JDHCS 2010
Volume 1 Number 2

Features From Frequency: Authorship and Stylistic Analysis Using Repetitive Sound

Christopher Forstall, Department of Classics, State University of New York at Buffalo
Walter Scheirer, Department of Computer Science, University of Colorado at Colorado Springs

1. Introduction

A growing number of studies in the humanities now use the tools of authorship attribution to answer traditionally "subjective" questions of literary style. However, scientists still for the most part develop these tools 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.

The basis of this work exists in the repetitive stylistic nature of sound oriented texts. Authors make use of repetitive sound, either consciously or subconsciously, to emphasize an idea or phrase, or to construct a poetic form. For example, turning to D.H. Lawrence, we find in his work numerous instances of repetition in both narration and dialogue. Lawrence, in the face of critics, justified this style as a natural product of the human mind, which is inherently repetitive in its process of generating thoughts and language.[1] The following passage from *Women in Love* highlights this notion:

> "He's got *go*, anyhow."
>
> "Certainly, he's got go," said Gudrun. "In fact I've never seen a man that showed signs of so much. The unfortunate thing is, where does his *go* go to, what becomes of it?"
>
> "Oh I know," said Ursula. "It goes in applying the latest appliances!"
>
> Lawrence, *Women in Love,* Ch. 4

To be able to capture repetitive sound in a feature for authorship and stylistic analysis is of great interest. In this paper, we present the functional *n*-gram as a feature well suited to the analysis of poetry and other sound-sensitive material, working toward 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 Principal 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.

---

[1] Stewart. 1996.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

2. Related Work

Machine learning approaches to authorship and stylistic analysis have been very popular in the past few years. In Abassi and Chen's work,[2] C.45, a powerful decision tree classifier is used to identify and track authors on message boards frequented by extremists. Neural Networks have also been quite popular in this problem domain, as demonstrated in the work of Luyckx and Daelemans,[3] whereby the authors use "shallow features," those features at the token, lexical, and syntactical levels as input units to their neural net. Both C.45 and Neural Networks show promise, but are consistently outperformed by Support Vector Machines, which, unlike the previous two techniques, are able to efficiently process hundreds of thousands of features.

The first, and most important work regarding SVMs for authorship attribution is that of Diederich et al.[4] Using a corpus of newspaper articles in German, Diederich et al. show a success rate of between 60% - 80% in matching an article to a target author. In comparison to other learning techniques (Naive Bayes, Neural Nets, kNN, Decision trees), SVMs provide as much, if not more accuracy, and are shown to be more computationally efficient.

Something examined in Diederich et al., and expounded upon in Argamon and Levitan,[5] is the selection of *function words*, the components of stylometry to be used as features. Function words represent the small number of most frequently used words in a language. This phenomenon corresponds to the observation made by Zipf - that is, "…in a corpus of natural language utterances, the frequency of any word is roughly inversely proportional to its rank in the frequency table." From the basic idea of the function word, we can build more complex constructs to use as features. Diederich et al. conclude that bi-grams (two words that fall next to each other in the text) carry less information than the full function word information over the entire text. The results of Argamon and Levitan are consistent with those of Diederich et al., with success rates of between 93.2% and 99% for function word counts, and 84% - 94% for bi-grams.

More approaches to function words, *n*-grams, and other style markers for authorship attribution have been investigated in detail. Beyond lexical markers (function words), Stamatatos et al.[6, 7, 8] suggest style markers based on non-lexical measures. Utilizing text that has already been analyzed by a natural language processing tool, markers are derived from the information left by the tool (at the token-level, phrase-level, and analysis-level). The authors suggest a fusion of lexical and non-lexical

---

[2] Abassi and Chen. 2008.

[3] Luyckx and Daelemans. 2005.

[4] Diederich et al. 2003.

[5] Argamon and Levitin. 2005.

[6] Stamatatos et al. 1999.

[7] Stamatatos et al. 2000.

[8] Stamatatos et al. 2001.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

approaches to achieve the highest accuracy. Peng et al.,[9] Keselj et al.,[10] and Houvardas and Stamatatos[11] all present methods that take advantage of character level *n*-grams. Character level *n*-grams are shown to be useful not only for languages built of words composed of characters, but also for pictogram based written languages such as Chinese.

Several groups have also explored interesting applications of SVMs. Abassi and Chen are concerned with comparing SVMs to C.45 in performing authorship analysis on extremist web postings including English and Arabic language material. As mentioned above, SVMs are shown to outperform C.45 by 20% on all feature set combinations tested. Berger and Merkl[12] apply SVMs to the problem of email classification, specifically for spam filtering. The authors claim a 90% success rate in classification. Fung[13] highlights the utility of SVMs for the historian. Of the 12 "disputed" Federalist Papers, SVMs suggest James Madison is the author of all of them – a finding that is consistent with previous examination by experts. Zhao and Zobel[14] present some interesting results on a large corpus (634 texts by 55 authors) of "classic literature." The authors report success rates as high as 85%.

For specific work on sound oriented features, Plamondon[15] has suggested that a signal processing approach to "computational phonostylistics" is a useful tool for analysis. In Plamondon's work, a two-dimensional clustering technique is used to show phoneme influence in adjacent phonemes and phonemes above and below the considered poetic line. Further, Plamondon proposes a theory of "phonemic persistence," whereby a phoneme's effect on the reader will carry through to subsequent phonemes. The data is treated as a phonemic accumulation waveform for analysis. This articulation of sound oriented stylistics is more similar to popular approaches in voice biometrics for speaker recognition, in contrast to the more traditional textual analysis approaches described above.

3. Functional *n*-grams

Function words are often cited as the weakest form of style marker, but serve as the basis for more powerful style markers that improve classification drastically. Diederich et al. formalize the utility of function words, through the use of Zipf's observation[16] of word frequency distribution in large texts. In its original form, Zipf's law is given as:

$$f(r) = \frac{A}{B + r} \tag{1}$$

[9] Peng et al. 2003.

[10] Keselj et al. 2003.

[11] Houvardas and Stamatatos. 2006.

[12] Berger and Merkl. 2005.

[13] Fung. 2003.

[14] Zhao and Zobel. 2007.

[15] Plamondon. 2009.

