Neil Coffee/James Gawley

How Rare are the Words that Make Up Intertexts? A Study in Latin and Greek Epic Poetry

1 Introduction

Among the new approaches to literary criticism made possible by the digital humanities, macroanalysis has been perhaps the most popular and powerful. Digital tools have been used with considerable success to study authorship, textual influence, and stylistics, within large individual works and across multiple works.[1] In classics, macro-scale studies have begun to expand our understanding of intertextuality, further illuminating how ancient authors adapted and reused other texts. Classicists are beginning to identify large numbers of localized intertexts automatically, then look at the overall trends they contain, as exemplified by the contribution of Neil Bernstein to this volume.[2]

The notion of macroanalysis was developed as a digitally-enabled counterpart to microanalysis, or traditional "close reading" of individual literary passages.[3] This study employs digital methods to pursue a different form of microanalysis. "Micro" refers here not to the small number of words or works considered, but rather to linguistic features not readily evident in the course of the ordinary or even intensive reading of literature.[4] Through a microanalytical approach to intertextuality, we can begin to identify more robustly the formal features that make up localized intertexts. Theoretical discussions of intertextuality have long been concerned with the role of the author or reader in determining whether a particular piece of text recalled another, and so constituted an intertext. The formal study of intertextuality instead begins by studying intertexts recognized as such by scholars and studies their linguistic properties, such as

[1] Jockers (2013) 22.

[2] Coffee *et al.* (2012) offer an analysis of large-scale intertextuality within one work, Lucan's *Civil War*. The German eTRAP group has carried out some measurements of text reuse over a large corpus of classical Latin (Büchler *et al.* 2013).

[3] Moretti (2005) contrasts "close reading" with "distant reading," his term for what Jockers calls "macroanalysis."

[4] The linguistic reading of poetics had its earliest distinguished modern exponent in Jakobson (1960), esp. p. 351.

https://doi.org/10.1515/9783110602203-019

similarity of lemma, meaning or sound. In this way, we can determine what features make a text marked enough to constitute a meaningful intertext.[5]

In this article, we investigate the formal features of intertextuality using two sets of recognized parallels between classical epic poems. The first is a collection of 317 Latin-language parallels between Book 1 of Lucan's epic *Civil War* and the whole of Vergil's *Aeneid*, compiled from commentators and inspection of Tesserae search results.[6] The collection was edited to include only intertexts consisting of at least two lemmata shared among the two texts, where in each text the lemmata fell within the bounds of one sentence.[7] In previous work, such bigram lemma intertextuality has been estimated to account for 67% of the intertexts between *Civil War* book 1 and the *Aeneid* and so represent of a major type of intertextuality.[8] The second set consists of 376 ancient Greek parallels between Book 3 of the *Argonautica* of Apollonius Rhodius and the *Iliad* and *Odyssey* of Homer, compiled from Richard Hunter's commentary on *Argonautica* 3, also consisting only of minimum two-word parallels. The intertexts in these sets include instances of language that recalls other passages, whether to generate specific comparison of the two, or just as artistic repurposing.[9]

Within these sets, we focus one formal question: how common or rare are the words that make up intertexts? Word frequency was one important consideration in ancient poetic practice. In his treatise on *Poetics*, Aristotle advises poets to use rare words, but only occasionally. Vergil comes in for criticism in antiquity for creating a new kind of poetic affectation, or *cacozelia*, produced by using common words in uncommon combinations. Words could have marked poetic effects because they were rare in the lexicon overall. Or they could be rare within a certain literary register, as when prosaic words were used in epic poetry, or within a certain genre, as when words that Homer used only once, *hapax legomena*, were used by later epic poets like Apollonius.[10]

[5] Morgan 1977, 3, discussed by Hinds (1998) 19. Fowler (2000) 122 writes of "markedness" as one criterion for what makes up an intertext, along with "sense" (i.e., that someone has to recognize a piece of language as marked and make some sense of it for it to be an intertext). With digital methods: Büchler *et al.* (2010), Coffee *et al.* (2012), Büchler *et al.* (2013), Forstall *et al.* (2014).

