [Porter 1951] argued that such tendencies established “patterns of expectancy” or norms in the mind of poet and audience, and philologists have begun to explore the potential intertextual, stylistic, cognitive, and literary effects of these patterns, for example, in [Giangrande 1967], [White 1986], [Hagel 1994], [Kahane 1994], [Martin 2000], [Forstall and Scheirer 2012], [Schein 2015, 93–116], and [Forstall and Scheirer 2019]. [4]

3 These studies have provided useful data and promising applications, but their potential has been somewhat constrained by limitations in scope, objects of analysis, and at times computational media. O’Neill treated only a relatively small number of lines, though his survey has been expanded and the data verified by [Porter 1951], [McDonough 1966], [Hagel 2004], and [Forstall and Scheirer 2012]. A lack of aggregate statistics on repeated words by *sedes* has restricted intertextual readings to the examination of infrequent words, *hapax*, or *dis legomena*, for example in [McLennan 1977]. [5] Even inquiries into anomalous behavior such as breaks in Hermann’s Bridge (that is, avoidance of word end between the biceps [two short elements] of the fourth foot [Martin 2000], [Schein 2015]) have been confined to a few authors or passages, due to a lack of available data. Existing digital tools for identifying repetitions in hexameter poetry — for example, the Thesaurus Linguae Graecae [TLG] or Perseus Project under PhiloLogic [PhiloLogic], for searching corpora by word list, lemma, or text string; the [Chicago Homer], for repeated phrases; and [Tesserae], for intertexts by *n*-gram matching [Forstall et al. 2015] — do not take into account metrical position. Future work on metrical position in Greek hexameter needs new methods and access to data in order to identify repetitive phenomena in meter more accurately and completely. [6] With such information, scholars can then better discern their linguistic causes and potential poetic effects.

4 In this article, we introduce SEDES, a software system for identifying, quantifying, and visualizing metrical position in ancient Greek hexameter poetry. Like intertextualist studies of metrical position, SEDES at present focuses on the repetition of words — more specifically, of lemmata. Our present purpose is to describe the design of the SEDES system and demonstrate a sample of its utility. [7] Engineering the system required bringing together existing resources in the processing of Greek poetry, as well as the creation of new processes, measurements, tests, and visualizations, all of which aid researchers of Greek literature and otherwise in the processing of poetic texts for computer and literary analysis and, more generally, provide a method that benefits the quantitative analysis of repetitive language in meter. In what follows, we show how SEDES generates these data, before giving examples of its application to Greek hexameter poetics, in particular the fields of oral formularity and intertextuality.

Processing Pipeline

5 The core function of SEDES is to assign a numeric value — an expectancy score — to every word in a work of verse. The expectancy of a word is a quantification, in a sense we will define later, of how frequently that word’s lemma appears at a given *sedes* in a corpus of works. After computing expectancy scores, SEDES provides ways of visualizing and interacting with them. The SEDES system is composed of a pipeline of programs, each of which performs a portion of the processing task. See Figure 1.

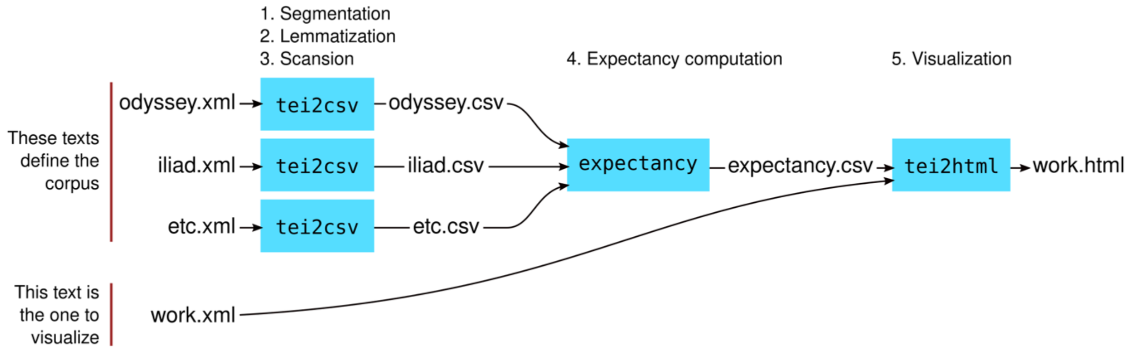

Figure 1. The SEDES processing pipeline comprises three programs: tei2csv, expectancy, and tei2html. Visualizing a text requires analyzing not only it, but all other texts of the corpus.

The input to the pipeline is a set of XML documents marked up according to the TEI Guidelines [TEI]. The selection of documents defines the *corpus* relative to which expectancy scores are computed. In developing SEDES, we have used 12 TEI texts from the Perseus Project [Perseus], totaling about 73,000 lines, with a minimum length of 479 lines and a maximum of 21,356. See Table 1 for a listing of texts, which span more than a millennium and comprise a variety of styles. We often define the corpus as all 12 works, but the system makes it possible to work with only a subset, in order to include or exclude certain authors or eras, for example. 6

Work	Date (circa c.)	Lines	Words
Iliad, Homer	8th BCE	15,683	111,865
Odyssey, Homer	8th BCE	12,107	87,185
Homeric Hymns	7th–6th BCE	2,342	16,022
Theogony, Hesiod	8th BCE	1,042	7,040
Works and Days, Hesiod	8th BCE	831	5,856
Shield of Heracles, Hesiod	6th BCE	479	3,298
Argonautica, Apollonius Rhodius	4th BCE	5,834	38,841
Hymns , Callimachus	3rd BCE	941	6,480
Phaenomena, Aratus Solensis	3rd BCE	1,155	7,752
Idylls, Theocritus	3rd BCE	2,527	18,071
Fall of Troy , Quintus Smyrnaeus	4th CE	8,801	60,098
Dionysiaca, Nonnus of Panopolis	5th CE	21,356	126,870
Total		73,098	489,378

Table 1. Works in the full SEDES corpus.

Though similar techniques could be applied to other forms of verse, SEDES is specially adapted to hexameter. In hexameter, vowels are distinguished by length. A line consists of six feet, which each may be either a long syllable followed by another long syllable (a spondee, denoted "– –"), or a long syllable followed by two short syllables (a dactyl, "– ⏑ ⏑"). The metrical shape of a word — its pattern of long and short syllables — limits the places (*sedes*) in a line it may appear. (Though there is some flexibility: a word may, subject to certain rules, sometimes be reinterpreted to have a different metrical shape than it would otherwise have, in order to fit the meter.) Even given the range of permissible "slots" for each word, most words tend to appear at some *sedes* more often than others. It is this non-uniform distribution of word placements that SEDES is designed to analyze. 7

The first program in the SEDES pipeline, tei2csv, assigns a lemma and a *sedes* to every word in the corpus. (We consider multiple instances of the same word separately — the same word in different lines may appear at different *sedes*, for example.) Later parts of the pipeline will operate, not on the original words of the text directly, but on the "lemma–*sedes* pairs" to which they are mapped by tei2csv. Concretely, tei2csv does the following three steps of segmentation, lemmatization, and scansion for every document in the corpus: 8

1. Break the document into lines and words.
2. Look up a lemma for each word.
3. Scan every line metrically, thereby finding a *sedes* for each word.

The output of tei2csv is a collection of CSV files, one for each TEI file in the corpus. The CSV files map every instance of every word in the corpus to the lemma and *sedes* of that instance, what we call a lemma–*sedes* pair. The next program in the pipeline, called expectancy, takes these CSV files as input and does the statistics necessary to compute expectancy scores: 9

4. Compute an expectancy score for every unique lemma–*sedes* pair.

The output of the expectancy program is another CSV file that maps every observed combination of lemma and *sedes* to a numeric expectancy score. The final program in the pipeline, tei2html, takes as input the expectancy CSV file and the TEI of any single work — which may or may not be part of the corpus 10 that was used to define expectancy — and produces an output visualization of the work:

5. Produce an HTML file that depicts differences in expectancy across words.

We will describe the above steps in detail in the subsections that follow. But first, we will walk through an example of applying the SEDES processing pipeline to book 1, line 1 of the *Odyssey* [Perseus – Odyssey] from start to finish. 11

In the source TEI file, the line looks like this: [8] 12

```
<l>a)/ndra moi e)/nnepe, mou=sa, polu/tropon, o(\s ma/la polla\</l>
```

The <l> and </l> XML tags delimit a line. The text between the tags is Beta Code [TLGBetaCode], a method of representing written ancient Greek using ASCII characters: "a" represents the letter alpha; ")", "/", "\", and "=" are various diacritics; and so on. The tei2csv program isolates the line from the XML markup and decodes the Beta Code to yield a line of verse represented in Unicode: 13

```
ἄνδρα μοι ἔννεπε, μοῦσα, πολύτροπον, ὃς μάλα πολλὰ
```

("Tell me, Muse, about the versatile man, who very many…") The program then scans the line metrically, marking each syllable as long (notated "–") or short (notated "⏑"). A long syllable counts for 1 position and a short syllable counts for 0.5. The scansion of the example line is: 14

```
ἄν δρα   μοι  ἔν νε πε,  μοῦ σα,  πο λύ τρο πον,  ὃς   μά λα   πο λλὰ
–   ⏑     ⏑   –  ⏑  ⏑    –   ⏑    ⏑  –  ⏑   ⏑     –    ⏑  ⏑    –  –
1   2    2.5  3  4 4.5   5   6   6.5 7  8  8.5    9   10 10.5  11 12
```

The *sedes* of a word is the *sedes* of its first syllable. In this example, *ἄνδρα* is at *sedes* 1, *μοι* is at *sedes* 2.5, and so on. Separately from scansion, every word in the line is mapped to its lemma. Here, the word *ἄνδρα* maps to the lemma *ἀνήρ*, *μοι* maps to *ἐγώ*, and so on for the rest of the words. The output of this process is a lemma and a *sedes* for each word: a lemma–*sedes* pair. The calculation of *sedes* and the finding of a lemma are independent operations: it is the syllables of the original words of the line, not their derived lemmata, that determine *sedes*. 15

The tei2csv program emits a CSV file with one row per input word, representing the word's lemma, *sedes*, and other metadata. 16

work	book_n	line_n	word_n	word	lemma	sedes	metrical_shape	scanned	num_scansions	line_text
Od.,	1,	1,	1,	*ἄνδρα*,	*ἀνήρ*,	1,	–⏑,	auto,	1	"*ἄνδρα μοι*..."
Od.,	1,	1,	2,	*μοι*,	*ἐγώ*,	2.5,	⏑,	auto,	1	"*ἄνδρα μοι*..."
Od.,	1,	1,	3,	*ἔννεπε*,	*ἐνέπω*,	3,	–⏑⏑,	auto,	1	"*ἄνδρα μοι*..."
Od.,	1,	1,	4,	*μοῦσα*,	*μοῦσα*,	5,	–⏑,	auto,	1	"*ἄνδρα μοι*..."
Od.,	1,	1,	5,	*πολύτροπον*,	*πολύτροπος*,	6.5,	⏑–⏑⏑,	auto,	1	"*ἄνδρα μοι*..."
Od.,	1,	1,	6,	*ὃς*,	*ὅς*,	9,	–,	auto,	1	"*ἄνδρα μοι*..."
Od.,	1,	1,	7,	*μάλα*,	*μάλα*,	10,	⏑⏑,	auto,	1	"*ἄνδρα μοι*..."
Od.,	1,	1,	8,	*πολλὰ*,	*πολύς*,	11,	––,	auto,	1	"*ἄνδρα μοι*..."

