Homeric Formulas and Meter

Sophia Sklaviadis  
Tufts University  
sklaviadis@gmail.com

James K. Tauber  
Eldarion  
jtauber@jtauber.com

Abstract

We present a computational simulation of 20th century theories of metrics and stylistics. We construct a visualization of the formulae in the Homeric corpus, through which it is possible to explore stylistic features such as the correspondence between formulae boundaries and metrical boundaries. We use a Python script to count all repeated n-grams in the Homeric corpus. We visualize n-grams at a high level across the entire corpus. We construct a reading environment in which repeated n-grams in the text are indicated by color. The reader is able to jump to various instances of a given n-gram or to its close variants. Most significantly, n-gram variants can be defined in terms of different types of equivalence classes: n-gram equivalence class may be set by the user, based on tokens or on any function of surface forms. For example, n-grams can be visualized over orthographic normalization, lemmatization, part of speech, and meter. Our n-gram analysis is potentially useful for further purposes, including semi-automatic treebank correction based on partial n-gram matches at different equivalence classes, and a computational characterization of epic prosody.

1 Formulas as n-grams

Formulaic language is pervasive in the Iliad and the Odyssey. Formulas are repeated sequences of words that have been the object of intense scrutiny by centuries of Homerists. The patterns displayed by the formulas are complex and they have given rise to many hypotheses about the authorship of the two epics: Were there individual authors or a diffused tradition? What is the relation between the composition of the Iliad to its “sequel”? In a recent book M. Mueller analyzes formulas in terms of computationally derived n-grams (2009, The Iliad). This is a natural way of detecting the formulas in the two texts. Moreover it provides systematic data that may allow us to take a more scientific approach to the questions relating to formulas that have been posed in the history of Homeric scholarship.

Even before encountering Mueller’s work, thinking about visualizing the formulas in digital editions of the Homeric epics, we used a naïve algorithm for detecting n-grams
in the poems. Our results confirm Mueller’s, which, in turn, quantitatively confirm various well known patterns such as the high degree of similarity between the first and last books of the *Iliad* and the first book of the *Odyssey*. In FIG. 1, we can see that the last book of the *Iliad* and the first book of the *Odyssey* have among the highest frequencies of common formulas.

The rest of the paper is structured as follows. First, we discuss some of the more technical issues that we encountered in computing the *n*-grams. Then we discuss some of the 20th ce. scholarship that leads to our final visualization of the relation between *n*-grams and metrical position on the hexameter.

2  *N*-gram generation with location information

The first step in the identification of repeated *n*-grams in a text is the generation of all *n*-grams in that text for a given *n*. The *n*-gram generation code takes as input a sequence of tokens organised into lines such that each line has a unique reference ([https://github.com/jtauber/homer-ngram](https://github.com/jtauber/homer-ngram)). Neither the lines nor the references contribute to the *n*-gram generation itself but they are used to label the location of each *n*-gram for later visualisation in terms of both the line reference and the start and end offset within that line.

For example, the following input

```
ref1 A B C
ref2 D E F G H
```

results in the generation of 5-grams with location information:

```
A B C D E (ref1 offset 1 of 3 to ref2 offset 2 of 5)
B C D E F (ref1 offset 2 of 3 to ref2 offset 3 of 5)
C D E F G (ref1 offset 3 of 3 to ref2 offset 4 of 5)
D E F G H (ref2 offset 1 of 5 to ref2 offset 5 of 5)
```

Note that as well as formulas, repeated *n*-grams (particularly longer ones) may be quotations from an earlier passage. In this analysis we made no attempt to distinguish such quotations from traditionals formulas.

3  Normalisation of tokens before *n*-gram generation

The tokens in the *n*-grams (A, B, C, etc. above) need not be the word forms in the text. A normalisation process may be employed depending on the desired equivalence between *n*-grams. For example, if one wishes to consider two *n*-grams identical if they contain the same sequence of lexemes (regardless of their specific inflection), then lemmatization can be performed prior to *n*-gram generation. Normalisation may be as simple as stripping punctuation or folding case. It may, however, involve replacement of the word form with any property of that word form such as part-of-speech.
The same normalisation need not apply to all tokens. For example, one might convert all proper names to a single token leaving all else untouched if the intention is to find repeating \(n\)-grams where the specific proper name used does not matter.

## 4 Subgram removal

One problem with naïve \(n\)-gram generation for multiple values of \(n\) is that all the sub-\(n\)-grams of a repeated \(n\)-gram will show up as well, and this is generally not what is wanted. For example, the string A B C D A B C A B C D has a repeated 4-gram A B C D. Naïve generation of 3-grams gives three A B C and two B C D. The A B C is helpful to know because it not only occurs as part of A B C D but also apart from it. In contrast, B C D is only repeated because it forms part of the longer sequence A B C D. In this case, we would not want B C D to be considered a repeating \(n\)-gram. The heuristic we developed is to output an \(n\)-gram if and only if the raw \(n\)-gram count is greater than one and is greater than the count from subgrams of repeated \((n+1)\)-grams.

## 5 Data format for visualisation

The output of the \(n\)-gram generation code (with subgram removal) is a list of the following information:
• repeated n-gram id
• tokens making up that repeated n-gram (which may not be the word forms in the text)
• a list of instances of the n-gram each consisting of
  • a start reference and offset
  • an end reference and offset

We can take just the list of start/end references and offsets to produce the following visualisation, where each column represents each of the book of the *Iliad*, and the horizontal lines of each column correspond to lines in the *Iliad* (FIG. 2).

The same visualisation code can apply to visualising any ranges within the text whether they are repeated n-grams or the results of some entirely different process. For example, if search results were output with similar start/end references and offsets, they could be visualised exactly the same way. It was a deliberate choice in our implementation to support this possibility.