[6] The full set is available at: http://tesserae.caset.buffalo.edu/blog/benchmark-data/. For more information on the methods used to obtain it, see Coffee *et al.* (2012).

[7] Where semicolons, in addition to the other usual marks of punctuation, were understood to mark sentence endings.

[8] Coffee *et al.* (2012) 415.

[9] That is, the sets include types 3–5 of the scale given in Coffee *et al.* (2012) 392–398.

[10] Aristotle *Poetics* 22 1458a – 1459a. Vergil's detractor is one M. Vipranius, quoted in Donatus's *Life of Vergil* section 44, available in Ziolkowski/Putnam (2008) 186. For prosaic words in Latin

The question of rarity appears elsewhere in this volume. Chiara Battistella and Lavinia Galli Milić focus on the way allusions shape characterization in Valerius Flaccus's *Argonautica*, but they acknowledge that previous scholarship has relied on rare words to identify some of those intertexts. Raymond Marks also relies on the importance of word-rarity as a marker of intertextuality in his study of Silius Italicus. In his quantitative study of Silius, Neil Bernstein's formula for measuring the rate of intertextuality between authors assumes that rare words close together distinguish allusions from other types of shared language. It is important to note that all of these studies involve Latin word rarity. Bernstein's contribution in particular assumes the importance of rarity based on evidence derived from previous analysis of Lucan's relationship to Vergil. As Stephen Hinds points out in his survey of scholarship on Flavian epic in this volume, we must be careful not to develop a Latin-only focus when we deploy quantitative tools to study Flavian intertextuality. To properly follow this advice, we must avoid importing assumptions about word rarity to the study of Greek intertextuality which were originally built upon studies of Latin literature. This chapter represents a first step in that direction, by comparing the intertexts of one Latin and one Greek epic poem.

Common sense suggests that intertexts are likely to be made of relatively rare words, since rare words stand out and can bring to mind the unusual passages in which they have appeared. Our results confirm this intuition, but also give it specificity, by showing the particular level of rarity of intertext words. They also reveal a contrast between Latin and Greek epic.[11]

These conclusions provide support for other work on intertextuality in this volume that relies, explicitly or implicitly, on rarity as a marker of significant intertexts in Latin, including the contributions of Battistella and Milić, Bernstein, and Marks. At the same time, our results provide a further, possibly corrective view, as Stephen Hinds warns may be necessary, by defining the difference in rare word usage in intertexts between Greek and Latin authors in the same genre.

poetry, see Axelson (1945). Wills (1996) 2 remarks upon word frequency as a marker of intertextuality, with reference to Vergilian *hapax legomena* referring to corresponding *hapax legomena* in Theocritus and Homer.

[11] The intuition about word frequency and intertextuality has for several years been incorporated into the Tesserae Project website (http://tesserae.caset.buffalo.edu/), which scores highest the n-gram matches with words that are relatively rare and close together in their respective phrases.

2 Overall Word Frequency Comparison: Apollonius and Lucan Intertexts versus Random Words

To begin, we provide a picture of the frequencies of the words in the intertexts of *Argonautica* and *Civil War* compared to the frequencies of randomly chosen words from the poems. For the random sets, two-word phrases were chosen randomly amounting to the same number of instances as contained in the respective intertext sets. In each case, we calculated word counts by first identifying the lemma for each word in a bigram, then finding the number of times it appears in the Tesserae corpus. The counts for the words of each phrase were then averaged, to give a single number representing the whole phrase. The results are illustrated in Figure 1.