Table 2.

Expectancy is a function of lemma and *sedes* jointly: it measures, for each lemma, how that lemma is distributed over the possible *sedes*, in the context of the overall corpus. Having run tei2csv on the *Odyssey*, we run it again on the other works of the corpus to obtain the necessary context, in the form of a collection of CSV files that give a lemma and a *sedes* for every word. The expectancy program takes all these files as input, analyzes the statistical distribution of *sedes* for each lemma, and outputs another CSV file that maps lemma–*sedes* pairs to their expectancy scores, using mathematical formulas we will define below. An excerpt of the output of the expectancy program is shown below. The x column counts the number of occurrences of a lemma–*sedes* pair throughout the corpus, and z shows the expectancy of that pair. From this sample output, we see that the expectancy of the lemma *ἀνήρ* ("man") at *sedes* 1 is +1.11, that of *ἐγώ* ("I") at *sedes* 2 is −0.13, and so on. 17

lemma	sedes	x	z
[...]			
ἀνήρ	1	452	+1.1089031679816
ἀνήρ	2	35	-1.60788923391802
ἀνήρ	2.5	100	-1.18440840388571
[...]			
ἐγώ	2	430	-0.132108322471802
ἐγώ, , ,	2.5	1027	+1.79559085302027
ἐγώ	3	457	-0.0449259477008033
[...]			
ἐνέπω	3	8	-1.32295427519852
[...]			

Table 3.

18 Finally, the tei2html program takes as input the TEI file of the *Odyssey* along with the expectancy CSV file. tei2html renders a human-readable HTML representation of the text, optionally highlighting each word according to its expectancy. One visualization style is shown below, with words shaded according to their expectancy: low expectancy is dark and high expectancy is light. Other styles will be explored in the section on visualization.

19 ἄνδρα μοι ἔννεπε, μοῦσα, πολύτροπον, ὃς μάλα πολλὰ

20 The programs that constitute SEDES are written in Python 3. Running SEDES from start to finish on every work in our corpus (12 TEI files, 73,000 lines, and 490,000 words) on a 2019 MacBook Pro takes about one minute. About 45% of the time is spent on segmentation, lemmatization, and scansion; 9% on expectancy computation; and 46% on generating visualization output.

Segmentation into Lines and Words

21 Scansion is an operation on individual lines, and lemmatization is an operation on individual words. SEDES begins by breaking a TEI document into lines, and its lines into words.

22 Segmentation into lines is fairly straightforward. Lines are delimited by either the l element, which encloses lines; or the lb element, which separates them: [9]

```
<l>mh=ter e)pei/ m' e)/teke/s ge minunqa/dio/n per e)o/nta,</l>
<l>timh/n pe/r moi o)/fellen *)olu/mpios e)gguali/cai</l>
<l>*zeu\s u(yibreme/ths: nu=n d' ou)de/ me tutqo\n e)/tisen:</l>
<l n="355">h)= ga/r m' *)atrei/+dhs eu)ru\ krei/wn *)agame/mnwn</l>
<l>h)ti/mhsen: e(lw\n ga\r e)/xei ge/ras au)to\s a)pou/ras.</l>
```

Example 1. *Iliad* 1.351–56 [TLGBetaCode][Perseus – Iliad]

```
<lb/><milestone ed="P" unit="para" />kou/rh d' *)wkeanou=, *xrusa/ori karteroqu/mw|
<lb rend="displayNum" n="980" />mixqei=s' e)n filo/thti poluxru/sou *)afrodi/ths,
<lb />*kalliro/h te/ke pai=da brotw=n ka/rtiston a(pa/ntwn,
<lb />*ghruone/a, to\n ktei=ne bi/h *(hraklhei/h
<lb />bow=n e(/nek' ei)lipo/dwn a)mfirru/tw| ei)n *)eruqei/h|.
```

Example 2. *Theogony* 979–83 [Perseus – Theogony]

23 Having isolated a line, we decode its text contents from Beta Code to Unicode. We extract words from the line by splitting on sequences of non-letter, non-diacritic characters. For this purpose, the apostrophe character (’), used to mark elision in words like *ἀλλ’*, is considered to be a letter character. Some TEI documents use apostrophes to mark quotations, which complicates word extraction by giving the character a second meaning as punctuation. When we discovered uses like this, we amended it in the source document using quotation markup.

Lemmatization

24 SEDES needs a way of defining which words count as “the same” for the purpose of computing distributions over *sedes*. Our instrument for deciding which words are considered equivalent is the lemma. [10] In lexicography, a lemma is the dictionary headword with which all its morphological variants are associated. Ancient Greek is a highly inflected language, one in which words take on many forms according to their grammatical function. The lemma unifies the many related forms of a word and bundles them under one label. For example, the genitive singular word *ἀοιδῆς* and the dative plural *ἀοιδαῖς* both have the same lemma, *ἀοιδή* (“song”); the two words therefore fall into the same bucket for the purpose of computing statistics. (Note, however, that lemma assignment does not affect the scansion procedure that will be described in the next section; scansion uses the words of the original text, pre-lemmatization.) As a semantic matter, lemma corresponds to lexeme, an item of meaning removed from a language’s inflectional rules; we may think of a lemma as a representative of a lexeme, standing for a group of related words. [11]

25 Mapping a word to its lemma is not trivial [Heslin 2019, 391]. But strictly speaking, we do not care about a word’s lemma per se. All we care about is when two words have the same lemma. In mathematical terms, equality of lemmata is an equivalence relation that partitions the universe of words into equivalence classes: all words that have the same lemma are in the same class, and no two words in the same class have different lemmata. [12] As long as a lemmatization algorithm is consistent, it may as well output abstract tokens like “x1234” as literal Greek words. [13] This is to say that the consequence of a word being lemmatized incorrectly is that the word ends up in a different class than it should: it is missing from its “true” class and is instead grouped with some other class of words (or a class by itself). Whether such an error makes a meaningful difference in the statistics depends on how many times the word occurs, and the sizes of the other lemma classes. In SEDES, we use the backoff Greek lemmatizer [14] of the Classical Language Toolkit [CLTK]. The backoff lemmatizer employs a chain of sub-lemmatizers (which are various forms of dictionary lookup and regular expression transformations [Burns 2018], [Burns 2020, 161–62]), trying each in turn until one returns a result. If no lemma is found by any of these techniques, the last-resort fallback is to use the word itself as the lemma. The fallback occurs for about 2% of words in the corpus (7% of unique words). A random sample of 100 lines suggests a 98.6% accuracy rate: we identified 10 misattributed lemmata among the 721 words in *Iliad* 13.563–662.

26 We augment the CLTK lemmatizer with a preprocessing step that attempts variations on accent marks: if the initial lemmatization fails, the lemmatizer applies one of a small set of accent transformations, namely, changing a grave accent to an acute accent, or removing an acute accent.[15] The pre-transformation step makes a few thousand words lemmatizable that would not be otherwise. We further have the lemmatizer first consult a hardcoded list of our own, manually determined lemmata, which we use this list for words that are common in our corpus but for which automatic lemmatization fails.

Scansion

27 The computation of expectancy requires a word’s lemma and *sedes*. In the previous step we found the lemma; now we tackle the *sedes*. *Sedes* is a byproduct of scansion, which, in the context of Greek hexameter, means determining the metrical value (long or short) of every syllable in a line.[16] Once a scansion is settled, the *sedes* of a word is determined by the metrical values of the words preceding it in the line. For automatic scansion, we use the Python hexameter module [Ranker 2012]. Here we will briefly describe how the algorithm works.

28 The first step is to isolate syllables by looking for vowels and diphthongs. Each lexical syllable is assigned a preliminary metrical value according to language rules that are specific to Greek. An example of these rules is that, unless modified by other rules, the vowels “ε” and “ο” are short, “η” and “ω” are long, and other vowels are (for now) undetermined. At this point in the analysis, a syllable may be classified as definitely long, definitely short, or one of a few other values that govern how the syllable may be resolved into long or short in the next step.

29 Resolution of preliminary metrical values into final metrical values is accomplished using a finite automaton. A finite automaton is a computational state machine that represents (or recognizes) a restricted set of inputs — in our case, the set of properly formed lines of Greek hexameter. The automaton matches lines of hexameter the same way that a regular expression matches a subset of strings. The automaton encodes various rules of the poetic form: for example, that a line contains six feet total; that the first half of a foot must be a long syllable; and that the second half of every foot but the last may be either one long syllable or two short syllables. Each transition in the automaton is marked with both the preliminary metrical values that permit the transition to be taken, and the final metrical value the syllable is to be assigned, when that transition is taken. In general, there is more than one way of fitting a line into the schema of hexameter. For example, one might change a short syllable to a long early in the line, making compensating adjustments to following syllables in order to maintain the same total length. This fact is reflected in the automaton by there being multiple paths from start to finish when there are multiple feasible scansions. Not all scansions are equally good, however. The automaton’s output is biased towards parsimony by weights attached to state transitions, which penalize metrical value assignments that require uncommon metrical rules. The associated final metrical values of the path with lowest weight becomes the scansion of the line.

