

Features from Frequency: Authorship and Stylistic Analysis Using Repetitive Sound

C. W. Forstall¹ and W. J. Scheirer²

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)

Abstract

A growing number of studies in the humanities now use the tools of authorship attribution to answer traditionally “subjective” questions of literary style. However, these tools still for the most part are developed by scientists with more traditional classification tasks in mind, and ultimately most scholars of literature still believe that quantified data cannot tell the whole story. A common model for digital literary analysis is to move from literature to text, from text to feature set, from feature set to index, and then finally to make an inductive leap back to the subjective world of literature. We aim to hone the tools of textual analysis to literary goals, to make the expression of digital analysis more flexible, and to strengthen that tenuous connection between feature set and literature upon which stylistics depends.

We present the *functional n-gram* as a feature well-suited to the analysis of poetry and other sound-sensitive material, working toward a stylistics based on sound rather than text. Using Support Vector Machines (SVM) for text classification, we extend the expression of our results from a single marginal distance or a binary yes/no decision to a more flexible receiver-operator characteristic curve. We apply the same feature methodology to Principle Component Analysis (PCA) in order to validate PCA and to explore its expressive potential. Having classified texts, we return to the most useful features and attempt to explain their relationship to the text in linguistic and literary terms.

Previous work leveraging word-level n-grams for authorship attribution has attempted to use as many features as possible, leaving vectors of uneven lengths. With this, a secondary problem of vector normalization must be overcome, in order to achieve accurate results with SVM learning. Moreover, it is unclear how much value is added by including seldom-used features (for example, how meaningful, stylistically, is a bi-gram composed of two proper nouns that appears once in a large text?).

Previously, n-gram smoothing has been a popular method for regularizing data of uneven feature length. Various smoothing techniques exist (Laplace, Good-Turing Discounting, Interpolation, Backoff, etc.) to estimate the probabilities of possible n-grams that are absent in a text. For the problem of predicative text analysis, this is a desirable solution; for deep stylistic analysis, it is not appropriate. It is likely that speculative feature generation will introduce stylistic inaccuracies into a text, leading to an increase in misclassification (recall the false positives rates of today’s predicative text tools, as a function of the desired word in one’s mind).

In the literature, it has been shown that with vectors of just three function words, accurate authorship attribution can be achieved. The power of small feature vectors relies on the amount of information carried by the elements at the left side of the Zipfian distribution (assuming the x axis is organized from most frequent to least frequent). It is highly probable that a limited set of any features taken from this portion of the Zipfian distribution will occur across texts that are not particularly tiny.

The functional n-gram is a new feature for authorship and stylistic analysis, whereby the power of the Zipfian distribution is realized by selecting the n-grams that occur most frequently as features, while preserving their relative probabilities as the actual feature element. The functional n-gram thus serves two purposes. On one hand, the feature vector carries a large amount of information about the text, as it reflects the most commonly occurring elements (be they words or sound). On the other, using only common features alleviates the need for feature vector normalization, thereby reducing error in the classification as well as overhead in the processing in general. We show that by using more primitive, sound-oriented features, namely, character- and phoneme-level n-grams, we are able to build accurate classifiers with the functional n-gram approach.

The texts considered span a range of times and cultures. We compared the British poets Chapman, Shakespeare, and Milton, writing in the sixteenth and seventeenth centuries. From the nineteenth century we examined the British poets Coleridge, Byron, Keats, and Shelley, and the Americans Longfellow and Poe. We examined the prose authors Jane Austen and D. H. Lawrence from the nineteenth and early twentieth centuries. Moving beyond English texts, we tested the same methods on Latin poets writing during the century or so around the birth of Christ, including Vergil and Ovid. From the Archaic and Classical Greek periods, we tested the works of Homer and Thucydides.

Features based on sounds were tested with SVMs for the first time, and produced results that were at least as good, if not better than, function-words in every experiment performed. Several rounds of experiments were performed, including a small set of novels across two authors, a large set of poems across eleven poets, and sets of works from a single author. From our results, it can be concluded that sound can be captured and used effectively as a feature for attributing authorship to a variety of literary texts. For instance, character level functional bi-grams achieve a correct binary classification rate of 99.5% with SVMs when comparing Longfellow to Poe, where using just function words achieved only 86.5% accuracy on the same test.

Principal components were used in the Greek study to provide some finer-grained analysis of the internal heterogeneity of the Iliad and Odyssey than the simple “either/or” SVM classification of phoneme and character n-grams. Not only were the two Homeric poems separable using as few as two principal components, but the arrangement of the 48 “books” which compose the two poems could be brought to bear on several long-standing arguments over the authorship of particular books. For example, claims that Odyssey XXIV was a later addition have been put forth, and disputed, since the beginning of recorded Homeric scholarship. Our results place this book squarely in the middle of the Odyssey cluster. Iliad X has been said by some to be too Odysseian, and, while the SVM classed it with the rest of the Iliad, PCA showed it near the fringe of the Iliad cluster.

PCA also shed some light on the argument that Homer’s poetry was composed without the aid of writing, another issue that has been contended for centuries. When the works of Thucydides, a literate prose historian, were projected using the principal components derived from Homer, Thucydides’ work not only clustered together but had a much smaller radius than either of the Homeric poems. This result agrees with philological arguments for the Homer’s works having been produced by a wholly different, oral mode of composition.

Based on our work in sound and analysis, we suggest several interesting areas of research for those engaged in literary studies and pattern recognition, well beyond the basic research problem of attribution. First, the detection of variation, or stylistic progression, in a single author’s work is important. We show a 79% variation between two works by D.H. Lawrence using SVMs, and a statistically significant variation between the poems of Homer with PCA. The methods presented here are well-suited to further work in style. Second, the possibility of enhancing SVM classification through marginal thresholding enhances the accuracy of the learning. A traditional ROC curve can be used as a tuning tool for operational classification systems. Third, we show that the techniques introduced are not particular to the English language. Investigating other languages is intriguing, especially foreign language poetry (how are sounds treated in other poetic forms?), as well as poetry in translation.