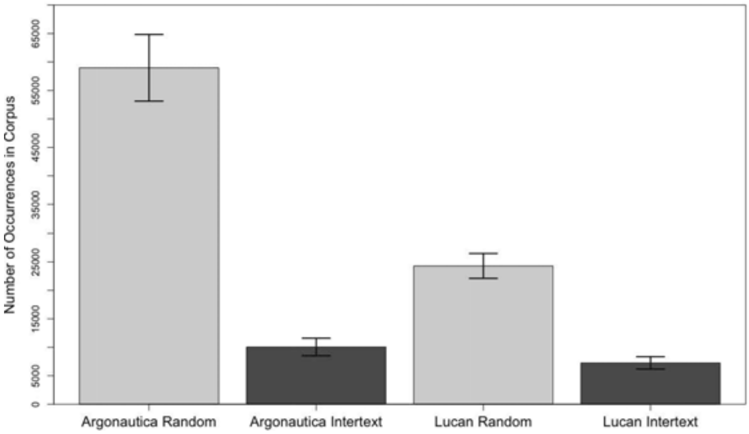

Fig. 1: Occurrences in respective Greek and Latin corpora of intertext vs. non-intertext words from *Argonautica* 3 and *Civil War* 1

The leftmost column represents the average number of occurrences in the Greek corpus of the random words from Apollonius' *Argonautica*, the second column the *Argonautica* intertext set, and so likewise for Lucan's *Civil War*. Whiskers represent the standard error measurement for these broadly distributed data sets. The words that make up intertexts are clearly rarer in the corpus than words that do not. Apollonius's intertext words occur less than a quarter as often as his non-intertext words. Lucan's occur less than half as often. Closer analysis will help to explain the difference between the intertextual practice of the two poets.[12]

3 Latin Epic: Lucan and Vergil

Figure 1 raises the question of what sort of distribution lies behind the average rarity of the intertext words of the poets. That is, given that the poet's intertext words are relatively rare overall, we might wonder whether this is because, for example, he uses many words that are at some point below the median frequency, or because he instead uses both very rare and very common words, whose frequency averages out to that point.

To answer the question, we can compare the frequencies of the individual bigrams that make up the intertext sets and random sets, in order from most common to most rare, as shown in Figures 2 and 3.[13]

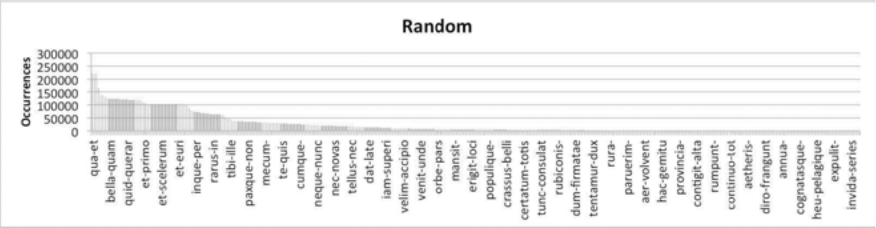

Fig. 2: Average number of appearances in Tesserae Latin corpus of words in bigrams randomly sampled from Lucan *Civil War* 1

[12] The Tesserae Latin corpus is identical with the Perseus corpus, with a small number of additional texts. For its contents, see http://tesserae.caset.buffalo.edu/sources.php.

[13] For the intertext sample, this meant dividing by two or more, depending upon the total number of words in the phrase. For the random sample, this always meant dividing by two, for the two random words selected.

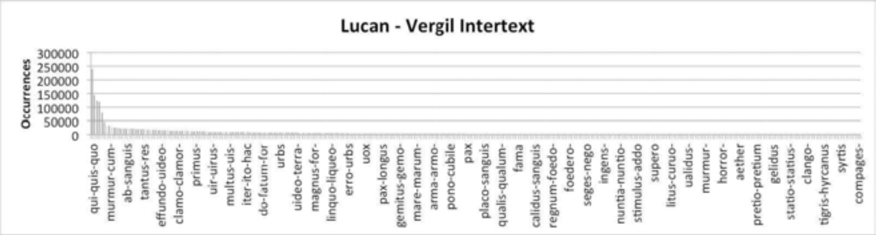

Fig. 3: Average number of appearances in Tesserae Latin corpus of words in n-grams in set of recognized parallels in Lucan *Civil War* 1 with Vergil *Aeneid*