6 A selective review of the tradition

In the late 1920s Milman Parry produced a groundbreaking analysis of Homeric verse as structured by formulas. Parry catalogues the possible ways of splitting the hexametric line in phrases that fit its 32 possible permutations of 4 dactyls or spondees plus the more rigid last two feet, a dactyl and a spondee (Parry 1971, *The Making of Homeric Verse*). Parry famously defines the formula as “an expression regularly used, under the same metrical conditions, to express an essential idea” (Parry 1971: 13). The simplest type of formula is the combination of epithet with a proper noun, of which two notable examples are δῖος Οδυσσεύς [godlike Odysseus] (98 times), and δῖος Ἀχιλλεύς [godlike Achilles] (55 times) (Parry 1971: 84). More complex combinations of formulaic phrases are exhaustively catalogued in Parry’s papers. For example, particular epithet-noun combinations with verbs of speaking, such as πολύτλας δῖος Ὀδυσσεύς [crafty Odysseus] is used 72 times with προσέφη [said] or μετέφη [said among] and 9 times otherwise (Parry 1971: 51).

The line of the hexameter is regularly broken by a caesura at several locations: after the first syllable of the third foot, after the second syllable of the third foot if it is a dactyl, after the first syllable of the fourth foot, after the fourth foot (bucolic diaeresis), and after a run-over word at the beginning of the line (i.e. after the first syllable of the second foot). Albert Lord’s overview, *The Singer of Tales* (1960: 142), summarizes Parry’s extensive categorization of formulas as follows: “One can ... expect to find formulas of one foot and a half, two feet and a half, two feet and three quarters, two feet and three quarters, two feet and three quarters,” (Parry 1971: 13–14).
three feet and a half, four feet, and six feet in length measured from the beginning of the line, and complementary lengths measured from the pause to the end of the line.” Lord himself is most interested in establishing the oral nature of the Homeric poems. According to his arguments the first of three defining marks of oral poetry is the use of formulas. We can visualize formulas at a high level as in FIG. 2, at an even higher level to include both poems as in FIG. 1, as well as at lower levels: individual Book/Rhapsody, episode, etc.

Lord’s second mark of oral poetry is enjambment, which, according to Parry, is usually not necessary in oral poetry (compared to Virgil or Apollonius). Enjambment is more difficult to visualize because it requires extensive manual correction. However, we propose the following idea about semi-automatic detection of enjambment. Computationally, we can search for the formulas that end at the beginning of the line (i.e. before the end of the second foot). In FIG. 4 below, we see that the hexametric position with the second highest frequency of n-grams ending in that position is the first syllable of the second foot. A next step that we hope to pursue in future work is to identify which of these enjambments are syntactically/semantically necessary. Further, it will be interesting to identify syntactic and formulaic patterns that precede the enjambments.
7 Meter

Pursuing Meillet’s 1923 hypothesis that Greek Lyric is cognate to Sanskrit Vedic—to which Parry is sympathetic—Gregory Nagy derives prominent formulas of Homeric hexameter from formulas attested in Greek lyric verse (1974, Comparative Studies in Greek and Indic Meter). Nagy notes the central difficulty with his argument: “the prime of Greek Epic precedes the attested phases of Greek Lyric by a considerable span of time and, what is more, features a highly complex meter of mysterious origins” (Nagy 1974: 5).

Meillet argues that Greek Lyric is cognate with Sanskrit Vedic. The earliest attested Greek Lyric is from the 7th ce. BCE, which is later than Greek Epic. Moreover, Greek Lyric has variable and rigid meters that are relatively sparsely attested. This is in contrast with the few (approximately 6) flexible meters of the Rig-Veda’s over one thousand religious hymns (dating as far back as 2000 BCE) ritually transmitted within a priestly society. Nagy traces the development of formulaic phrases alongside reconstructed lyric predecessors of the hexameter via the Indic meters. This constitutes the bridge between later Greek Lyric and Vedic that is required by Meillet’s hypothesis. Nagy catalogues evidence that lyric meters, such as the Aeolic, which are attested only after the Homeric epics, are the building blocks of the hexameter.

Nagy makes a compelling case “that epic formulas are derived from lyric formulas appropriate mainly to Pherecratic meters”, which underlies the claim that “epic meter itself is derived from a lyric meter, the Pherecratic” (Nagy 1974: 140). In part, these hypotheses are based on the frequency of word and n-gram breaks located at the positions where dactyls would have been added to simplest Pherecractic meter (Nagy 1974: 62).

Aeolic meters are based on the choriamb: −−−−−. In its simplest form Lyric Pherecractic has the following structure:

 ListItem

 In a more complex form, Pher\textsuperscript{3d}, another choriamb is added to the simple Pherecractic, and it is from this metrical pattern that Nagy argues the hexameter is derived:

 ListItem

 In order to explore Nagy’s argument we produced the data summarized in FIG. 4.

 In support of Nagy’s argument we see the higher relative frequencies of n-grams ending at the end of the second foot of the hexameter, where the first dactyl would have occurred to produce Pher\textsuperscript{3d}, and after the first (necessarily long) syllable of the second foot (both natural locations to look for enjambment):

 ListItem

 The counts underlying FIG. 4 are calculated by first merging the metrical analysis of D. Chamberlain (2017) with the text used for repeated n-gram generation. This was
non-trivial in small part due to occasional discrepancies in the text though more significantly due to the fact that word boundaries for metrical purposes do not always correspond to the typical tokenisation used for n-gram generation (due to clitics, etc). However, once the two data sources were joined, it was possible to look at the metrical position of the ending of each repeating n-gram. These positions were labeled according their foot number within the line, their syllable number within the foot, and the length of the syllable. Hence, for example, 3.2S refers to the second syllable of the third foot where it is short.
References