30 Occasionally, the automatic scansion algorithm yields either no paths through the automaton, or two or more paths of equal weight. We scanned these problematic lines by hand and added them to a list of scansion overrides, which take precedence over automatic scansions. There are 1,526 entries in the list of overrides, about 2.1% of the lines in the corpus. Manually scanning lines whose automatic scansion is undefined or ambiguous requires human expertise in the form of meter being analyzed. The CSV output of tei2csv records the source of the scansion in the scanned column — either manual or auto. The num_scansions column counts the number of equally weighted automatic scansions found, usually 1.

31 Once every syllable in a line has been assigned its metrical value, it is straightforward to determine the *sedes* of each word. Start at the beginning of the line and *sedes* 1. Advance the *sedes* by 1 for a long syllable, and by 0.5 for a short syllable. The *sedes* of a word is the *sedes* of its first syllable. The *sedes* of words in a line are not all necessarily distinct. For example, in *Iliad* 1.241 [Perseus – Iliad], *σύμπαντας· τότε δ’ οὔ τι δυνήσεαι ἀχνύμενός περ*, the two adjacent words *δ’* and *οὔ* share *sedes* 5, because they are pronounced as one word.

32 To evaluate the accuracy of the automatic scansion algorithm, we randomly sampled 1,200 lines from the entire corpus and manually checked their automatic scansion. 1,195 lines (99.6%) were scanned correctly.

Expectancy Computation

33 We now come to the central question: how expected is each word, given its lemma and the *sedes* at which it appears? Following [Porter 1951], we call the measure of expectedness “expectancy.” Expectancy is not a property of lemma only or of *sedes* only, but of both considered jointly. For each lemma–*sedes* pair, we ask: given the distribution of this lemma in the corpus, is this *sedes* usual or unusual?

34 Our metric of expectancy is the *z*-score, which measures distance from the mean of a distribution in units of the standard deviation [Hoover 2013, 520–21], [Allen 1988, 6–7]. The formula for the *z*-score of an observed value x is $z = (x - \mu)/\sigma$ where μ is the mean of a population and σ is its standard deviation. A *z*-score of +1.5, for example, describes a value that is 1.5 standard deviations greater than the mean. Values greater than the mean have positive scores, and values less than the mean have negative scores. We count how many times a certain lemma appears at a certain *sedes*, and the *z*-score expresses whether that count is greater or less than would be expected, given the counts at that and other *sedes*.

35 In our definition of expectancy, different lemmata do not interact. [17] The expectancy of a lemma at a particular *sedes* depends only on the distribution of that lemma. For example, in our corpus, the lemma *βοῦς* (“cow”) appears 448 times, while *χέλυς* (“tortoise”) appears only 8 times. It is more “expected,” then, in some sense, for any given word to be *βοῦς* than to be *χέλυς*. But we are not interested in comparing *βοῦς* versus *χέλυς*; rather we weigh *βοῦς* against other instances of βοῦς, and χέλυς against other instances of χέλυς.

36 To establish corpus-wide distributions of lemmata and *sedes*, we repeat the three previous steps — segmentation, lemmatization, and scansion — for every work in the corpus, using the tei2csv program to create a collection of work-specific CSV files with a lemma and *sedes* for every word. Then a separate program, called expectancy, takes as input the whole collection, computes the expectancy of every unique lemma–*sedes* pair across the entire corpus, and writes a table of computed expectancy values to another CSV file. This table will be used in the next step to produce a visualization for a single work.

37 The expectancy program begins by counting the number of occurrences of each unique lemma–*sedes* pair. Then, for each lemma, it examines the lemma’s distribution over the possible *sedes*. Before applying the *z*-score formula to the table of *sedes* counts, we weight each count by itself. For instance, the weighted mean of the distribution of counts [1, 2, 2, 15] is not (1 + 2 + 2 + 15) / 4 = 5, but rather (1×1 + 2×2 + 2×2 + 15×15) / (1 + 2 + 2 + 15) = 11.7. This is because the counts represent 20 total words, not 4; since the majority of words appear at a *sedes* having a count of 15, a representative count should be closer to 15 than it is to 1 or 2. We similarly weight the standard deviation calculation. To be precise, we use the following formulas for the weighted mean and standard deviation of a vector of *sedes* frequency counts [c_1, …, c_n]:

38

$$\mu = \frac{\sum_{i=1}^{n} c_i \times c_i}{\sum_{i=1}^{n} c_i}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} c_i \times (c_i - \mu)^2}{\sum_{i=1}^{n} c_i}}.$$

39 The output of the expectancy program is a CSV file that maps lemma–*sedes* pairs to their respective *z*-scores.

40 The *z*-score formula is not defined for distributions whose standard deviation is zero. Such a situation arises when the frequency counts for a lemma are exactly equal in every *sedes* where it appears. Examples are *ἑκηβόλος* (“far-shooter”), which occurs 42 times but always in *sedes* 6.5, and *Κίλλα* (the astyonym “Killa”), which occurs 4 times total, twice each in *sedes* 1 and 11). Important special cases of this general phenomenon are words that appear only once or twice in the corpus, such as the *hapax legomenon ἑλώρια* (*Iliad* 1.4, lemma *ἑλώριον*). Expectancy is also undefined for lemmata that never appear in the corpus — those whose table of frequency counts is everywhere zero. This last situation may occur only when when the work under consideration is not part of the corpus that defines expectancy: if it had been, all its lemmata would necessarily have a nonzero count in at least one *sedes*. For the purposes of visualization, we treat undefined *z*-scores as if they were zero; i.e., neither more nor less frequent than expected. Figure 2 shows the distribution of *z*-scores for every word in the entirety of our corpus.

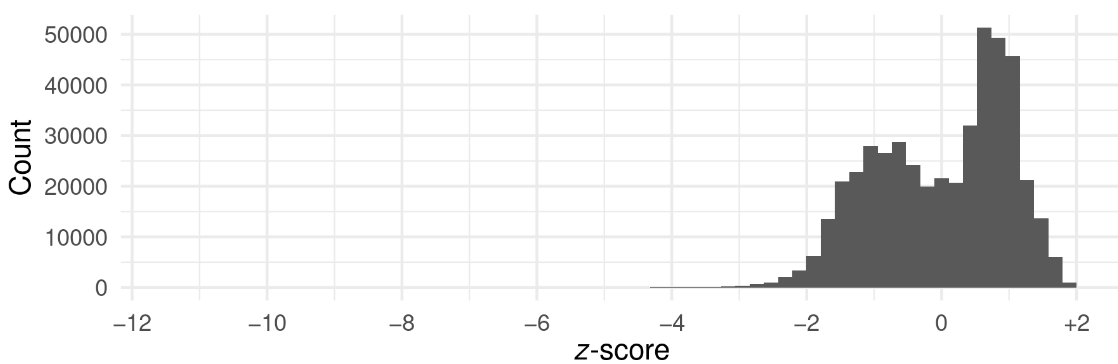

Figure 2. Histogram of *z*-scores for all words across our entire corpus. This chart excludes about 33,000 words with undefined *z*-scores. Over 95% of *z*-scores lie in the interval [−1.75, +1.75], though the tail of negative values extends as far as −11.5.

41 Even disregarding lemmata, it is not the case that all *sedes* are equally likely to be the site of the start of a word. A simple demonstration of this fact is that every line of hexameter contains a word at *sedes* 1 (the first word), but whether there is a word at *sedes* 2 depends on the how long the first word is. Figure 3 shows the distribution over *sedes* for the works in our corpus. One might reasonably suggest applying a correction to the expectancy formulas to adjust for an assumed *a priori* distribution over *sedes*. We have chosen not to do so, reasoning that a word’s appearing at an uncommon *sedes* is itself a sign of unexpectedness.

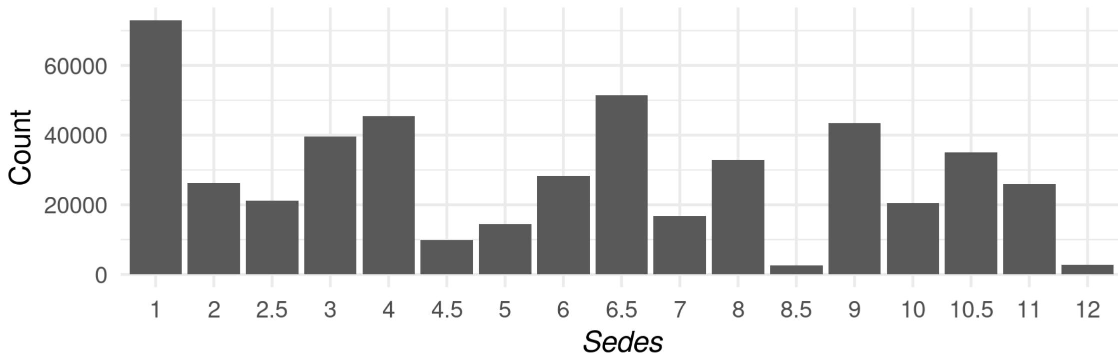

Figure 3. Distribution of *sedes* cross the full corpus. Not all *sedes* are equally likely.