[16] Zipf. 1949.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

where $f(r)$ is the frequency of the term of rank r in a text and A and B are positive parameters. The distribution of words based on frequency in a text is extremely uneven. For example, Figure 1 shows a sample of 10 words from the first chapter of Lawrence's *Sons and Lovers*. The most frequently used words fall to the left of the plot, and tend to be articles, adverbs, conjunctions, and pronouns. The set of these words, compared to the entire set of words in a text (composed mostly of nouns, adjectives, verbs, and adverbs), is quite small. In practice, half of the words in a text occur only once. This set is designated *hapax legomena* by linguists.

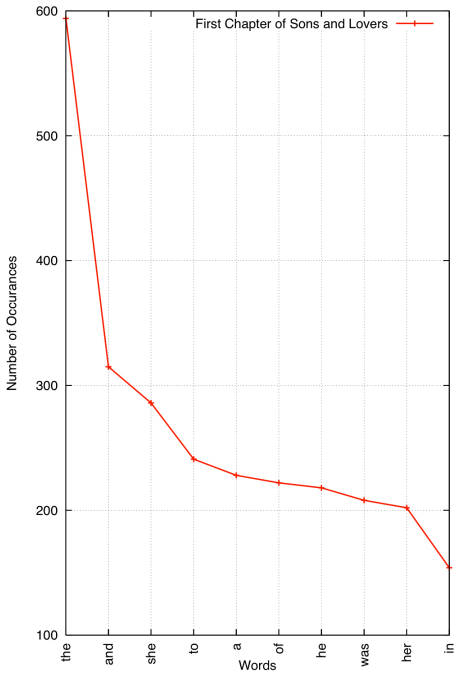

Figure 1. Selection of words ordered by occurrence from the first chapter of Lawrence's novel *Sons and Lovers*. Note that frequently used words, falling to the left of the plot, tend to be articles, adverbs, conjunctions, and pronouns. As Zipf noted, the most frequently used words represent a subset that is comparatively tiny to the complete set of words in a text.

Zipf's law can be generalized to the Zipf-Mandelbrot law:[17]

$$f(r) = \left(\frac{A}{B+r}\right)^{\frac{1}{\gamma-1}} \tag{2}$$

Equation 2 contains Equation 1 as a special case. Zipf's law is generally considered an empirical law, while the Zipf-Mandelbrot law is a tighter theoretical law defining a discrete probability distribution. Bringing all of this back to function words, and their application to authorship analysis, the idea is

[17] Mandelbrot. 1953.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

that the nature of the Zipfian distribution will be particular to an author of a text, or body of texts, as a function of their age, education, and stylistic abilities. By fitting the observed distributions of function words to known distributions by author, a probabilistic authorship determination can be made. With an understanding of what is "functional," we can turn to the actual features to be used for machine learning or other statistical classification.

In general, *n*-grams are a probabilistic feature of language: the purpose is to compute the probability of an element *e* given some history *h*, or P(*e* | *h*). The element *e* could be a word, a character, a part of speech, a punctuation mark, or any other feature that could possibly be derived from a text. If we are interested in words as a feature, we might be presented with the following: (3)

$$P(\text{wish}|\text{I}) = \frac{C(\text{I wish})}{C(\text{I})}$$

Equation 3 tells us that we're trying to find the probability of the bi-gram "I wish" occurring in a text. To calculate this probability, we simply divide the count of bi-gram occurrences by the count of all bi-grams that begin with the same word (in this case, "I"). If we were using bi-grams as an input feature for a machine learning system, the feature vector would be a collection of these calculated probabilities. To generalize Equation 3 for bi-grams: (4)

$$P(e_n|e_{n-1}) = \frac{C(e_{n-1}e_n)}{C(e_{n-1})}$$

We can generalize further for *n*-grams of any length: (5)

$$P(e_n|e_{n-N+1}^{n-1}) = \frac{C(e_{n-N+1}^{n-1}e_n)}{C(e_{n-N+1}^{n-1})}$$

Recall, that *e* need not be a word, but something more primitive, or more complex. Thus, we can compute a character level (or phoneme level, as we'll show later) bi-gram just as we did with the word level bi-gram: (6)

$$P(h|t) = \frac{C(th)}{C(t)}$$

The probability of "h" occurring, given that "t" has already occurred is calculated by dividing the count of "th" occurrences by the count of all bi-grams beginning with the character "t". The above-generalized formulas of Equations 4 & 5 apply directly to character level *n*-grams as well.

Why would *n*-grams be immediately useful for the problems of authorship and stylistic attribution? Language, by its very nature, exhibits probabilistic patterns. As has already been noted with function words (or, we could say, *uni-grams*), by fitting the observed distributions of function words to known distributions by author, a probabilistic authorship determination can be made. By doing the same with the relative probabilities of *n*-grams where *n* is greater than 1, reliable authorship determinations can be achieved. Moreover, by considering the distribution of *n*-gram frequencies, a fusion of the Zipfian approach to handling function words and the power of *n*-grams as features

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

leads to a new technique presented in this paper, the *functional n-gram*, which also solves some of the problems inherent in *n*-grams as data varies.

Poetry presents an interesting class of text for authorship attribution. While sound is important for certain novelists, it is of critical importance for poets. The job of the poet is to craft sound, not just words, in a structured or unstructured manner. Like words, the sounds of a poem may be considered features – features that are numerous (advantageous if a poem's lines are short), providing an excellent basis for statistical analysis. In the preceding discussion of character-level *n*-grams, a model is developed for a stylistic feature more primitive than a word.

The poetry of Milton, as exemplified by his epic poem *Paradise Lost*, takes the form of blank verse iambic pentameter. The iamb, in English, is commonly found to be an unstressed syllable followed by a stressed syllable. The iambic pentameter form requires that a line consist of five iambic feet in a row (ten syllables). For example, the lines:

> But God left free the Will, for what obeys
> Reason, is free, and Reason he made right,
> But bid her well beware, and still erect,
> Lest by some fair appearing good surpris'd
> She dictate false, and misinform the Will
> To do what God expressly hath forbid.
>
> Milton, *Paradise Lost* IX. 351-356

demonstrate the nature of this poetic form. Each line is unrhymed, but maintains the rhythmic flow of the iambic pentameter. Understanding the syllabic nature of this sort of poetry allows one to apply the character-level *n*-gram method to authorship attribution with great success. A bi-gram, or tri-gram can capture a discrete sound, representing a syllabic component of the line. In line 6 of the above quotation, the bi-gram "ex" captures the first syllable of the word "expressly", while the tri-gram "bid" captures the final syllable of the word "forbid".