Figure 2 represents the results for the random two-word phrases from *Civil War* 1. The y-axis shows the average number of occurrences of these words in the corpus. The x-axis lists the phrases in order from highest frequency on the left to lowest frequency on the right. A small number of phrases consist of words that are used very frequently, as shown in the left of the chart, while the majority of phrases consist of words are used relatively rarely and trail off to the right. Therefore the frequency pattern in our random sampling of Vergil is consistent with the commonly-cited analysis of Zipf, who posited that natural languages consist of a small number of words repeated often and a majority of words used rarely.[14]

Figure 3 shows the word frequencies of the *Civil War* 1 intertext phrases. The difference from Figure 2 is most evident on the left of the chart. The rapid flattening of the graph from left to right shows that intertext phrases in *Civil War* 1 are almost never made up of very common words, in contrast with random two-word phrases, which more often are. Figures 2 and 3 show that part of the rarity of Lucan's intertext vocabulary is caused by the absence of common words that occur elsewhere in his poem.

The scale of Figures 2 and 3 makes it difficult to see any difference between the rightmost part of their curves, which represents the rarest words. Figure 4 offers a closer look at this end of the frequency distribution.

[14] Zipf (1949).

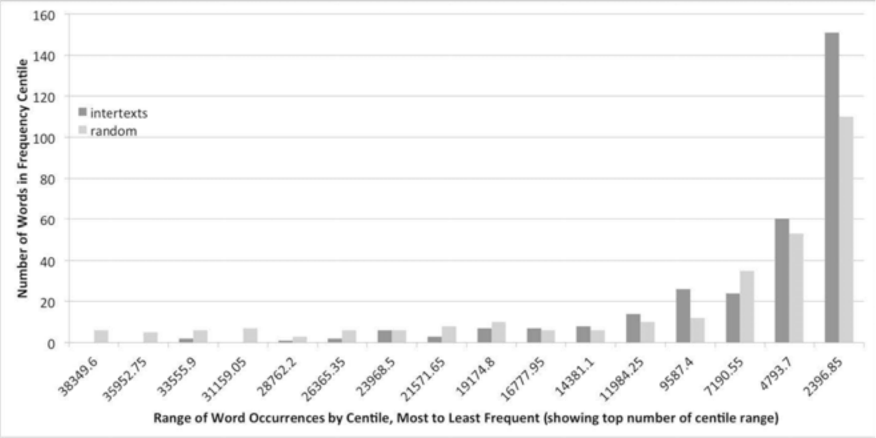

Fig. 4: Occurrences of Intertext and Random Phrases By Centile – Top 16 Centiles

Figure 4 represents the right end of the curves in Figures 2 and 3, but in a different way. Each two-word phrase, again, was given a number corresponding to the average number of occurrences of the words that made it up. These figures ran from the equivalent of 239,716 appearances in the corpus, the most frequent, to 31 appearances, the least frequent. This range was then divided into 100 equal parts. All the phrases that occur in the corpus between 239,716 and 237,289 times were counted and put into the first group, those that had occur between 237,288 and 234,890 times were counted and put into the second group, and so forth. Of these 100 groups, Figure 4 shows just the 16 with the rarest words, ending with the very rarest on the right. There are few words in Latin that occur frequently and many more that occur rarely. The graph in Figure 4 thus rises toward the right, reflecting the increasing number of words that occur only rarely.

The greater height of the rightmost column for Lucan's intertexts shows that Lucan makes use of rare words, particularly those in the 100th centile, more often than they occur randomly. So far, then, we can conclude that Lucan's intertexts with Vergil do indeed consist of words that are relatively rare, in that they show a reduced use of frequent words and increased use of rare words compared to a random sample. Put in terms of poetic technique, we would say that Lucan used relatively rare words, consciously or unconsciously, when creating intertexts with the *Aeneid*.