Further exploration and justification of the utility of *z*-scores as a measure of expectancy appears below in the section “Interpretation of *z*-scores.” We now show how expectancy scores feed into the final step of the pipeline, namely visualizing numeric expectancy data. 42

Visualization

The data files emitted by the previous steps lend themselves to various forms of analysis. Our specific goal is to produce a readable visualization of the source text, with every word highlighted according to its expectancy.[18] Such a visualization enables a human reader to quickly visually peruse and identify words whose expectancy is much different than expected, aided by the context of the surrounding text. Words thus identified may be worthy of further investigation. 43

The processing pipeline terminates in the tei2html program, which takes as input a TEI text and the table of lemma–*sedes* expectancy scores output by the expectancy program, and produces a visualization that uses HTML, CSS, and JavaScript. Like tei2csv, tei2html isolates words from its TEI input and assigns each one a lemma and *sedes*. It looks up the expectancy scores of the lemma–*sedes* pairs, and emits an HTML document where every word is annotated with its expectancy and other metadata. 44

The HTML document provides an interactive selection of visualization styles. A simple and effective way to depict expectancy is to shade the color of words on a scale from dark to light.[19] In this style, words with higher expectancy fade into the background, and those with lower expectancy are visually emphasized (Figure 4). 45

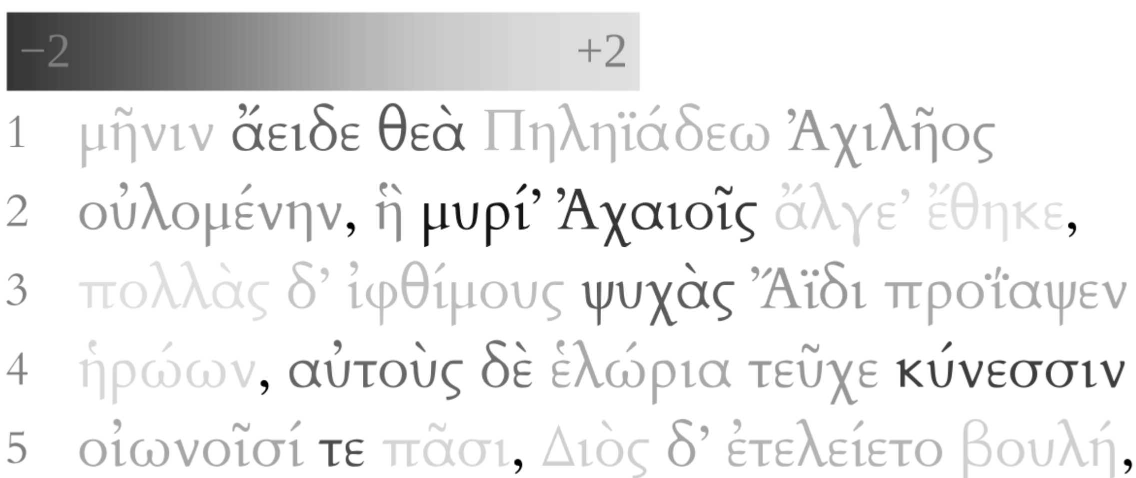

Figure 4. Words shaded according to *z*-score, dark to light.

A variation of this style is a diverging color gradient, which, rather than emphasizing the difference between negative and positive, emphasizes distance from the mean. In this style, *z*-scores close to zero are visually diminished, while those that are large positive or large negative are highlighted in saturated and contrasting colors. Here, negative *z*-scores are red and positive *z*-scores are blue (Figure 5). 46

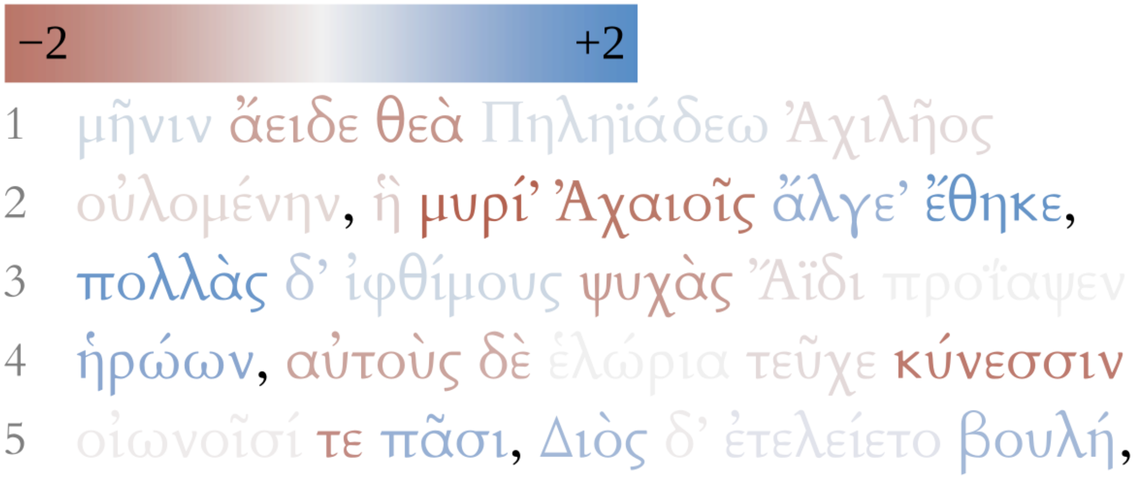

Figure 5. Words shaded on a diverging scale, red to gray to blue.

Words vary in their length and visual density; of two words shaded identically, the longer word may be visually more prominent, simply because up more space. In an attempt to mitigate this potential bias, we have provided alternative styles that attach filled circles of uniform size to words, where the circles are shaded rather than the text (Figure 6). Here, as well, the gradient may be sequential or diverging. 47

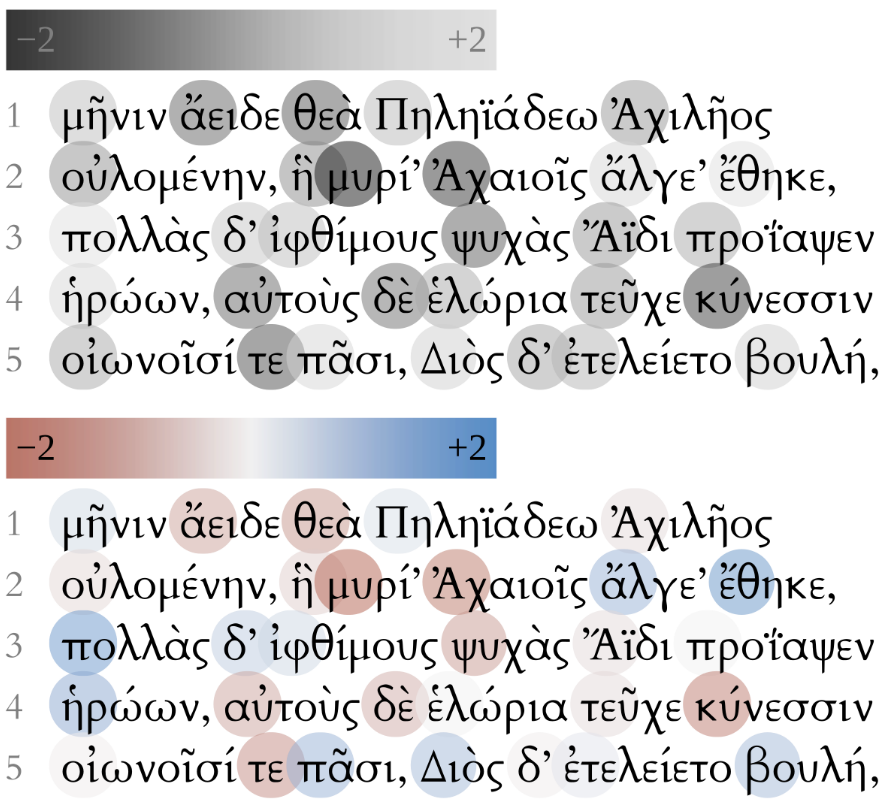

Figure 6. Shaded circles of uniform size are an alternative to text shading.

The visualization has an “align to *sedes* grid” option that vertically aligns words according to their *sedes* (Figure 7).[20] This option may be combined with any visualization style. 48

	1	2	2.5	3	4	4.5	5	6	6.5	7	8	8.5	9	10	10.5	11	12
1	μῆνιν		ἄειδε			θεὰ		Πηληϊάδεω						Ἀχιλῆος			
2	οὐλομένην				ἣ		μυρί’		Ἀχαιοῖς				ἄλγε’		ἔθηκε		
3	πολλὰς			δ’ ἰφθίμους				ψυχὰς			Ἄϊδι			προΐαψεν			
4	ἡρώων				αὐτοὺς			δὲ	ἑλώρια				τεῦχε		κύνεσσιν		
5	οἰωνοῖσί					τε	πᾶσι		Διὸς		δ’ ἐτελείετο					βουλή	

Figure 7. “Align to *sedes* grid” forces words in the same *sedes* into vertical alignment.

The HTML output provides a panel for selecting visualization options and displaying additional information (Figure 8). Clicking on a word shows its metrical shape, lemma, *sedes*, and expectancy, as well as the distribution of the lemma over *sedes* and the resulting *z*-scores. 49

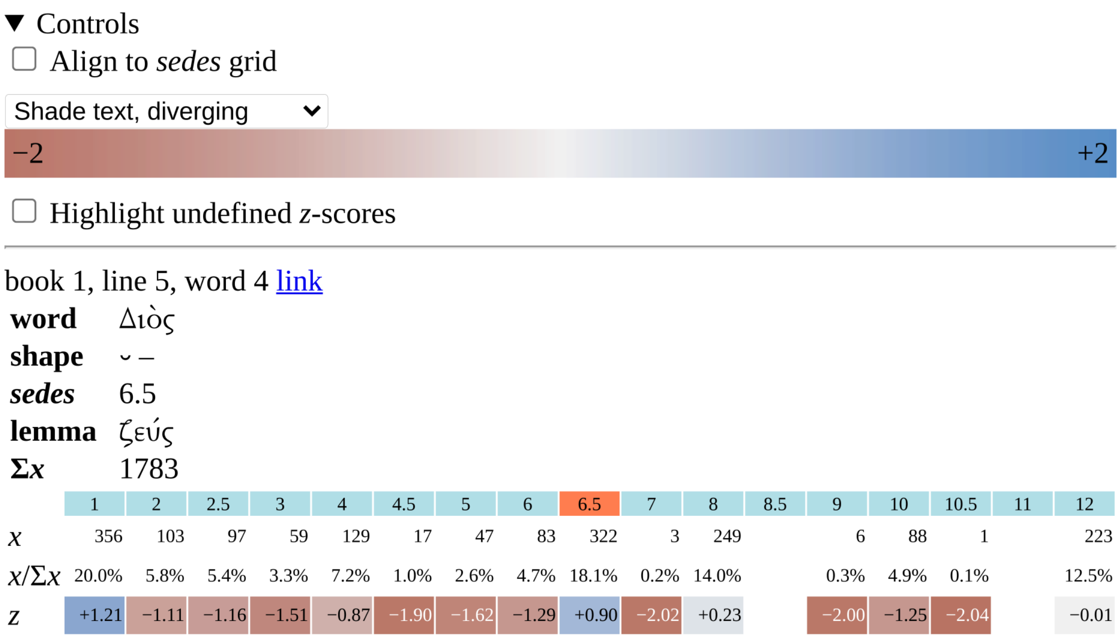

Figure 8. The interactive panel that controls visualization and shows information about selected words, here showing the *sedes* and *z*-score distribution of the lemma *ζεύς*, found in *Iliad* 1.5.

