Visualizing sound as functional n-grams in Homeric Greek poetry

C. W. Forstall[1] and W. J. Scheirer[2]

1. Department of Classics, State University of New York at Buffalo
(forstall@buffalo.edu)
2. Department of Computer Science, University of Colorado at Colorado Springs
(wjs3@vast.uccs.edu)

This work in progress attempts to examine internal heterogeneity in poetic language using the tools of computer-based authorship analysis. As stylometric tools become finer-grained, scholars such as Hoover (2007) and Andreev (n.d.) have turned their gaze from the characterization of an author or corpus as a whole to considerations of an author's stylistic evolution over time, and the differences between and even within individual works.

The question of the stylistic integrity of Homer's corpus is a venerable one. For centuries, diverse models, subjective as well as quantitative, have claimed to explain the composition of the Iliad and Odyssey: some scholars have seen it as the work of a single, literate genius (West, 2001, 3); others as a collective multitext, the superposition of generations of continually-changing performances handed down from one illiterate bard to the next (Nagy, 1996, 107 *ff.*). Often much of the support for these claims is the perceived homo- or heterogeneity of the text. And what is at stake in these examinations is larger than a nineteenth-century romantic notion of the artist and his genius; recent studies have used the structure of the Ancient Greek epics to examine how cognition structures spoken poetry, and how the sounds of poetry in turn give structure to our thought (Peabody, 1975, 168 *ff.*). A connection between low-level phonetic structure and larger-scale poetics is not unique to oral composition, but has been shown to be equally active in literate authors as well (Brierley & Atwell, 2010).

Previously, we have used character- and word-level functional n-grams to compare Homer's two epic poems to one another and to later written text (Forstall & Scheirer, 2010a). We have also adapted the functional n-gram to metrical data (Forstall, Jacobson, & Scheirer, 2010; Forstall & Scheirer, 2010b). In the present research, we attempt to characterize the internal sound structure of Homer's epics using functional n-grams at the word, character and metrical levels.

We ask,

- How homogeneous is the author signal within a large work?

- Do the areas frequently identified as later additions stand out?
- Can internal patterns help us understand a poem's composition?

While statistical studies of Homer have been made before, it is often difficult for the critic to move comfortably between the numbers and the subjective experience of interpreting poetry. David W. Packard, in pioneering computational work on sound patterns in Homer, cautioned that "we cannot expect to identify expressive passages merely by counting letters" (Packard, 1974). More recently, Marjorie Perloff, noting that the significance of sound is all too often overlooked in poetry criticism, has laid part of the blame on "'scientific' prosodic analysis,"' which

> has relied on an empiricist model that allows for little generalization about poetic modes and values: the more thorough the description of a given poem's rhythmic metrical units, its repetition of vowels and consonants, its pitch contours, the less we may be able to discern the larger contours of a given poet's particular practice, much less a period style or cultural construct. (Perloff & Dworkin, 2009, 2)

In this poster we focus on visualizing the data in ways that bridge the gap between empirical data and the subjective experience of interpreting poetry. We take our inspiration from work such as that of Plamondon (2009) and Mandell (n.d.) which has shown that innovation in how we visualize data is vital to connecting computing with humanities scholarship. Plamondon, in particular, used color to represent multi-parameter sound data over individual poems, allowing a subjective appreciation of the poem's structure based on objective values at a glance.

We divide the poem into samples of various sizes, and calculate n-gram frequencies for the most common features. We then use principal components analysis to concentrate the variance among fewer variables. The top three principal components are then assigned to three component color channels: red = PC1, green = PC2, blue = PC3. Each sample is visualized as a color which simultaneously represents three parameters, each potentially comprehending the most important aspects of a much larger feature set. The flow of sound in the poem may be seen as a gradient with local and large-scale variation (see Figure 1). As a control, we also treat a text of Homer's poetry in which the order of the lines has been randomized. This is in part a response to the sobering results shown by Eder (2010), who made a strong case that authorship analysis was unreliable at samples fewer than several thousand words, and was improved with randomization in sampling. It may be that smaller samples are less reliable at the author level precisely because they are sensitive to internal patterns in the text, which the randomization should smooth over.