4 Greek Epic: Apollonius and Homer

As with Lucan, the frequency of random two word combinations from Apollonius follows a Zipf distribution, as illustrated in Figure 5.

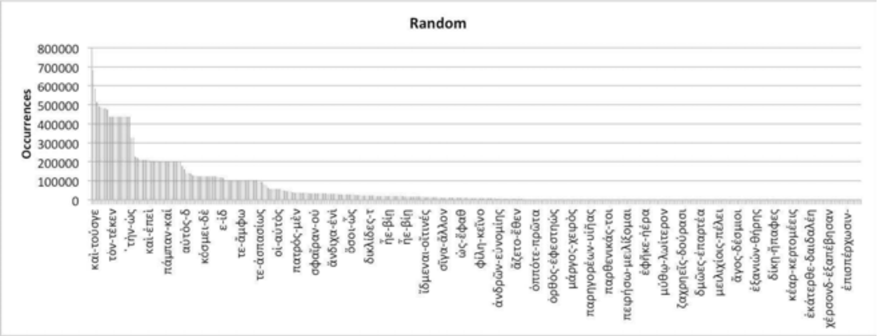

Fig. 5: Average number of appearances in Tesserae Greek corpus of words in bigrams randomly sampled from Apollonius *Argonautica* 3

As the larger range of the y-axis indicates, there are far more different words (types) in the ancient Greek lexicon than in the ancient Latin lexicon, a fact reflected in the Perseus corpora for both languages, from which the Tesserae corpus is primarily drawn. Apollonius also uses primarily Homeric language, while word frequencies were tabulated based on the entire Greek corpus.[15] Relative to each language, however, the distribution of random bigram combinations from Apollonius is similar to that for Lucan.

Likewise, the Apollonius intertext phrases also contain fewer very common words than the random sample, as illustrated in Figure 6.

[15] On Apollonius using Homeric language, see Hunter (1989) 38 with further references.

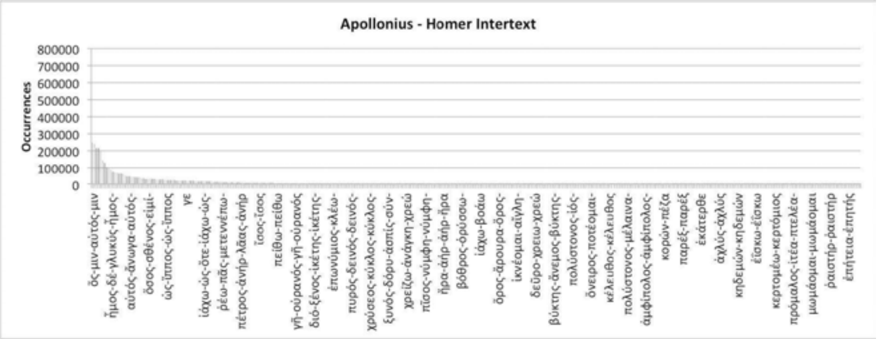

Fig. 6: Average number of appearances in Tesserae Greek corpus of words in n-grams in set of recognized parallels in Apollonius *Argonautica* 3 with Homer *Iliad* and *Odyssey*

Closer consideration shows that Apollonius prefers rare words even more than Lucan does. To begin with, none of the Apollonius parallels contain the very most frequent words, while at least a few of Lucan's do. The highest frequency words contained in Apollonius's parallels appear about 200,000 times in the Greek corpus, while the highest frequency words in our random sample of his poem appear about 400,000 times in the corpus. By contrast, the highest frequency words in Lucan's intertextual phrases appear roughly 240,000 times in the Latin corpus, while the highest frequency words in our random sample of his poem appear 220,000 times. In other words, even the most frequent words that appear in Apollonius intertexts occur only half as often as the most frequent words in his poem overall, while the most frequent words in Lucan's intertexts occur equally as often as the most frequent words in his poem overall.