Interpretation of *z*-scores

The essential aspect we hope to capture with SEDES is how expected, in some human sense, a lemma is at a particular metrical position. There are many possible ways to quantify expectancy (e.g., frequency charts à la [O’Neill 1942]); the one we have chosen is the *z*-score. The *z*-score is an adaptive measure that does not assume a specific prior distribution of counts over *sedes*. Because the *z*-score is normalized to the standard deviation of a distribution, it provides a common basis of comparison for all lemmata, whether frequent or rare. Let us take two examples to explore the meaning of the *z*-score and how it is affected by the underlying *sedes* distribution. The two lemmata we will look at both occur over 50 times, but with markedly different distributions. 50

First, consider the lemma *μορφή* (“shape; beauty”). It overwhelmingly appears at *sedes* 11, and only rarely occurs at four other *sedes*. *Sedes* 11 is assigned a *z*-score of +0.33, whose nearness to zero indicates that it is expected, or unmarked, at this *sedes*.[21] In contrast, the rarely occurring *sedes* are assigned large negative *z*-scores, less than −3.00, which indicate a low degree of expectancy. 51

Sedes	x	*z*-score
1	8	−3.01
2	2	−3.15
4	5	−3.08
6	1	−3.18
11	150	+0.33

Table 4. *Sedes* distribution and expectancy of *μορφή*.

Next, consider the lemma *δένδρεον* (“tree”). Its distribution is less skewed. It occurs with roughly equal frequency at *sedes* 3, 7, and 9; still more often at *sedes* 1, but not overwhelmingly so. This lemma is more likely to appear at *sedes* 1 than at any other single *sedes* — but *sedes* 1 accounts for fewer than 52 half of total instances. In this distribution, a count of 11 is closer to “typical” than a count of 19. The evenly distributed *sedes* accordingly get negative *z*-scores that are relatively close to zero. The *sedes* with a larger count gets a positive *z*-score, farther from zero.

Sedes	x	*z*-score
1	19	+1.30
3	10	−1.02
7	12	−0.51
9	11	−0.76

Table 5. *Sedes* distribution and expectancy of *δένδρεον*.

The extreme *z*-scores — those that are most negative and most positive — naturally stand out as being worthy of attention. The most negative scores are the rare outliers, the *sedes* at which a lemma appears much less frequently than usual. The first example above illustrates such a situation. A large negative score is a hint (though not a guarantee) that something unexpected is happening compositionally: perhaps a common word is used in an uncommon way, or as part of an uncommon poetic structure. In the other direction, the most positive *z*-scores show where a lemma appears *more* frequently than usual. Note an apparent contradiction: in order for a lemma to be more than expected at a certain *sedes*, it must appear there a large number of times — does that not make the lemma actually expected at that *sedes*? But there is no contradiction, really: large positive *z*-scores arise only with counts that are larger than the mean, but which are outweighed by the sum of smaller counts at other *sedes*. The second example above is such a case. In contrast to the rare outliers signified by large negative *z*-scores, large positive *z*-scores represent occurrences that occur frequently yet are still unusual in the larger context of the lemma. The diverging color scale visualization options bring out both phenomena. 53

It is important not to confuse *z*-scores with pure frequency counts. A low *z*-score does not mean that a lemma itself is uncommon. Just the opposite: it means that *a particular sedes* is an uncommon one for the given lemma. Intuitively, a lemma’s appearance at one *sedes* can only count as unusual if there are sufficiently many instances of the lemma at other *sedes* to establish what “usual” is. Large negative *z*-scores are only possible with frequently occurring lemmata. To reach a *z*-score as low as −2, there must be at least 5 total instances of a lemma; for −5 there must be at least 26; and for −10 there must be at least 101. (Compare to the empirical distribution of *z*-scores in Figure 2.) 54

It may happen that the demands of poetic form affect word placement and therefore the distribution of *sedes*. For instance, some lines in a poem may be more or less formulaic than others, and therefore be more or less constrained in their composition. Suppose that in a special subset of lines (invocations, say), a certain lemma falls at a *sedes* at which it does not commonly appear in other, non-invocation lines. The relatively low frequency of the lemma at its place in an invocation is, then, partially a reflection of the relative rarity of invocation lines. In the subset of lines that are invocations, the lemma may in fact be perfectly expected at a *sedes* that is unusual in the wider context of the poem. It is possible to imagine specializations of *sedes* expectancy that take into account additional factors, such as the functional or literary purpose of the lines in which words appear. Doing so, however, requires presupposing criteria by which a work is to be subdivided. Our present purpose is different: to use *sedes* as a tool to *discover* where the placement of words has been affected, whether by formula or by any other cause. 55

SEDES in Action: Formula and Intertext

We have engineered SEDES to be useful for diverse types of analysis and, by making the code available as free software, to be adaptable to the interests of future researchers of metrical position in Greek hexameter, by itself or in tandem with other digital humanities tools. Most readily, it is designed to compute and visualize the frequency of lemmata. Users can explore the corpus with the visualization (either online or by downloading the program), look for dark shading among the sea of gray, and click for statistics on the lemma. Or, after downloading the program, they can create the CSV file of underlying data and organize it in a spreadsheet according to their own interests, for example by ordering rows by lemma and *z*-scores. Once the user finds a word in an unexpected position, they then can interpret its literary significance (see, e.g., [Sansom 2021, 454–60]). As a sample of its functionality in relation to established ways of interpreting Greek hexameter, in what follows we briefly gesture towards the integration of *sedes* analysis with two prominent interpretative modes: formularity and intertextuality. 56

SEDES adds granular, novel information to our understanding of formulaic regularity. In oral-poetic theory and Homeric philology, the formula serves as one of the most basic and useful units of composition, though the quality and nature of the formula in epic diction is highly contested.[22] As defined by [Parry 1930, 80], a formula is “a group of words which is regularly employed under the same metrical conditions to express a given essential idea” ([Parry 1971, 272]; cf. [Parry 1928, 16] “une expression qui est régulièrement employée, dans les mêmes conditions métriques, pour exprimer une certaine idée essentielle”). Despite its focus on single and not groups of words, SEDES helps to identify unexpected behavior in constituent words of the group, behavior that should inform our conception of formulaic regularity. Take the phrase *πατρίδα γαῖαν*, “(to one’s) fatherland.” Using a collated lemma search in the [TLG] or a phrase search in the [Chicago Homer], we find that the phrase *πατρίδα γαῖαν* occurs 21 times in the *Iliad*, with 19 instances positioned at line end (after the “bucolic diaeresis” at *sedes* 9–12).[23] Iliad 15.499 is typical: 57

> *οἴχωνται σὺν νηυσὶ φίλην ἐς πατρίδα γαῖαν*
> (*Iliad* 15.499, *πατρίς* *z* = +0.51 at *sedes* 9, 19 instances at line end)

There is variance in the preceding elements: most often the phrase is introduced with the preposition *φίλην ἐς* (“to the beloved (fatherland)”), though three instances at line end lack preceding *φίλην ἐς* (*Iliad* 7.335, 13.645, and 15.706). Although the phrase *πατρίδα γαῖαν* is highly regular and predictable in other respects, two exceptional instances in the *Iliad* shift its metrical position away from line end: 58

> *ἐμβαδὸν ἵξεσθαι ἣν πατρίδα γαῖαν ἕκαστος*
> (*Iliad* 15.505, *πατρίς* *z* = −1.81 at *sedes* 7, Ajax to Achaians)

σὴν ἐς πατρίδα γαῖαν, ἐπεί με πρῶτον ἔασας
(*Iliad* 24.557, *πατρίς* *z* = −2.37 at *sedes* 3, Priam to Achilles)

59 In the two exceptions, by collating the paired lemmata search or phrase search with the SEDES visualization, we see that there is also a change in the expectancy of the phrase's constituent parts, *πατρίς* and γαῖα. *Sedes* 9 is a natural place for the lemma *πατρίς*, whether or not part of a formula. While γαῖα remains relatively expected in each position, in *sedes* 3 and 7 *πατρίς* is placed unexpectedly (Figure 9).

	1	2	2.5	3	4	4.5	5	6	6.5	7	8	8.5	9	10	10.5	11	12
x	3			2			1			22			105				
$x/\Sigma x$	2.3%			1.5%			0.8%			16.5%			78.9%				
z	−2.34			−2.37			−2.40			−1.81			+0.51				

Figure 9. Table of lemma πατρίς ("of one's fathers") by *sedes* in the Archaic subset of the corpus, with expectancy. The total number of instances of the lemma is Σ*x* = 133.

60 When comparing the instances of the phrase at *Iliad* 15.505 and 24.557 with the 19 others at line end, there is a difference in formularity. If you were to ask which instances of the formula *πατρίδα γαῖαν* are most "formulaic," it seems reasonable to claim that the 19 instances of the phrase at line end are more "formulaic" than the two that occur elsewhere, due to their regularity in that position of the line. This in itself suggests that not all instances of even a single formula are equally expected in a given metrical position. But we could have seen this shift of phrase even without SEDES, using, for example, the repeated phrase tool of the Chicago Homer or a proximity search for the two lemmata in the *Thesaurus Linguae Graecae*. What SEDES allows us to see for the first time is that the placement of the phrase at *sedes* 3 and 7 not only deviates from normal metrical placement for the *phrase*, it places the *word πατρίς*, in particular, in a less expected place. In other words, while the phrase is "formulaic" in all instances, it is less so in *Iliad* 15.505 and 24.557, where it is out of place and where its constituent words, especially *πατρίς*, stray from their typical positions.

61 This new information about the differing metrical behavior of constituents of formulae has at least three effects. First, it nuances the notion of the "flexibility" of formulaic language [Hainsworth 1968]. SEDES shows that words within a phrase may behave in different ways *qua* words, even if they remain contiguous within the phrase and the phrase is not otherwise "expanded" or "contracted." Future work on epic formularity can adapt and incorporate SEDES to fit different models of concepts central to the idea of the formula. If scholars of epic language who are committed to the empirical investigation of formularity are to define the quality and nature of this fundamental component of epic composition, the *sedes* expectancy of constituent elements of formulae — individual words — should factor into notions associated with formularity, such as typicality, regularity, expectation, economy, and flexibility.