With even more precision for capturing sound, the phoneme-level *n*-gram model is more powerful than character level-grams in some circumstances. Instead of considering characters that may or may not capture a distinct sound element ("ap" vs. "pp" in the word "appearing" – "pp" does not provide any additional sound information about "p", while "ap" captures the transition between the "a" and "p" sounds), the sounds can be considered by themselves, and treated as word-level *n*-grams during processing.

The patterns that are detectable by phoneme-level *n*-grams immediately reflect the style of the poem. The second line above, converted to phonemes, takes the following form:

R IY1 Z AH0 N IH1 Z F R IY1 AE1 N D R IY1 Z AH0 N HH IY1 M EY1 D R AY1 T

Note the repetition of the bi-gram "R IY" - it occurs three times: twice from the word "reason" and once, subtly, in the word "free." In this case, the character-level *n*-gram model can also capture the sound made by "re", but it would capture more information not directly related to the essential sound at hand (the "a" in "reason"). That is not to say that the phoneme-level *n*-gram is always superior to the character-level *n*-gram. If two phonetic texts closely resemble each other statistically,

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

the character-level *n*-gram may outperform the phoneme-level *n*-gram by capturing both sound and word content.

Bringing all of these ideas together, we can define the functional *n*-gram. Previous work leveraging word-level *n*-grams for authorship attribution has attempted to use as many features as possible, leaving vectors of uneven length. With this, a secondary problem of vector normalization must be overcome, in order to achieve accurate results with machine 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, Fung has 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, as is shown in Figure 1). 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.

4. Receiver Operator Characteristic Analysis for Support Vector Machines

Section 2 presented a brief survey of authorship and stylistic analysis approaches that rely upon Support Vector Machines for classification. While high accuracy has been reported for different data and features, accuracy still tends to be variable as a function of the experiment. What we would like is a general technique to boost accuracy in nearly all experiments. For SVMs, the placement of the margin on the hyperplane after training determines the accuracy of the classifier. In practice, we often find that classification can benefit from an adjustment of the margin if we observe a small bias towards one of the sides of the hyperplane. Figure 2 highlights this notion by showing (7) marginal adjustments from the original margin established during training (the line of separation on the bottom left of the diagram). In the first adjustment, two additional elements of the first class (represented by red circles), that had been misclassified initially, are correctly classified. In the second adjustment, all of the elements of the first class are correctly classified, but four elements of the

second class are misclassified – thus representing a trade-off in the analysis. Depending on our hypothesis, we might be willing to accept a misclassification trade-off on one side of the hyperplane, in order to increase the accuracy of the other. Ideally, we would like to increase the accuracy of both. We can understand the nature of the classification space by representing this data as a curve.

The Receiver Operator Characteristic (ROC) curve is a common fixture in engineering analysis for many different types of classification systems. By choosing a point on the curve, an analyst can tune a classification system to reduce *false acceptance*, whereby fewer misclassifications attributed to the hypothesized target class occur, or reduce *false rejection*, whereby more correct classifications for the hypothesized target class occur. To calculate the false acceptance rate (FAR) and the false rejection rate (FRR), the following simple formulas are used:

$$FAR = \frac{C_2}{(n/2)}, \quad FRR = \frac{C_1}{(n/2)}$$

where C_1 represents the count of all incorrectly classified test samples of class 1, C_2 represents the count of all incorrectly classified test samples of class 2, and n represents the total number of test samples. To collect C_1 and C_2, the ground-truth (*a priori* knowledge of the test sample's correct class) of the test samples must be known. To build an accurate curve, the number of test samples for classes 1 & 2 should be the same. The counts C_1 and C_2 vary as the margin is adjusted, and test samples are compared against the new marginal threshold. An example ROC is given in Figure 3. While this paper is the first work to present this technique for SVM based literary analysis, Halteren[18] has proposed ROC analysis for a novel literary classification system unrelated to standard machine learning techniques.

[18] Halteren. 2004.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

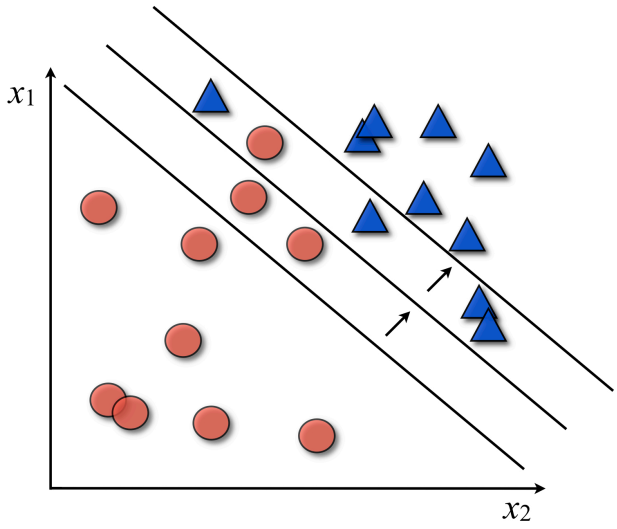

Figure 2. By adjusting the margin of an SVM hyperplane after training, it is possible to boost classification accuracy. Two marginal adjustments are depicted in this figure. The first adjustment classifies two additional elements of the first class correctly. The second adjustment classifies all of the elements of the first class correctly, and four of the elements of the second class incorrectly, thus demonstrating the trade-off that is made at certain adjustment points.