We find an analogous difference if we consider the appearance of the rarest words in Apollonius's intertexts with Homer, as illustrated in Figure 7.

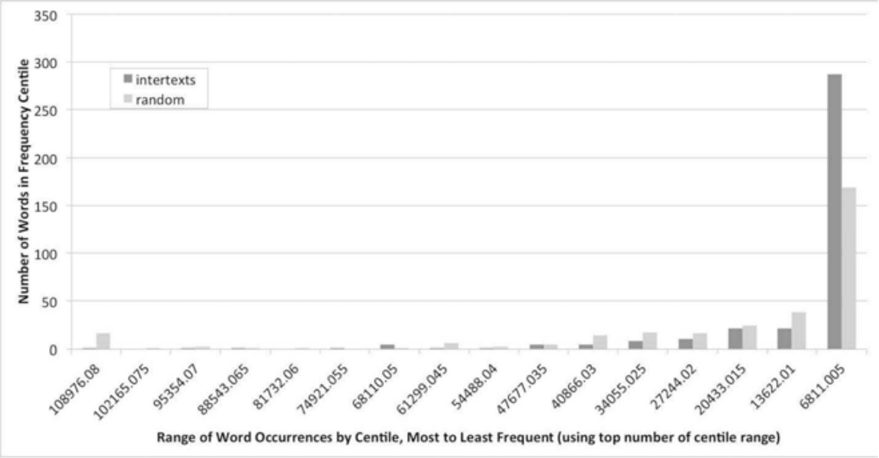

Fig. 7: Frequency Counts of Intertext and Random Phrases By Centile – Top 16 Centiles

The procedure for the creation of Figure 7 was the same as that for the creation of Figure 3, substituting the words of Apollonius for Lucan's. Once again the poet uses the rarest words more often in intertexts, which indicates that the difference in average word count seen in Figure 1 comes from two formal characteristics of intertexts: the appearance of fewer common words, and the more frequent use of extremely rare words.

In comparison to their respective random samples, the Apollonius parallels contain a greater number of words in the rarest centile than do the Lucan parallels. As comparison of the final paired columns in Figures 4 and 7 illustrates, there are only 137% more Lucan intertexts in the rarest centile than found in the random sample, while there are 170% more Apollonius intertexts in the rarest centile than found in the random sample. In this subset of rarest words, the intertextual data set is substantially larger than the randomly selected data set. This difference is larger than the analogous difference found when we compare the random and intertextual data sets in the rarest centile of words from Lucan.

5 Conclusions

The results of this study support the intuition that, for a phrase to be intertextual, it often consists of words that are relatively rare. This is not the case for every intertext. At the opening of Lucan's *Civil War*, at Book 1 line 8, his narrator asks "what madness" it was that drove Roman citizens to kill one another: *quis furor?* As commentators have observed, this is the same question asked by Ascanius in Book 5 line 670 of the *Aeneid*, deploring the madness of Trojan women burning their own ships in hopes of ending the miserable wanderings of their people. Although the word *quis* is among the most frequent in the Latin language, the phrase remains distinctive. Yet our results show that the words that make up intertexts are overall rarer than average. Furthermore, this average level of occurrence results from a particular distribution, in which the rarest words in the language appear more often and most frequent words less often. Our study also demonstrates that the intertexts of Apollonius are more strongly marked by rarity than those of Lucan. The difference might be explained by the poets' distinctive approaches to epic. Our study did not account for Apollonius's fondness for Homeric *hapax legomena*, but his frequent use of these rarest of words shows the Hellenistic-era poet had a general preference for recherché Homeric language, which extended into this longer intertexts.[16]

Lucan likewise forms his intertexts with rare words, but two factors likely kept him to a more common dictional register overall. Vergil had provided an authoritative model within the Latin epic tradition for making marked tropes from common words, his special form of *cacozelia*. Lucan may have picked up this habit in re-using the language from Vergil's epic.[17] Lucan also wrote historical, rather than mythical epic, and accordingly employed more prosaic words than Vergil did.