62 Second, unexpected constituents of a formula may interact with other literary features of a text. Scholars have shown how the meaning of individual elements of formulae, such as an epithet, can become more or less salient at different moments in the text [Elmer 2015], [Visser 1988], [Bakker and Fabbricotti 1991]. *Sedes* expectancy may also do something similar when one or more elements of a formula occurs in an unexpected position. Of the two instances above, the more statistically marked instance of *πατρίδα* is at *Iliad* 24.557, in which it occurs at a *sedes* (3) that is more than two standard deviations away from the mean ($z = -2.37$). In the context of book 24, the unexpected placement of *πατρίς* ("of one's father[land]") concords with the book's emphatically paternal semantics, several of which align with the *sedes* of *πατρίδα*. The final book of the *Iliad* features a clandestine visit by Priam, the king of Troy, to Achilles' tent through divine aid. Priam appears suddenly at Achilles' camp and kisses Achilles' hands (24.477–9), astounding Achilles and his companions (*θάμβος…θάμβησεν…θάμβησαν* Iliad 24.482–4), and utters his first words of petition: "remember your father" (*μνῆσαι πατρὸς σοῖο Iliad* 24.486). As [Edwards 1991, 10] states, "Priam's first words are a culmination of (the *Iliad*'s) theme (of father–son relationships)." In this sudden, incipient statement, *πατρός* occupies *sedes* 3, a less expected position for the lemma *πατήρ* ($z = -1.24$) and the same unexpected *sedes* as *πατρίδα (γαῖα)* later. It is notable that of the twenty instances of words with the root *πατρ-* in *Iliad* 24, only this *πατρὸς*, the errant *πατρίδα (γαῖαν)* of 24.557, and one final instance (24.592) occur in *sedes* 3. The last time that Achilles addresses Patroklos, whose name etymologizes as "glory (*kleos*) of the father (*patr-*)," the name of his dead companion occurs in *sedes* 3 (*μή μοι Πάτροκλε σκυδμαινέμεν*, Iliad 24.592) for the only time in the vocative case — a case particularly marked for Patroklos, as the narrator famously speaks it in apostrophe eight times [Allen-Hornblower 2020]. Moreover, as a phrasal search in [Chicago Homer] shows, the lines immediately surrounding the shifted *πατρίδα γαῖα* are suffused with paternal referentiality, including an interformulaic reference to the fatherly petitions of Chrysês in book 1 (*δέξαι ἄποινα Iliad* 24.555 ≈ *δέχθαι ἄποινα* 1.23) and the often parental formula, *ζώειν καὶ ὁρᾶν φάος ἠελίοιο*, spoken by Thetis about Achilles twice (18.61, 442; cf. 24.562) and by the Trojan leader Anchises after pleading for his "flourishing progeny" (*θαλερὸν γόνον*) in the *Homeric Hymn to Aphrodite* (104–5).[24] *Sedes* expectancy draws our attention to potential cross-effects of the unexpectedly placed *πατρίς*, related words in the same *sedes*, and paternal thematics in the passage.

63 Third, the fact that otherwise "formulaic" phrases, lines, or passages may contain words that are unexpectedly placed can reveal craft where before the reader may have assumed strict regularity. We see this especially at work in passages that are repeated verbatim in two or more places. Take, for example, Ajax's speech to the Achaians in *Iliad* 15.560–64, three lines of which are repeated from book 5 (15.562–64 = 5.530–32). We may visualize *sedes* expectancy and repeated phrases together using SEDES and notation from the [Chicago Homer] (*Iliad* 15.560–64):

Ἀργείους δ' ὄτρυνε [1[2[3 [4Τελαμώνιος3] Αἴας.2]4]
"[5[6[7ὦ1] [8[9φίλοι7] ἀνέρες9] ἔστε,8] καὶ6] αἰδῶ θέσθ' ἐνὶ θυμῷ,5]
[10ἀλλήλους τ' αἰδεῖσθε [11κατὰ [12κρατερὰς ὑσμίνας.11]12]
αἰδομένων δ' ἀνδρῶν πλέονες σόοι ἠὲ πέφανται·
φευγόντων δ' οὔτ' ἂρ κλέος ὄρνυται οὔτέ τις ἀλκή10]."

Great Telamonian Ajax roused the Argives: "Oh friends, be men, and put shame in your heart, and consider one another throughout the fierce struggles. When men are considerate more of them end up safe but when they flee there is neither glory nor any warcraft."

It is immediately apparent that within the passage's formulae there is variation in the expectancy of individual words. Some phrases contain words that are 64 especially expected in their given *sedes*, such as the pronoun and epithets (phrases 1–4) on line 560. We see greater disparity beginning on line 561, where, although they occur elsewhere in repeated phrases, words such as *ἀνέρες* ("men") and *αἰδῶ* ("shame") are less expected. Perhaps what is most interesting is that the lines (passage 10) repeated verbatim from book 5 contain several unexpected placements of words at *sedes* 4 (*αἰδεῖσθε, ἀνδρῶν,* and *οὔτ'*). Here, *sedes* expectancy joins other remarkable stylistic features, such as the pragmatic focus on the participles *αἰδομένων* and *φευγόντων* which begin both their lines and sentences opposed rhetorically in antithesis (fighting vs. fleeing). Rhetoric affects word order which, with SEDES, we can now see shifts words from their more typical positions and can result in "stacking" of unexpectedly placed words along particular *sedes*. SEDES also reveals that formulaic analysis with frequency data is insufficient in articulating the regular or erratic composition of constituent features of this otherwise repetitious material of epic diction, as is especially apparent in longer passages repeated verbatim.

There are likely more ways that *sedes* expectancy can alter our understanding of epic formularity, the identification and study of which will occupy future 65 research with SEDES. For example, it is clear from the brief look at *sedes* expectancy and formulae above, in particular the regularity of *πατρίδα γαῖα* at line end, that formulae too have *sedes* expectations of their own. Adaptation of SEDES code and methodology could assist in articulating these expectations quantitatively, a function that could then better identify moments when even formulae subvert expectation.

SEDES allows scholars to better locate and qualify intertextual relationships between different passages based on the *sedes* expectation of lemmata. While 66 the utility of *sedes* expectancy to intertextuality is argued more extensively in [Sansom 2021], below is a synopsis of the significance of *sedes* expectancy to words that appear in the same or different *sedes* in two texts (Figure 10).

		Sedes of lemma in two texts: Same	*Sedes* of lemma in two texts: Different
Expectancy	High / Mixed	*modello-codice* "generic model" [Conte 1986]	unclear significance
	Low	*modello-esemplare* "individual model" [Conte 1986]	unforeseen intertexts

Figure 10. Matrix of *sedes* intertextuality by expectancy.

Passages that feature the same lemma in the same *sedes* have different interpretations given the expectancy of the lemma. If a word is expected in a 67 particular position, such as *πατρίδα* in *sedes* 9 (z = +0.51 in the Archaic corpus, cf. above), its appearance there in two or more passages should not in and of itself suggest intentional reference or allusion. Rather, it is likely to be the result of typical compositional practice for hexameter poetry, what [Conte 1986] calls *modello-codice* ("generic model"). But if a word occurs in the same unexpected *sedes* in both passages, it is more likely to be a direct reference, what [Conte 1986] calls *modello-esemplare* ("individual model"). An example of this can be found in the highly unexpected placement of the lemma *σελήνη* ("moon"), which occurs only twice at *sedes* 6.5 (x = 2, Σx = 157, z = −8.80 in the whole corpus): in the Shield of Achilles passage in both *Iliad* 18.484 and Quintus of Smyrna's *Fall of Troy* 5.8. SEDES provides readers of Greek hexameter the quantitative information necessary for better distinguishing between generic and individual types of intertextuality when regarding the shared metrical position of words.

In the past, scholars would often disregard the possibility of an intertextual relationship if a word occured in different *sedes*. *Sedes* expectancy, however, 68 provides a way to identify new intertexts even in cases when there is a difference in *sedes*. When a word occurs in different *sedes* that are both relatively unexpected, in both instances the unexpected *sedes* reflects a break in the pattern of expectancy and thus merits further attention. For example, the name of the goddess *Ἀφροδίτη* occurs at *sedes* 6 in *Iliad* 9.389 (*οὐδ' εἰ χρυσείῃ Ἀφροδίτῃ κάλλος ἐρίζοι*) and at *sedes* 2 in *Od*. 20.73 (*εὖτ' Ἀφροδίτη δῖα προσέστιχε μακρὸν Ὄλυμπον*), both of which have *z*-scores of −10.0 in the whole corpus, which are among the lowest of any lemma–*sedes* pair; the other 99% of the time *Ἀφροδίτη* occurs exclusively at *sedes* 10 (x = 200, Σx = 202, z = +0.10).[25] Despite their different *sedes*, both instances thwart expectations. This fact should cause the researcher to investigate the two further, at which point other collective features emerge: shared language such as the "works of Athena" (*ἔργα δ' Ἀθηναίῃ*) in adjacent lines (*Iliad* 9.390, *Od*. 20.72) or their presence in character speech by notoriously versatile speakers, Achilles and Penelope; cf. [Martin 1989]. Conversely, in cases in which the expectancy of the word is high in both *sedes*, or mixed high and low, there is perhaps less reason to suppose intertextuality by *sedes* is a helpful mode of interpretation. Such statistical data combining word and meter could inform future quantitative intertextuality such as that of [Tesserae] and as described in [Forstall and Scheirer 2019]. By focusing interpretive efforts on statistically marked instances of words, SEDES gives readers of hexameter an additional tool with which to investigate repetitive phenomena in Greek epic such as formulae and intertexts.

Conclusion

SEDES reached its current state through a long process of evolution and iterative development. It originated as a tool to investigate sonic patterns in poetry, 69 by identifying lines with similar tonal patterns, counting letter *n*-grams, and quantifying alliteration. It later became specialized for the purpose of finding the metrical positions of all occurrences of forms of a lemma, which enabled analysis of the kind we have described in this article, albeit of only one lemma at a time. At that point, we were using texts sourced from the Thesaurus Linguae Graecae [TLG] and Diogenes [Diogenes] as a tool for lemma lookup. A concrete vision began to take shape of what we wanted to accomplish with SEDES: to compute metrical positions and other statistical information about every word in a large corpus, in an automated fashion. The difficulty of automating Diogenes lookups, and a general preference for freely available data sources, caused us to switch our source of texts from the Thesaurus Linguae Graecae to Perseus, and to begin developing programs around the direct processing of source TEI files. Automating all the steps of the processing pipeline made it feasible to start producing visualizations of an entire work at a time.