Considering the example of Figure 3 for authorship attribution, we can see how to use the ROC curve to improve classification accuracy. With the original margin after training for a classifier meant to separate the poets Longfellow and Poe, with Longfellow as our hypothesized target poet, we see a FAR of 30%, and a FRR of 10%. Articulating these rates in literary terms, we say that the FAR is the rate at which Poe is misclassified as Longfellow, and the FRR is the rate at which Longfellow is misclassified as Poe. By looking at the curve, we find a point that is suitable for minimizing the FAR, while not impacting the FRR. Thus, the FAR is reduced to 20%, and we have improved accuracy by a significant measure. We emphasize that this analysis relies on the ground-truth to readjust the SVM margin, though any data can be supplied to the enhanced classifiers after re-tuning.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

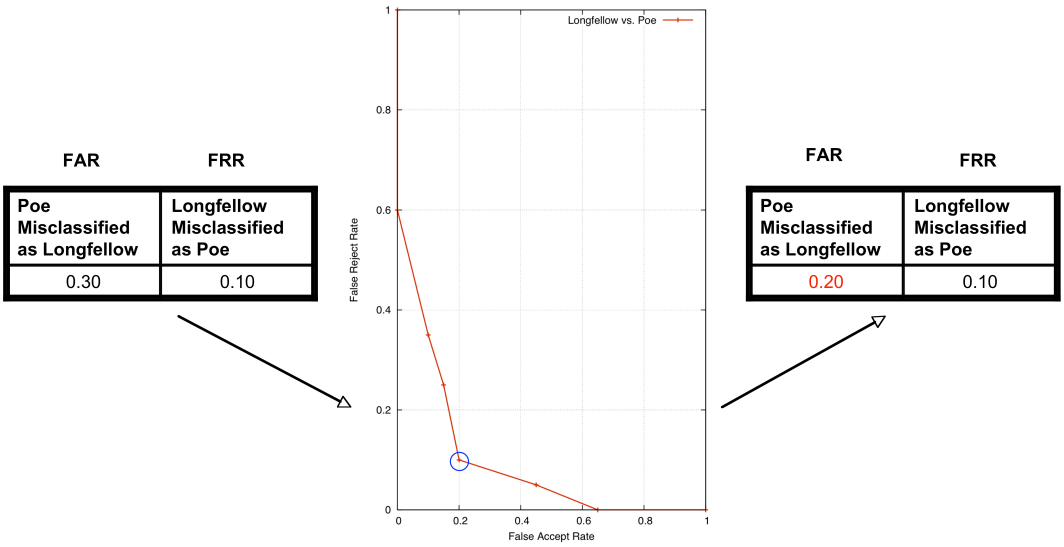

Figure 3. An example of ROC analysis for authorship attribution, with a reduction of the False Accept Rate from 30% to 20%.

SVM Experiments

To validate the functional *n*-gram, a first test using literary novels was performed. Project Gutenberg[19] is an online archive of electronic texts in plaintext and HTML formats. A small test corpus was constructed from the project's plaintext versions of the following works:

Sons and Lovers by D.H. Lawrence
Early 20th century British novel, 21,978 lines, 160,035 words.
Sense and Sensibility by Jane Austen
Early 19th century British novel, 14,731 lines, 118,542 words.

The line counts above reflect the post-processing of each text to remove the extraneous header and footer information, as well as chapter headings, included by Project Gutenberg. For testing, Austen is somewhat similar in prose format, but far enough removed from Lawrence in time to be distinctive. Thus, clear separation on the hyperplane should be observed between these two authors during the testing of each style marker. The texts were also selected based on their length. With line counts in the thousands, each text can be split into *n* texts, allowing for flexibility in generating training and test sets.

For each experiment, 10 randomized sets of input files were generated. For each author on the positive side of the hyperplane, a list of their top 10 function words was generated to produce a baseline, and utilized as the feature vector. The data for each random test is split into two sets, with 100 training samples, and 20 test samples per author (roughly 71% training, and 29% test). The

[19] http://www.gutenberg.org/wiki/Main_Page

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

SVM*light*[20] package was used for all experiments. Table 1 shows the results for the classification tests for the novels. Lawrence is on the positive side of the hyperplane, and Austen is on the negative side of the hyperplane. With an ample amount of textual data, the Project Gutenberg corpus performs perfectly, with a misclassification rate of 0%.

Beyond the baseline, the next set of experiments tested character level *n*-grams. The texts used are the same as in the first experiment, and the experimental methodology is identical. Interestingly, tri-grams outperform bi-grams at the same level of training data (100 samples). In an information theoretic sense, tri-grams capture more information about a word than bi-grams, and thus, have the potential to capture something closer to a function word (for this corpus, our best performing style marker). For these two novels, function words outperform the *n*-gram methods, however, the functional sound oriented *n*-grams worked well, and were thus validated as features. The next set of experiments will show where character-level *n*-grams excel.

Test	**Function Words** Training Vectors	**Function Words** % Misclassified	**Functional Char.-level Bi-grams** Training Vectors	**Functional Char.-level Bi-grams** % Misclassified	**Functional Char.-level Tri-grams** Training Vectors	**Functional Char.-level Tri-grams** % Misclassified
Lawrence vs. Austen	90	0.0	100	0.0575	100	0.0275

Table 1. Experimental results for two novelists. Function words perform the best here, but the results for character level *n*-grams validate the sound oriented feature approach.

Poetry proves to be a difficult challenge because of the often-limited nature of poetic texts, and the similarity of poets in select genres. Like the samples used for the literary novel experiments, Project Gutenberg was used to collect poetry samples, along with the Latin Library[21] for Latin poems. Each poet was selected based on their representation in one of three periods (Romantic, Renaissance, and Classical), as well as the availability of a large amount of their work in electronic form. Overall, the amount of text is less per poet over a span of works than for a novelist's single long novel. The poetic corpus used in the following experiments consists of the following:

Byron - Romantic British poet, 18,074 lines, 125,623 words.
Shelley - Romantic British poet, 18,652 lines, 126,383 words.
Coleridge - Romantic British poet, 2,745 lines, 17,614 words.
Keats - Romantic British poet, 2,652 lines, 19,031 words.
Longfellow - Romantic American poet, 6,081 lines, 31,065 words.
Poe - Romantic American poet, 3,082 lines, 17,495 words.
Chapman - Renaissance British translator and poet, 8,872 lines, 71,253 words.
Milton - Renaissance British poet, 10,608 lines, 79,720 words.
Shakespeare - Renaissance British poet and playwright, 2,309 lines, 17,489 words.
Ovid - Classical Latin poet, 11,998 lines, 80,328 words.
Vergil - Classical Latin poet, 10,260 lines, 65,686 words.

[20] http://svmlight.joachims.org/

[21] http://www.thelatinlibrary.com/

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

For each experiment, 10 randomized sets of input files were generated. For each author on the positive side of the hyperplane, a list of function words or functional character or phoneme-level bi-grams was generated, and utilized as the feature vector. For the phoneme-level bi-grams, each text was translated to phonemes found in the CMU Pronouncing Dictionary[22] before feature extraction and processing. Each random test is split into 50 training samples, and 20 test samples per author (roughly 56% training, and 44% test). An exception to this is the Milton vs. Chapman test, which performed poorly at 50 training samples, but had enough text to train at 100 samples, in order to yield acceptable results. Again, the SVM*light* package was used for all experiments.