Alternatively, the stronger pattern of rarity in the intertexts of Apollonius might reflect a more general difference between Greek and Latin epic. Future work should replicate the current study using new benchmark sets for comparable works in each language. If intertexts appear more strongly marked by word rarity in multiple works of Greek epic, then another level of explanation will be necessary beyond the artistic choices of individual authors.

[16] For Apollonius's use of Homeric *hapax legomena*, see Fantuzzi (1988), chapter 1 and Kyriakou (1995).

[17] See Roche (2009) 51–53. This study did not consider phrase frequency, or how often two words occur together, beginning instead from the frequency of individual words that make up two-word phrases and employing as an index the average of their frequencies. Measuring phrase frequency instead could yield different results, since, as in Vergil's practice, it might be most unusual to find certain common words used together.

Bibliography

Axelson, Bertil (1945), *Unpoetische Wörter, ein Beitrag zur Kenntnis der lateinischen Dichtersprache*, Lund.

Büchler, Marco/Geßner, Annette/Eckart, Thomas/Heyer, Gerhard (2010), "Unsupervised Detection and Visualization of Textual Reuse on Ancient Greek Texts", in: *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science* 1.

Büchler, Marco/Geßner, Annette/Berti, Monica/Eckart, Thomas (2013), "Measuring the Influence of a Work by Text Re-use", in: *BICS Supplement* 122, 63–79.

Coffee, Neil/Koenig, Jean-Pierre/Poornima, Shakthi/Ossewarde, Roelant/Forstall, Christopher/Jacobson, Sarah (2012), "Intertextuality in the Digital Age", in: *TAPhA* 142.2, 381–419.

Fantuzzi, Marco (1998), *Ricerche su Apollonio Rodio: diacronie della dizione epica*, Rome.

Forstall, Christopher/Coffee, Neil/Buck, Thomas/Roache, Katherine/Jacobson, Sarah (2014), "Modeling the Scholars: Detecting Intertextuality through Enhanced Word-Level N-Gram Matching", in: *Literary and Linguistic Computing* 10.1093/llc/fqu01.

Fowler, Don (2000), *Roman Constructions: Readings in Postmodern Latin*, Oxford.

Hinds, Stephen (1998), *Allusion and Intertext: The Dynamics of Appropriation in Roman Poetry*, Cambridge.

Hunter, Richard L. (1989), *Apollonius of Rhodes*, Argonautica, *book III*, Cambridge.

Jakobson, Roman (1960), "Closing Statement: Linguistics and Poetics", in: Sebeok (1960) 350–377.

Jockers, Matthew L. (2013), *Macroanalysis: Digital Methods and Literary History*, Champagne, IL.

Kyriakou, Poulheria (1995), *Homeric* hapax legomena *in the* Argonautica *of Apollonius Rhodius: A Literary Study*, Stuttgart.

Moretti, Franco (2005), *Graphs, Maps, Trees: Abstract Models for a Literary History*, London/New York.

Morgan, Kathleen E. (1977), *Ovid's Art of Imitation: Propertius in the* Amores, Leiden.

Roche, Paul (2009), *Lucan*, De bello civili. *Book 1*, Oxford.

Sebeok, Thomas A. (ed.) (1960), *Style in Language*, New York.

Wills, Jeffrey (1996), *Repetition in Latin Poetry: Figures of Allusion*, Oxford.

Ziolkowski, Jan M./Putnam, Michael C.J. (2008), *The Virgilian Tradition: The First Fifteen Hundred Years*, New Haven.

Zipf, George K. (1949), *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, Cambridge, MA.