The decision to switch to Perseus texts was an important one, enabling us to take control over the entire processing pipeline. Moreover, open-access texts facilitate reproducibility of statistical results, and lower the barriers to participation by people who are not specialists in classical studies. Regardless of the source of texts, a project of this nature requires paying substantial attention to textual and data-format issues. 70

SEDES was developed in the context of a set of texts we wished to analyze and does not automatically generalize to other Greek hexameter texts. Analyzing texts outside the set we have considered will likely require manual tweaking of the processing programs, as well as adding to the tables of manual lemmatizations and scansions for words and lines that resist automatic processing. When we expanded the corpus from six to twelve texts, to incorporate Hellenistic and Imperial Greek works, we had to make adjustments to the TEI parser to account for previously unseen markup variations. [26] A beneficial side effect of this work has been the discovery and correction of various errors in the source texts. 71

SEDES provides groundwork for future corpus-level analysis of Greek hexametrical poetry, including by combining textual features (lemma, *sedes*, and metrical shape) in different ways and by expanding to additional features. More generally, it provides a framework for the analysis of metrical position in metrical language, by which the researcher assembles a corpus, analytical pipeline, and visualization of basic statistics for further development of the program, model, concept, and interpretation and appreciation of the text. Due to the limited size of the Greek hexameter corpus, the method results more from the slow work of philology using basic statistics than from the capacious abilities of machine learning. In some respects, this methodology reflects the disciplinary background of the authors. Yet the benefits of such an interest in poetic language within a limited corpus have focused our attention on a particular approach to lemmatized word frequencies correlated with metrical position. This attention to metrical position, a long-standing interest in classical philology, demonstrates a new route for future studies that seek strong correlations between linguistic phenomena and other formal features of Greek poetry. 72

Availability

SEDES source code and documentation are available at https://github.com/sasansom/sedes. The project home page, with sample visualization outputs, is https://sasansom.github.io/sedes/. 73

Acknowledgments

We express gratitude to Patrick Burns, author of CLTK's Greek lemmatization; Kyle P. Johnson of CLTK; Hope Ranker, author of the hexameter scansion module; Lisa Cerrato of the Perseus project; David Mimno; Rachel Greenstadt; Annie Lamar; Simeon Ehrlich; Stanford's Center for Spatial and Textual Analysis (CESTA); and the *Digital Humanities Quarterly* reviewers. 74

Notes

[1] [van Ophuijsen 1987, 66]: "in Hephaestion (*χώραι*) are places for feet;" "it seems most correct to say that a metron as it is actually found, a specimen, contains feet, while a metron as a species contains a definite number of places for feet…" (17). For Hephaestion's *oeuvre*, see Suda η 659 [Heath 2001], [Westphal 1866].

[2] For competing notations, see [Janse 2003], cf. [Bozzone 2014].

[3] *Caesura* ("cutting"), as in a pause or word break in a verse, with *τομή* as the Greek equivalent (cf. *LSJ* s.v. *τομή* IV.2).

[4] For localization of word-types in iambic trimeter, see [Dik 1998] and [Schein 1979].

[5] Cf. [Sansom 2021, 449–450].

[6] Other digital analyses of metrical Greek language focus more exclusively on natural language processing, such as the Classical Language Tool Kit [CLTK]; or grammatico-semantic corpus linguistics, such as The Ancient Greek and Latin Dependency Treebank [AGDT], the Homeric Dependency Lexicon [HoDeL] (cf. [Zanchi 2021]), and Ancient Greek WordNet [AGWN].

[7] For additional theorization and application of SEDES to close reading and intertextuality, see [Sansom 2021].

[8] The Perseus project is undergoing a transition toPerseus version 5.0, in which, among other changes, TEI texts contain Unicode, rather than Beta Code. At this point, SEDES sources texts from Perseus 4.0(Perseus Hopper), which require a preliminary step to decode Beta Code.

[9] TEI element l (verse line), TEI element lb (line beginning).

[10] For recent work on lemmata in digital philology, see [Forstall et al. 2015, 505–6], [Heslin 2019], [Passarotti et al 2020].

[11] Although particular inflections of a lemma may vary in metrical shape and thus only be able to occupy particular *sedes* (cf. [O'Neill 1942]), we primarily seek to identify potential patterns of expectancy generated by the repetition of a lemma's various inflections in the *sedes* they happen to occupy. To accomplish this empirical study, we need only quantify the frequency of lemmata and *sedes* at which they do, in fact, occur; cf. [Sansom 2021, 444–45].

[12] The idea of grouping words according to a definable equivalence relation immediately suggests generalizations of SEDES beyond lemmata. Other definitions of equivalence could be coarser (more words per class) or finer (fewer words per class). Our focus on lemma in this article is due to lemma being a ready means of uniting the many inflected forms of words in Ancient Greek, as well as its being a relatable concept in itself.

[13] That is, for the purpose of statistics. For the purpose of interpretation, it is nice to have the actual lemma at hand.

[14] Source code for cltk.lemmatize.grc.

[15] #29 Add a transform-ultima fallback step to the lemmatizer.

[16] [Schumann et al. 2021, 4] give a summary of approaches to automatic scansion. For ancient syllabification and scansion, see, for example, Hephaestion *Encheiridion* ("On Meter") 1.1–10 [Westphal 1866], [van Ophuijsen 1987, 31–48].

[17] Just as a lemma aggregates the various inflections of a word and allows for analysis of their metrical distribution, future work could aggregate words into coarser or finer classes. For example, one potentially useful coarser classification is word root. The separate lemmata *ἀοιδή/ᾠδή* ("song"), *ἀοιδός/ᾠδός* ("singer"), and *ἀείδω/ἀοιδιάω* ("to sing") all share the same root related to the possibly Indo-European **h2ueid* ("sing," cf. [Beekes 2010, 23]); this root may also generate patterns of expectation based on metrical position that could be assessed by *sedes* analysis.

[18] For theory of visualization, see [Braun 2018]; for other attempts by philologists to visualize repetition in Greek hexameter, see [Martin 1989, 167–70], [Chicago Homer], and [Pavese and Boschetti 2003].

[19] The steepness of the shading gradient around zero is a configurable parameter. We have tuned the default value to provide good discrimination in the interval [−2, +2], where mostz-scores are (Figure 2).

[20] The alignment of inflected words to the beginning of their first syllable reflects their membership in lemmata with less regard to the quantity of its morphological ending and, as a result, better represents the program's quantification of lemmata and interest in their patterns of expectancy.

[21] For markedness in this sense, cf. [Schein 2015, 93–116].

[22] For a helpful review of predominant definitions of formula in context of oral poetics and quantitative intertextuality, see [Forstall and Scheirer 2019, 35–7]. For the ("structural") formula and metrical position, see [Russo 1963], with the response by [Packard 1976], and for attention to metrical position and formula *passim*, [Tate 2010].

[23] *Iliad* 2.140, 158, 174, 454, 4.180, 5.687, 7.335, 460, 9.27, 47, 414, 11.14, 13.645, 15.499, 505 (*sedes* 7–10), 706, 16.832, 18.101, 23.145, 23.150, 24.557 (*sedes* 3–6); note also the variants, "in the fatherland" (*ἐν πατρίδι γαίῃ*) at *Iliad* 3.244, 8.359, and 22.404, and "from one's fatherland" (*γαίης ἄπο πατρίδος*) at *Iliad* 13.696 and 15.335.

[24] For the idea of interformularity, see [Bakker 2013, 157–69].

[25] This information was discovered by using SEDES to create a CSV file of the metrical and frequency statistics of words in the Archaic Greek epic corpus, and then organizing the spreadsheet byz-score (lowest to highest).

[26] #9: Update TEI parser to handle Hellenistic and Imperial texts.

Works Cited

AGDT Celano, Giuseppe G. A., Crane, Gregory, and Almas, Bridget. (2006–2021) "The Ancient Greek and Latin Dependency Treebank". Available at: https://perseusdl.github.io/treebank_data/

AGWN Short, William Michael, Forte, Alexander, Tauber, James, Luraghi, Silvia, and Zanchi, Chiara. "Ancient Greek WordNet". Available at: https://greekwordnet.chs.harvard.edu

Allen 1988 Allen, Robert F. (1988) "The Stylo-Statistical Method of Literary Analysis", *Computers and the Humanities* 22, pp. 1–10.

Allen-Hornblower 2020 Allen-Hornblower, Emily. (2020) "Revisiting the Apostrophes to Patroclus in *Iliad* 16". *Donum natalicium*. Available at: https://chs.harvard.edu/allen-hornblower-revisiting-the-apostrophes-to-patroclus-in-iliad-16/

Bakker 2013 Bakker, Egbert. (2013) *The Meaning of Meat and the Structure of the Odyssey*. Cambridge: Cambridge University Press.

Bakker and Fabbricotti 1991 Bakker, Egbert and Fabbricotti, F. (1991) *Peripheral and Nuclear Semantics in Homeric Diction: The Case of Dative Expression for "Spear"*, *Mnemosyne* 44, pp. 63–84.

Barnes 1986 Barnes, Harry R. (1986) "The colometric structure of Homeric hexameter", *Greek, Roman and Byzantine Studies* 27, pp. 125–150.

Beekes 1972 Beekes, R. S. P. (1972) *On the Structure of the Greek Hexameter. Glotta* 50, pp. 1–10.

Beekes 2010 Beekes, R. S. P. (2010) *Etymological Dictionary of Greek. Volume One*. Leiden: Brill.

Bozzone 2014 Bozzone, C. (2014) *Constructions: A New Approach to Formularity, Discourse, and Syntax in Homer*, Diss. University of California in Los Angeles.

Braun 2018 Braun, Steven. (2018) "Critically engaging with data visualization through an information literacy framework", *Digital Humanities Quarterly* 12(4). Available at: http://www.digitalhumanities.org/dhq/vol/12/4/000402/000402.html

Burns 2018 Burns, Patrick J. (2018) *Backoff Lemmatization as a Philological Method*, In *Digital Humanities*. Available at: https://dh2018.adho.org/en/backoff-lemmatization-as-a-philological-method/

Burns 2020 Burns, Patrick J. (2020) "Ensemble lemmatization with the Classical Language Toolkit", *SSL* 58(1), pp. 157–76. Available at: https://www.studiesaggilinguistici.it/ssl/article/view/273

CLTK Johnson, Kyle P.; Burns, Patrick, Stewart, John, Cook, Todd, Besnier, Clément, and Mattingly, William J. B. (2014–2021) "CLTK: The Classical Language Toolkit". Available at: https://github.com/cltk/cltk

Chicago Homer Chicago Homer, Eds. Ahuvia Kahane and Martin Mueller. Available at: http://homer.library.northwestern.edu

Conte 1986 Conte, Gian Biagio. (1986) *The Rhetoric of Imitation: Genre and Poetic Memory in Virgil and Other Latin Poets*, Ithaca: Cornell University Press.

