Visualizing sound as functional n-grams in Homeric Greek poetry

Christopher W. Forstall  
State University of New York at Buffalo

Walter J. Scheirer  
University of Colorado at Colorado Springs

Introduction
In this work we are looking for new ways to identify stylistic heterogeneity within the Iliad and Odyssey. As oral-formulative poetry, the Greek epic may contain special evidence of the mutual relationships between poetics, cognition, and creativity. At the same time, scholars of the digital humanities have long recognized that a successful digital criticism will find ways to return to statistics in more substantive understanding. Here, we assign n-gram counts to red, green, and blue color components in order to visualize patterns of sound within the poems. The resulting images demonstrate viscerally that several well-known “set-piece” episodes within Homer’s epics have distinct n-gram distributions.

Text, Sampling, and Controls
* Iliad and Odyssey downloaded from the Perseus Project® in XML
* concatenated, then broken into 20-line samples.

This was done 11 times: once without alteration; (original series; 10 more times, each time randomly reordering the lines of the concatenated poems before sampling (series r0...r9)

n-gram Distribution
Trying to detect which might be the most interesting n-grams, we calculated s, the number of samples in which a given n-gram occurs.

When n-grams cluster in certain samples, other samples go without; s is then lower in the original than in the r series. The lower s, the more interesting the n-gram.

We quantify this by

\[ \text{interest} = s_{\text{mean}} - \text{mean}(s_r) \]

Exactly how interesting this is depends on the variability of s; so we also consider the standard deviation of \( s_r \).

<table>
<thead>
<tr>
<th>n-gram</th>
<th>interest</th>
<th>sdev(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mm</td>
<td>-146</td>
<td>8</td>
</tr>
<tr>
<td>Su</td>
<td>105</td>
<td>9</td>
</tr>
<tr>
<td>sq</td>
<td>101</td>
<td>9</td>
</tr>
<tr>
<td>hu</td>
<td>-10</td>
<td>10</td>
</tr>
<tr>
<td>pi</td>
<td>09</td>
<td>7</td>
</tr>
<tr>
<td>li</td>
<td>116</td>
<td>7</td>
</tr>
<tr>
<td>sq</td>
<td>122</td>
<td>2</td>
</tr>
<tr>
<td>pi</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>cu</td>
<td>-26</td>
<td>5</td>
</tr>
</tbody>
</table>

Example I: three 2-grams for "horse"

Two of the most interesting 2-grams, hu and mm, are part of the word frmc, “horse.” To these we added mrfr, which had a lower interest value.

n-gram counts were scaled and translated into color values:

- red: 50%
- green: 25%
- blue: 25%

The large bright region in series original corresponds to a set-piece, the chariot race held during the funeral games for Patroclus.

Example II: Three Heroes

Next we consider three 3-grams related to three independent content elements—each is a component of a Greek hero’s name:

- red: 50% Odysseus
- green: 25% Achilles
- blue: 25% Diomedes

Each hero is foregrounded in a different part of the story.

Quantifying the Sound-Content Relationship
We measure both the number of words containing the n-gram, and the number of times each of those words occurs. The greater the lexical diversity of an n-gram, the less content-driven it is likely to be.

For example, compare the 3-grams mrfr and ovfr both frequent and of high interest values.

These graphs show the number of times each word containing a given n-gram occurs in the text as a function of that word’s rank. The right is a log-log version of the left.

Below are the top 10 words for each. ovfr shows a far greater diversity than mrfr, in large part because it contributes to some common noun and verb inflections.

Example III: A Content-Independent Pattern

Here, we used shades of grey to represent counts of a single 3-gram, ovf. The bright region corresponds to the “Catalogue of Ships” inset piece at Iliad 2.484–761.

Notes
5. By way of his patronym, “Son of Tydeus.”
6. In compiling the n-grams, we transcribe rough breathing as Latin i, soft s as subscript as (regular i); and final sigma as medial sigma as e.