The results for the function word, character-level *n*-gram, and phoneme-level *n*-gram experiments are found in Table 2. Overall, the results are very promising, with the sound features performing better than the function words in all but a single case. The length of each feature vector used fluctuates by experiment type, and between poets. The vectors were chosen to maximize separation on the hyperplane, while preserving the "functional" aspect of the features. The individual tests were designed to be realistic exercises in authorship attribution - and in some cases, represent very difficult scenarios. Some tests represent influences and relationships that are reflected in the biographies and work of the poets. All tests represent a single literary period, which lends to commonalities in style.

[22] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

Test	**Function Words** Vector Length	**Function Words** % Misclassified	**Functional Char.-level Bi-grams** Vector Length	**Functional Char.-level Bi-grams** % Misclassified	**Functional Phoneme-level Bi-grams** Vector Length	**Functional Phoneme-level Bi-grams** % Misclassified
Byron vs. Shelley	5	0.185	50	0.1775	20	0.1425
Chapman vs. Shakespeare	5	0.2025	70	0.1650	20	0.1025
Longfellow vs. Coleridge	5	0.0925	20	0.06	20	0.18
Longfellow vs. Poe	5	0.1350	20	0.005	10	0.1550
*Milton vs. Chapman	30	0.0675	70	0.0850	20	0.15
Shelley vs. Keats	20	0.20	-	-	18	0.15
Ovid vs. Vergil	50	0.0950	10	0.0375	-	-

Table 2. Results for the authorship attribution experiments on the poetry corpus. In all but a single case, the sound oriented functional *n*-grams outperform the baseline function words. 50 training vectors were used in all cases except the Milton vs. Chapman test, which used 100.

One of the more difficult tests is Byron vs. Shelley. Upon a standard reading, the texts appear remarkably similar:

> You gentlemen, by dint of long seclusion
> From better company, have kept your own
> At Keswick, and through still continued fusion
> Of one another's minds at last have grown
> To deem, as a most logical conclusion,
> That poesy has wreaths for you alone.
> There is a narrowness in such a notion,
> Which makes me wish you'd change your lakes for ocean.
> Byron, *Don Juan* 37-44

> Now Time his dusky pennons o'er the scene
> Closes in steadfast darkness, and the past
> Fades from our charmed sight. My task is done:
> Thy lore is learned. Earth's wonders are thine own,
> With all the fear and all the hope they bring.
> My spells are past: the present now recurs.
> Ah me! a pathless wilderness remains
> Yet unsubdued by man's reclaiming hand.
> Shelley, *Queen Mab* 138-145

These two selections are structurally similar—each is in the form of the iambic pentameter in eight lines, with Byron choosing to rhyme every other line. In terms of word content, both utilize standard romantic imagery (lakes, ocean, wilderness), and word forms of the period (poesy, o'er, thy). Thus, it is very easy to determine the period, but not nearly as easy to determine the author. The statistical features of these poets explain the weaker performance for this test in Table 2. The similarity between the functional phoneme-level and character-level bi-grams for Byron and Shelley is shown below (Byron is represented on the left, while Shelley is on the right). Phoneme-level bi-grams fair better for this test, with a misclassification rate of 14.25%. For very similar texts, it is possible that augmenting the functional model with additional feature information (such as *n*-grams that occur infrequently and meter) could achieve better performance.

0.2694040669200 ah0 n	0.2634725496800
0.4419285274183 dh ah0	0.4683208701563
0.6186898642414 ao1 r	0.5843537414965
0.1369433323703 t uw1	0.1079038768422
0.2185688405797 eh1 n	0.2256212256212
0.478233034571063 he	0.482253521126761
0.253358036127837 an	0.253488372093023
0.298937784522003 re	0.304950495049505
0.155569782330346 ha	0.141408450704225
0.148111332007952 ou	0.126984126984127
Byron	**Shelley**

The results of Table 2 do allow for an important conclusion—that there is value in fusing the results of multiple techniques applied to the same poets. In a severe instance, Table 2 shows no results for the test of Shelley vs. Keats. During testing, all feature vector lengths attempted resulted in a tremendously heavy bias toward the positive side of the hyperplane, enough that all produced results were not meaningful. Thus, with two other techniques available, an accurate determination can still be made for the test. In the other tests, there is variation between the results, allowing for a variety of interesting combinations and weighting schemes to be applied. This is an interesting avenue of research that has yet to be explored.

Beyond the English language, Latin poetry was also considered. The final row of Table 2 shows the results for function word and character-level *n*-gram experiments performed against Ovid's poem *Metamorphoses* and Vergil's poem *Aeneid*. Excellent results are achieved, with function words producing a misclassification rate of 9.5%, and character-level bi-grams producing a misclassification rate of 3.75%. With a phonetic Latin dictionary, the phoneme-level *n*-gram experiments could also be performed in a straightforward manner on Latin texts.

For the final set of experiments, the ROC analysis of Section 4 was performed on all of the classification data produced from the poetry experiments described above. Points along the curve generated for each classification test were chosen to minimize both FAR and FRR. Table 3 shows the improved results (marked "After") as a percentage of total classification error. The ROC analysis provides a significant reduction in error for all experiments. These results are some of the best ever reported for literary authorship attribution.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

Test	Function Words Before	Function Words After	Functional Char.-level Bi-grams Before	Functional Char.-level Bi-grams After	Functional Phoneme-level Bi-grams Before	Functional Phoneme-level Bi-grams After
Byron vs. Shelley	0.185	0.15	0.1775	0.035	0.1425	0.10
Chapman vs. Shakespeare	0.2025	0.165	0.1650	0.0375	0.1025	0.0875
Longfellow vs. Coleridge	0.0925	0.0575	0.06	0.0375	0.18	0.115
Longfellow vs. Poe	0.1350	0.105	0.005	0.0025	0.1550	0.1375
*Milton vs. Chapman	0.0675	0.04	0.0850	0.0525	0.15	0.12
Shelley vs. Keats	0.20	0.155	-	-	0.15	0.0725
Ovid vs. Vergil	0.0950	0.0575	0.0375	0.0125	-	-

Table 3. Results for the authorship attribution experiments on the poetry corpus after ROC analysis. The underlying data was the same as in Table 2. In all cases, ROC analysis provides a significant improvement in classification accuracy.