Dik 1998 Dik, Helma. (1998) "Words into Verse: The Localization of Some Metrical Word-Types in the Iambic Trimeter of Sophocles", *Illinois Classical Studies*

23, pp. 47–84.

Diogenes Heslin, Peter. *Diogenes Desktop*. Available at: https://d.iogen.es/d/

Edwards 1991 Edwards, Mark W. (1991) *The Iliad: A Commentary. Volume V: books 17–20*, Cambridge: Cambridge University Press.

Elmer 2015 Elmer, David. (2015) "The 'Narrow Road' and the Ethics of Language Use in the *Iliad* and the *Odyssey*", *Ramus* 44, pp. 155–83.

Forstall and Scheirer 2012 Forstall, Christopher and Scheirer, Walter J. (2012) "Revealing hidden patterns in the meter of Homer's *Iliad*", In *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science*.

Forstall and Scheirer 2019 Forstall, Christopher and Scheirer, Walter J. (2019) *Quantitative Intertextuality: Analyzing the Markers of Information Reuse*. Cham, CH: Springer.

Forstall et al. 2015 Forstall, Christopher, Neil Coffee, Thomas Buck, Katherine Roache, and Sarah Jacobson. (2015) "Modeling the scholars: Detecting intertextuality through enhanced word-level n-gram matching", *Digital Scholarship in the Humanities* 30(4), pp. 503–15.

Fränkel 1926 Fränkel, Hermann. (1926) "Der Kallimachische und der homerische Hexameter", *Göttinger Nachrichten* 2, pp. 197–229.

Giangrande 1959 Giangrande, Giuseppe. (1959) *Conjectural Emendations*, *Rheinisches Museum für Philologie* 102(4), pp. 365–76.

Giangrande 1967 Giangrande, Giuseppe. (1967) "'Arte Allusiva' and Alexandrian Epic Poetry", *Classical Quarterly* 17, pp. 85–97.

Hagel 1994 Hagel, Stefan. (1994) *Zu den Konstituenten des griechischen Hexameters*, *Wiener Studien* 107/108, pp. 77–108.

Hagel 2004 Hagel, Stefan. (2004) *Tables Beyond O'Neill*, In Spaltenstein, F. and Bianchi, O. eds. *Autour de la Césure*, Bern: Peter Lang, pp. 135–215.

Hainsworth 1968 Hainsworth, J. B. (1968) *The Flexibility of the Homeric Formula*, Oxford: Clarendon Press.

Heath 2001 Heath, Malcolm. (2001) "Hephaestion" *Suda On Line*, Translation, 2 March 2001. Available at: http://www.cs.uky.edu/~raphael/sol/sol-entries/eta/659 (Accessed 27 December 2021).

Heslin 2019 Heslin, Peter. (2019) "Lemmatizing Latin and Quantifying the Achilleid." In Coffee, Niel, Chris Forstall, Lavinia Galli Milic, and Damien Nelis eds. *Intertextuality in Flavian Epic Poetry*, Berlin: De Gruyter, pp. 389–408.

HoDeL Zanchi, Chiara. (2021) *Homeri Dependency Lexicon*. Available at: https://hodel.unipv.it/hodel-res/

Hoover 2013 Hoover, David L. (2013) "Quantitative Analysis and Literary Studies". In Siemens, R. and Schreibman, S. eds. *A Companion to Digital Literary Studies*, Oxford: Blackwell.

Janse 2003 Janse, Mark. (2003) "The Metrical Schemes of the Hexameter". *Mnemosyne* 56, pp. 343–48.

Kahane 1994 Kahane, Ahuvia. (1994) *The Interpretation of Order: A Study in the Poetics of Homeric Repetition*, Oxford: Clarendon Press.

LSJ Liddell, Henry George, Scott, Robert, Jones, Henry Stuart, and McKenzie, Roderick. (1996) *A Greek-English Lexicon (9th edition)*, Oxford: Clarendon Press.

Martin 1989 Martin, Richard P. (1989) *The Language of Heroes: Speech and Performance in the Iliad*, Ithaca: Cornell University Press.

Martin 2000 Martin, Richard P. (2000) "Synchronic aspects of Homeric performance: the evidence of the Hymn to Apollo", In A. M. González de Tobia, ed. *Una nueva visión de la cultura griega antigua hacia el fin del milenio*, La Plata, Argentina: Editorial de la Universidad Nacional de La Plata, pp. 403–32.

McDonough 1966 McDonough, Jr., J. T. (1966) *The Structural Metrics of the Iliad*. Diss. Columbia University.

McLennan 1977 Mclennan, G. R. (1977) *Callimachus: Hymn to Zeus: Introduction and Commentary.* Rome: Ateneo and Bizzari.

O'Neill 1942 O'Neill, Jr., Eugene. 1942. "The Localization of Metrical Word-Types in the Greek Hexameter: Homer, Hesiod, Alexandrians", *Yale Classical Studies* 8, pp. 102–76.

Packard 1976 Packard, David W. (1976) "Metrical and grammatical patterns in the Greek hexameter", In Jones, A. and Churchhouse, R. F., eds, *The Computer in Literary and Linguistic Studies. Proceedings of the Third International Symposium*, Cardiff: University of Wales Press, pp. 85–91.

Parry 1928 Parry, Milman. (1928) *L'Épithète traditionnelle dans Homère; Essai sur un problème de style homérique*, Paris: Société d'éditions "Les belles lettres".

Parry 1930 Parry, Milman. (1930) "Studies in the Epic Technique of Oral Verse-Making. I. Homer and Homeric Style", *Harvard Studies in Classical Philology* 41, pp. 73–147.

Parry 1971 Parry, Milman. 1971. *The Making of Homeric Verse: The Collected Papers of Milman Parry*. Ed. Adam Parry. Oxford: Clarendon Press.

Passarotti et al 2020 Passarotti, Marco, Mambrini, Francesco, Franzini, Greta, and Massimiliano Cecchini, Flavio. (2020) "Interlinking through lemmas. The lexical collection of the LiLa Knowledge, Base of linguistic resources for Latin", *Studi e Saggi Linguistici* 68, pp. 177–212.

Pavese and Boschetti 2003 Pavese, Carlo Odo and Boschetti, Federico. (2003) *A Complete Formulaic Analysis of the Homeric Poems*, Amsterdam: Adolf M. Hakkert.

Perseus Perseus Digital Library Project, Available at: https://www.perseus.tufts.edu/hopper/

Perseus – Iliad Homer. (1920) *Homeri Opera in five volumes*. Oxford: Oxford University Press. Perseus Digital Library, Available at: https://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.01.0133 (Accessed 14 April 2021).

Perseus – Odyssey Homer. (1920) *Homeri Opera in five volumes*. Oxford: Oxford University Press. Perseus Digital Library, Available at: https://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.01.0135 (Accessed 14 April 2021).

Perseus – Theogony Hesiod. (1914) *The Homeric Hymns and Homerica with English Translation, trans. H. G. Evelyn-White. Theogony*. Cambridge, MA: Harvard University Press. Available at: https://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.01.0129 (Accessed 14 April 2021).

PhiloLogic Dik, Helma. (2008–2021) "Perseus Project Texts Loaded under PhiloLogic". Available at: https://perseus.uchicago.edu

Porter 1951 Porter, Howard N. (1951) “The Early Greek Hexameter”. *Yale Classical Studies* 12, pp. 3–63.

Ranker 2012 Ranker, Hope. (2012) “Libraries and command-line tools for metrical analysis of epic Greek hexameter”, Available at: https://github.com/epilanthanomai/hexameter

Russo 1963 Russo, Joseph. (1963) “A Closer Look at Homeric Formulas”, *TAPA* 94, pp. 235–47.

Sansom 2021 Sansom, Stephen A. (2021) “*Sedes* as Style in Greek Hexameter: A Computational Approach”, *TAPA* 151(2), pp. 439–67.

Schein 1979 Schein, Seth. (1979) *The Iambic Trimeter in Aeschylus and Sophocles: A Study in Metrical Form*. Leiden: Brill.

Schein 2015 Schein, Seth. (2015) *Homeric Epic and its Reception: Interpretative Essays*. Oxford: Oxford University Press.

Schumann et al. 2021 Schumann, Anne-Kathrin, Beierle, Christoph, and Blößner, Norbert. (2021) “Using Finite-State Machines to Automatically Scan Classical Greek Hexameter”, Available at: https://arxiv.org/abs/2101.11437

TEI Text Encoding Initiative. *P5 Guidelines* 4.2.1. Available at: https://tei-c.org/Guidelines/P5/

TLG Thesaurus Linguae Graecae, Available at: https://stephanus.tlg.uci.edu/

TLGBetaCode Thesaurus Linguae Graecae. *The TLG Beta Code Manual*. Available at: https://stephanus.tlg.uci.edu/encoding/BCM.pdf

Tate 2010 Tate, Aaron Phillip. (2010) *Modularity and the Spectrum of Formularity in the Homeric Corpus*. Diss. Cornell University.

Tesserae Tesserae, Available at: https://tesserae.caset.buffalo.edu

Visser 1988 Visser, E. (1988) “Formulae or Single Words? Towards a New Theory of Homeric Verse-making”, *Würzburger Jahrbücher für die Altertumswissenschaft* 14: 21–31.

Westphal 1866 Westphal, R. (1866) *Scriptores Metrici Graeci. Vol. 1. Hephaestionis. De Metris Enchiridion et de Poemata Libellus cum Scholiis et Trichae Epitomis, Adjecta Procli Chrestomathia Grammatica*, Leipzig: Teubner.

White 1986 White, Heather. (1986) “Greek Epic Language and Callimachus’ Hymn to Delos”, *L’Antiquité Classique* 55, pp. 316–22.

Winnington-Ingram 1963 Winnington-Ingram, R. P. (1963) *Aristidis Quintiliani de musica libri tres*, Leipzig: Teubner.

Zanchi 2021 Zanchi, Chiara. 2021. “The *Homeric Dependency Lexicon*: What it is and how to use it”, *Journal of Greek Linguistics* 21, pp. 263–97.

van Ophuijsen 1987 van Ophuijsen, J. M. (1987) *Hephaestion on Metre. A Translation and Commentary*, Leiden: Brill.