The color gradient produced by this visualization of PCA is useful to the classical philologist precisely because of its subjective quality; yet a more difinitive analysis of the epics' internal heterogeneity is also desirable. Which sections are the "most different"? Are units which are functionally related, for example the type scences which are played out over and over by different characters

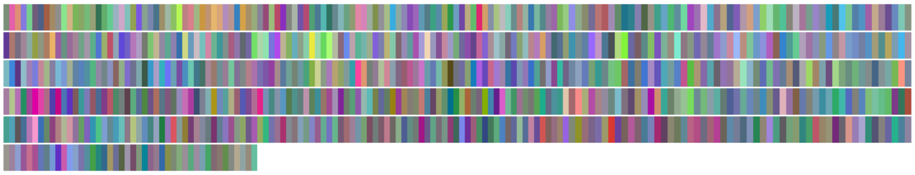

Figure 1: The entirety of the Iliad and Odyssey, in 35-line samples. The color of each sample represents the values of the first three principal components of a larger feature set composed of character bi-gram frequencies.

(Lord, 2000, 68 *ff.*), more consistent than the poem as a whole? Is the difference between the *Iliad* and *Odyssey* greater than the variation within each poem? In addressing such questions, the classicist cannot help but be biased: certain features, certain passages take on prominence at the expense of others. Here we turn to unsupervised classification for an answer which encompases all of the text at once, and has no literary bias. Using the same features as for PCA, we now perform k-means classification for various numbers of categories; again, the randomized text is used as a control. In forcing the algorithm to subdivide our sample set into an arbitrary number of classes, we make no specific assumptions about the structure of the poem. Rather we ask, how is variation distributed within the text? For example, with only two or three classes, we find that samples from all are relatively evenly distributed within and between the two poems. With more classes, we find that some tend to be found more in one poem or the other in the ordered version of the text, while in the randomized text samples from all classes are found througout. The computer-assigned, discrete classifications may be arbitrarily assigned to colors, which are then displayed alongside the continuously varying PCA data to contrast the more subjective, human-interpreted view of the poem's heterogeneity with the entirely objective computer-based analysis.

References

Andreev, V. S. (n.d.). *Patterns in style evolution of poets.*

Brierley, C., & Atwell, E. (2010). Holy smoke: vocalic precursors of phrase breaks in Milton's *Paradise Lost*. *Literary and Linguistic Computing*, *25*(2), 137–151.

Eder, M. (2010). *Does size matter? Authorship attribution, small samples, big problem.* Digital Humanities 2010. London, UK.

Forstall, C., Jacobson, S., & Scheirer, W. (2010). *Evidence of intertextuality: Investigating Paul the Deacon's* Angustae Vitae [Poster]. Digital Humanities 2010. London, UK.

Forstall, C., & Scheirer, W. (2010a). Features from frequency: Authorship and stylistic analysis using repetitive sound. *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science*, *1*(2).

Forstall, C., & Scheirer, W. (2010b). *A statistical stylistic study of Latin elegiac couplets* [Poster]. Chicago Colloquium on Digital Humanities and Computer Science. Chicago, IL.

Hoover, D. L. (2007). Corpus stylistics, stylometry, and the styles of Henry James. *Style*, *41*(2), 160–189.

Lord, A. B. (2000). *The Singer of Tales.* Cambridge, MA: Harvard University Press.

Mandell, L. (Ed.). (n.d.). *The Poetess Archive.* Retrieved 09/14/2010, from `http://www.poetessarchive.com`

Nagy, G. (1996). *Poetry as Performance: Homer and Beyond.* Cambridge, UK: Cambridge University Press.

Packard, D. W. (1974). Sound patterns in Homer. *Transactions of the American Philological Association*, *104*, 239–260.

Peabody, B. (1975). *The Winged Word: A Study in the Technique of Ancient Greek Oral Composition as Seen Principally Through Hesiod's "Works and days".* Albany, NY: SUNY Press.

Perloff, M., & Dworkin, C. (2009). *The Sound of Poetry, the Poetry of Sound.* Chicago, IL: University of Chicago Press.

Plamondon, M. (2009). *Computational phonostylistics: Computing the sounds of poetry.* Chicago Colloquium on Digital Humanities and Computer Science. Chicago, IL.

West, M. L. (2001). *Studies in the Text and Transmission of the Iliad.* Munich-Leipzig: K. G. Saur.