6. Case Study: The Homeric Epics

Authorship questions and stylistics have long been a part of Homeric scholarship. The origins of the *Iliad* and *Odyssey* are more ancient than any extant text of either poem, and speculation on their provenance and the identity of their author has accompanied their transmission throughout history. In the heyday of 19th century philology, some argued for a single author (or one for each work) who had penned both poems essentially in their current form; others saw stylistic evidence for multiple authorship, often imagined as the incompetent meddling of late "redactors" and the accretion of second-rate material around the original master's work.[23] Today, the emphasis has shifted somewhat, to the question of the composer's literacy and the degree of authorial deliberation. Some still view the Homeric poems as the creative product of a single, literate artist, while others see them as the culmination of a long line of illiterate and unself-conscious singers' recomposition-in-performance of traditional material.[24]

The evidence for every one of these positions is drawn from the same two poems, together comprising about 60,000 lines of dactylic hexameter, composed in a composite dialect, which mixes forms from a wide diachronic and geographic range. The arguments are essentially stylistic, often unabashedly subjective, and a clear consensus even on some basic points seems unlikely. For example, consider these two statements, made by two different scholars in a recent multi-authored

[23] The development of these views is summarized by Davison (1962) and Turner (1997).

[24] More common are that combine these two sources in various ways. See, for example, West (2001) 3ff., Powell (1997), Nagy (1996) *passim*.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

scholarly commentary on the *Odyssey*.[25] Russo, in his Introduction to Books 17—20, writes,[26] "I have assumed the text commented upon is almost entirely Homer's, and the overall cohesiveness has been created by a master storyteller who was usually in full control of his technique," while Fernández-Galiano, in his Introduction to Book 21 only a hundred pages later,[27] states with equal confidence, "it is now widely accepted that the poem had two main authors: the original poet whom critics call A, and one or more later poets known collectively as B." We here set out to delineate objectively the stylistic difference between the two poems and their internal heterogeneity, using functional *n*-grams as our feature set and a combination of SVM and PCA for the classification.

The texts used were downloaded from the Perseus Digital Library Project (where they were originally transcribed from turn-of-the-20th-century critical editions in the Oxford Classical Texts series). Samples of three sizes were made: in one run, samples were individual books of the poems. There were equal numbers of unevenly sized samples (24 books each) representing each poem. In two other runs, sample size was held constant: all books of a poem were concatenated and split into samples of 5,000 and 10,000 characters. Because the books of the *Iliad* are longer on average than those of the *Odyssey*, in these runs the *Iliad* was represented by more samples than was the *Odyssey*. The text of Herodotus, whose use is described below, was also taken from Perseus. Because the "book" length segments of this work are much longer than those of the Homeric poems, samples of 15,000 characters were created, as were the 10,000 and 5,000 character samples. This data is summarized in Table 4.

	n = 2	***n = 3***	***n = 4***
5,000	*176*	*115*	*7*
10,000	*257*	*402*	*66*
book	*323*	*926*	*354*

Table 4. Samples of Greek Works

The features used were character-level bi-, tri-, and 4-gram frequencies. Only those features common to all samples in a run were used. A reduced, "functional" *n*-gram set consisting of the most frequent features was excerpted from the full set. As with the poetry experiments described earlier, all machine learning based classification experiments here use SVM*light*. The number of features excerpted in each case was determined through trial and error, such that the lowest error was reported by SVM*light*'s cross-validation. A summary of the features is given in Tables 5 & 6.

[25] Heubeck and Hoekstra, eds. 1988.

[26] Ibid. 14.

[27] Ibid. 131.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

	***n* = 2**	***n* = 3**	***n* = 4**
5,000	130	110	7
10,000	200	240	40
book	300	430	150

Table 5. Number of *n*-grams common to all samples.

	Books (ca. 12,000—30,000 characters)	10,000 character samples	5,000 character samples
Iliad	24	57	114
Odyssey	24	41	82
Herodotus' *Histories*	64 samples of 15,000 characters	96	192

Table 6. Number of functional *n*-grams.

Each sample was classified successively, using all others as the training set. Success was calculated as the percent of all samples correctly thus classified. This was done with the full feature set and the functional feature set. The full feature set was subjected to PCA, and SVM*light* then reclassified the rotated vectors.

Graphs were prepared plotting several of the principal components. Onto graphs of the *Iliad* and *Odyssey* samples, similarly sized samples of the Histories of Herodotus, a 5th-century BC prose author were projected using the PCA models derived from Homer. This was meant to provide some comparison for both the distance between the Homeric poems and their internal variability.

Special consideration is given for character level *n*-grams in highly inflected languages, such as classical Greek. Character-level *n*-grams were proposed by Kešelj et al.[28] as a way to sidestep the need for pre-parsing inflected languages such as Modern Greek. Considering the *n*-grams common to our samples, it is possible to see how these features effectively cull common roots and endings without parsing. For example, the noun ἀνήρ, "man," can take 12 different forms in Ancient Greek, depending on its case and number. If features were identified at the word level, one would have to treat these as 12 different words, or else correctly parse all the inflected forms to identify them with a single lemma. However, because most forms of the word share the common base, ανδρ-, that 4-gram is much more common than any one inflected form. Meanwhile, the various case endings, because many nouns of the same declension share them, occur much more frequently in general than when attached to any particular base. Thus many character-level *n*-grams are common inflectional morphemes, while a few represent the bases of common nouns, adjectives or verbs. A few *n*-grams capture common preverbs and uninflected particles, which more closely resemble "function words": adverbs, conjunctions, prepositions, etc.

It is particularly in capturing the inflectional endings as independent parts of the language that character *n*-grams show their strength. Many of these morphemes play roles in Greek that in English would be played by function words: marking the case of nouns, like our prepositions;

[28] Kešelj *et. al.* 2003.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

marking person and number of verbs, like our personal pronouns; marking mood and tense, like our auxiliary verbs, among other things. Much of this information is lost at the word-level, unless one adds part-of-speech features – the result of parsing – back into the feature set, as noted by Diederich et al.

Because the feature set was restricted to features common to all samples, the size of the feature set was related to both sample size and *n*-gram length. Larger *n*-grams offer the possibility of a much larger feature set, but as sample size decreases, large *n*-grams become individually much less frequent than small *n*-grams. The morphophonetic structure of Greek limited the distribution of character *n*-grams as well. For example, the sounds that can occur at word-end in Greek are strictly limited: only the vowels and the consonants /r/, /n/, and /s/,[29] can end a word. This represents a bottleneck: since every word must have an end, these sounds are bound to occur relatively frequently. Of the most common *n*-grams, a great majority were common word endings, increasingly more with smaller *n*-grams.

	***n* = 2**	***n* = 3**	***n* = 4**
5,000 Full Feature Set	0.88	0.87	0.58
10,000 Full Features Set	0.81	0.95	0.70
Book Full Feature Set	0.88	0.98	0.98
5,000 Full Feature Set + PCA Pre-processing	0.87	0.82	0.57
10,000 Full Feature Set + PCA Pre-processing	0.94	0.98	0.73
book Full Feature Set + PCA Pre-processing	0.89	0.98	1.0
5,000 Functional Feature Set	0.89	0.87	0.58
10,000 Functional Feature Set	0.81	0.98	0.73
book Functional Feature Set	0.88	1.0	0.98

Table 7. Classification results for the full feature set, full feature set + PCA pre-processing, and the functional feature set.

Best results were achieved using book-sized samples, and tri- and 4-grams. PCA generally improved results, though not for the smallest samples. The functional feature set in many cases slightly out-performed the full one. While the best results were achieved with 4-grams and book-sized samples, 4-grams also showed the most sensitivity to sample size, becoming entirely ineffective with 5,000-byte samples, for which, conversely, bi-grams performed best. Tri-grams represented a good compromise, performing relatively well at all sample sizes, though better with larger samples.
In some cases, the PCA plots present remarkably well-separated data. In the case of book-sized samples with tri-gram frequencies, it is often possible to separate the data by eye using only two principal components. This visual approach to the data is useful for presentation to more traditional literary critics who prefer a qualitative/subjective apprehension of the texts. In the case of book-sized samples, it is possible to label the points with the familiar Ionic letters (see figure), allowing classicists to pick out those books historically identified as stylistically problematic, such as *Iliad* 10 (**Κ**), and *Odyssey* 24 (ω). We can approach, still somewhat subjectively, but now on an objective footing, literary questions like the following: "How do thematically related sections, such as the

[29] Smyth (1920) § 133. Final s includes not only *sigma*, but also the "double" consonants *psi* and *ksi*.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

books of the *Odyssey* pertaining to Telemachus, or those comprising Odysseus' narrative to the Phaiacians, group with respect to the poem as a whole," or, "Which is the most *Iliad*-like of the books of the *Odyssey*?"

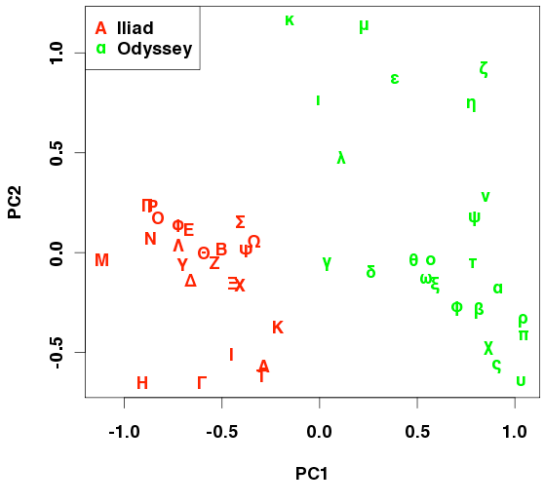

Figure 4. Books of the *Iliad* and *Odyssey*, plotted according to the first two components of the full set of tri-grams. Books of the *Iliad* are labeled with upper-case Greek letters, (A=1, B=2, etc.); those of the *Odyssey*, with lower-case letters.

The projection of Herodotus' prose onto the principal components of the Homeric poems showed that variability within the individual epics was often much greater than that within a work more uncontroversially attributed to a single, literate author, but that the difference between the two Homeric poems was at least as great as that between Homer and Herodotus.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

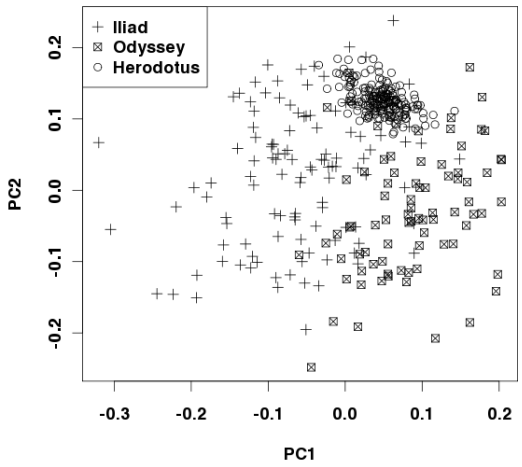

Figure 5. Herodotus' Histories (circles) projected onto principal components of the Homeric poems' bi-gram set for 5,000-character samples (*Iliad*--crosses, *Odyssey*--boxes).

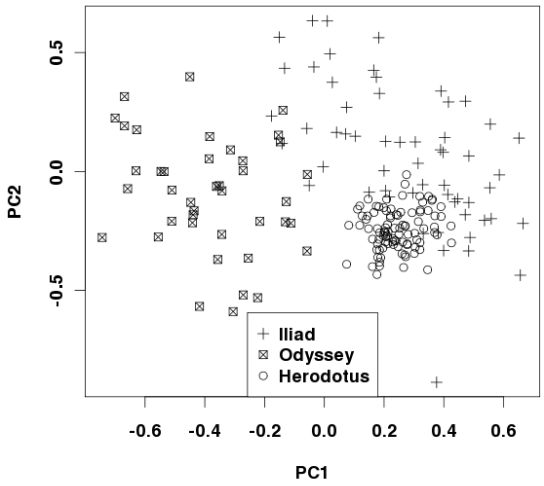

Figure 6. Herodotus' Histories (circles) versus *Iliad* (crosses) and *Odyssey* (boxes). First two principal components for full set of tri-grams among 10,000-character samples.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

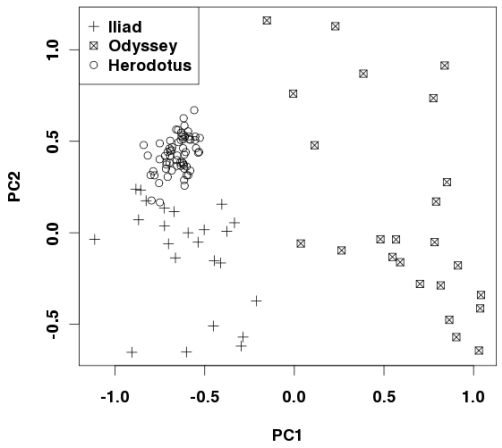

Figure 7. Herodotus' Histories (circles) versus *Iliad* (crosses) and *Odyssey* (boxes). First two principal components for full set of tri-grams among book-length samples. Herodotus has been artificially cut into 15,000-character samples.

Conclusion

In this paper, we examined the power of repetitive sound as a feature for literary authorship and stylistic analysis. We developed the functional *n*-gram as a feature well suited to the analysis of poetry and other sound-sensitive material. Using Support Vector Machines (SVM) for text classification, we extended 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 applied the same feature methodology to Principal Component Analysis (PCA) in order to validate PCA and to explore its expressive potential. Our experiments on a variety of texts spanning several different literary periods (Archaic Greek, Classical Greek, Classical Latin, Renaissance, Romantic, and Modernist) produced some of the best-reported results in the literature for sound oriented material.

Our future work will consist of further enhancements to classification accuracy and our feature set. The possibility of fusing the results from multiple tests will allow for higher accuracy, compared to a single test on its own. Rules for combining and weighting the results must be developed for such an approach. We are currently investigating the feasibility of feature-level and decision-level fusion for machine learning classification. We are also interested in understanding the influence of meter on sound, with plans to incorporate it as another feature. Finally, we are also turning our attention to sounds that occur infrequently in a text, in an approach that is opposite of the "functional" nature of this work. By fusing frequent and infrequent features, we hope to gain further insight into the role sound plays in style.

References

Abbasi, Ahmed and Hsinchun Chen. 2005. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems* 20 (5): 67-75.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

Argamon, Shlomo and Shlomo Levitan. 2005. Measuring the usefulness of function words for authorship attribution. *Proceedings of the 2005 ACH/ALLC Conference*.

Berger, Helmut and Dieter Merkl. 2005. A comparison of support vector machines and self-organizing maps for e-mail categorization. *Proceedings of AusDM 2005, the 4th Australasian Data Mining Conference*: 189-204.

Burges, Christopher. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2 (2): 121-167.

Davison, J. A. 1962. "The Homeric question." In *A companion to Homer*, 234-266.

Diederich, Joachim, Jörg Kindermann, Edda Leopold, and Gerhardt Paass. 2003. Authorship attribution with support vector machines. *Applied Intelligence* 19 (1-2): 109-123.

Fung, Glenn. 2003. The disputed Federalist Papers: SVM feature selection via concave minimization. *Proceedings of the 2003 Conference on Diversity in Computing*: 42-46.

Halteren, Hans. van 2004. Linguistic profiling for author recognition and verification. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*: 200-207.

Heubeck, Alfred and Arie Hoekstra, eds. 1988. *A commentary on Homer's Odyssey*. Volume 3: Books XVII--XXIV. Oxford University Press.

Holmes, David, Michael Robertson, and Roxanna Paez. 2001. Stephen Crane and the New York Tribune: A case study in traditional and non-traditional authorship attribution. *Computers and the Humanities* 35 (3): 315-331.

Houvardas, John and E. Stamatatos. 2006. N-gram feature selection for authorship identification. *Proceedings of the 12th Int. Conf. on Artificial Intelligence: Methodology, Systems, Applications (AIMSA'06)*, LNCS 4183: 77-86.

Keselj, Vlado, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-gram based author profiles for authorship attribution. *Proceedings of PACLING 2003, Pacific Association for Computational Linguistics*.

Luyckx, Kim and W. Daelemans. 2005. Shallow text analysis and machine learning for authorship attribution. *Proceedings of CLIN 2004, the Fifteenth Meeting of Computational Linguistics in the Netherlands*: 149-160.

Mandelbrot, Benoit. 1953. On the theory of word frequencies and on related Markovian models of discourse. *Proceedings of Symposia in Applied Mathematics* XII: 190-219.

Morris, Ian and Barry Powell, eds. 1997. *A new companion to Homer*. Brill.

Nagy, Gregory. 1996. *Poetry as performance: Homer and beyond*. Cambridge University Press.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago

Peng, Fuchun, Dale Schuurmans, Vlado Keselj, and Shaojun Wang. 2003. Language independent authorship attribution using character level language models. *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*: 267-274.

Perseus Digital Library Project. Ed. Gregory R. Crane. Tufts University. Accessed December 2008. http://www.perseus.tufts.edu.

Plamondon, Marc. 2009. Computational phonostylistics: computing the sounds of poetry. Paper presented at Chicago Colloquium on Digital Humanities and Computer Science.

Powell, Barry B. 1997. "Homer and writing." In *A New Companion to Homer,* 3-32.

Smyth, Herbert W. 1920. *Greek grammar*. Revised by Gordon M. Messing. Harvard University Press.

Stamatatos, Efstathios, Nikos Fakotakis, and George Kokkinakis. 1999. Automatic authorship attribution. *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*: 158–164.

Stamatatos, Efstathios, Nikos Fakotakis, and George Kokkinakis. 2000. Automatic text categorization in terms of genre and author. *Computational Linguistics* 26 (4): 471-495.

Stamatatos, Efstathios, Nikos Fakotakis and George Kokkinakis. 2001. Computer-based authorship attribution without lexical measures. *Computers and the Humanities* 35 (2): 193-214.

Stewart, Jack. 1996. Linguistic incantation and parody in women in love. *Style* (Spring).

Turner, Frank M. 1997. "The Homeric question." In *A New Companion to Homer,* 123-145.

Wace, Allan J. B., and Frank H. Stubbings, eds. 1962. *A companion to Homer*. Macmillan & Co.

West, Martin L. 2001. *Studies in the text and transmission of the Iliad*. K. G. Saur Verlag.

Zhao, Ying and Justin Zobel. 2007. Searching with style: Authorship attribution in classic literature. *Proceedings of the 30th Australasian Conference on Computer Science* 62: 59-68.

Zipf, George. 1949. *Human behavior and the principle of least-effort*. Addison-Wesley.

URL: http://jdhcs.uchicago.edu/
Published by: The Division of the Humanities at the University of Chicago