

# English Text Classification by Authorship and Date

Adam Belay, Mujde Pamuk, Tucker Sylvestro, and Keith Winstein

{abelay,mujde,tsyl1,keithw}@mit.edu

## Abstract

We performed two experiments with statistical techniques for classifying documents by date and author, using large bodies of publicly-available texts. In one experiment, we produced a Markov chain of every United States Supreme Court opinion ever written, and evaluated its ability to classify American judicial opinions by decade of authorship. In the other, we examined the performance of two sets of quasi-linguistic features in classifying op-ed articles from *The New York Times* among four authors with a support-vector machine. The results in each case were encouraging. With the Markov chain, we could correctly identify the decade of authorship of a Supreme Court opinion within one decade 85 percent of the time. With the two quasi-linguistic feature sets, we were able to measure the equivocation between pairs of authors and observe some interesting effects when more features were collected.

## 1 Introduction

English is changing, and people use English differently. Scholars have used these differences in stylistic expression in order to argue about the true authorship of Shakespearean manuscripts, the Federalist papers, and the Old Testament.

In modern times, even the simplest analysis reveals stylistic changes over time. Consider the average length of a top-40 paper’s title in the long-running Westinghouse, now Intel, high-school science contest (Figure 1). Every year for the last half-century, the best American high-school science students have been, on average, 0.7 letters more long-winded in ti-

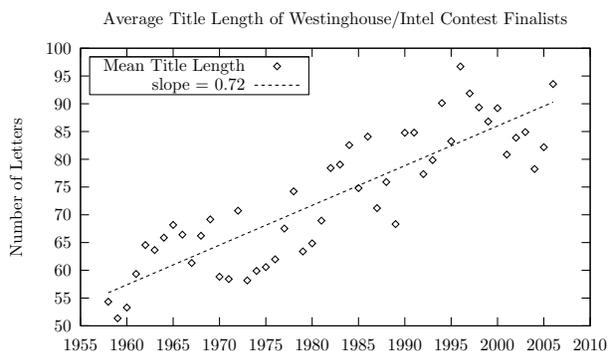


Figure 1: Trend in scientific paper title length.

ting their entries than the previous year.

We sought to answer four questions about two kinds of stylistic variations in English text: the change in American English over time since the 1790’s, and the differences between contemporary authors in the last ten years.

Regarding changes over time, we asked: (1) “Given an arbitrary English document, can we identify the year in which it was written?” (2) “How have different strains of English evolved differently over the last 215 years? Are some speakers consistently ‘ahead of the curve’ and others ‘behind the times’ in their modes of expression?”

To explore differences between authors, we looked at: (3) “Given a training set of writing from a group of authors, what features or techniques supply the most predictive information toward identifying the author of an unknown document?” And finally, (4) “How well can authors be distinguished by stylistic features?”

We focused our analyses on four corpora obtained

from the Lexis-Nexis database. For questions 1 and 2, we analyzed every signed plurality opinion from the United States Supreme Court since 1789 ( $N = 25,765$ ), every Supreme Court dissenting opinion (5,488 decisions included dissents), and every opinion issued by the trial-level federal District of Massachusetts since 1808 ( $N = 17,496$ ).

For questions 3 and 4, we examined every article published in *The New York Times* from 1995 through 2005 by four opinion columnists: William Safire ( $N = 991$ ), Maureen Dowd ( $N = 1,092$ ), Paul Krugman ( $N = 635$ ), and Thomas L. Friedman ( $N = 850$ ).

## 2 Related Work

The difficulty of finding features that can correctly classify texts with respect to properties such as chronological order has prompted considerable research. More specifically, author identification has been a well-studied problem for application in areas such as historical, religious, literature text studies and forensics with much recent emphasis on digital crimes.

The assumption behind text analysis is that individuals have differing ways of writing and there exist a set of useful features that given an anonymous text, one can derive certain characteristics of the author even if not the author himself. A methodological question arises in the quest to identify these features, namely for a given application domain and constraints within that domain, are there consistently universal features that can perform reasonably well across author identification, in an arbitrary fashion?

Based on this methodological questions, there are two opposing sides. Proponents of the first approach claim, given the constraints of an application domain, it is possible to find well-functioning identifiers. By application domain constraints, we refer to special markers used, limitations of corpora, topic constraints etc. For example, in forensics, the excerpts in analysis are very short in size, so that whatever classifier is found to differentiate between different suspects of a writing should perform well within that constraint. [6]

The other approach posits, the identifiers for each

text analysis should change their classifiers based on what is performing better in classification accuracy for the analysis at hand. The body of research by the proponent of this claim started out with Burrows, using Principal Component Analysis to identify which markers or combinations of markers can discriminate between authors in a given case. PCA aims to reduce a large set of markers to a smaller number of components which can account for as large a range of variance in the dataset as possible. The problem with this approach is it takes on no theoretical basis of what kind of identifier works in which case. [2]

There have been numerous features in texts that performed reasonably successful over different corpora and across different classification problems, which lead to a proliferation of multivariate techniques such as decision trees, neural networks, Bayesian probabilistic approaches, and support-vector machines. [8]

Particular interest has focused on support-vector machines, which are based on the concept of decision planes that define decision boundaries and have been used in text analysis studies with success. The SVM concept is based on the idea of structural risk minimization as opposed to empirical risk minimization used by conventional classifiers — i.e., minimizing training set error, which does necessitate a minimum generalization error. Therefore, SVMs have theoretically greater ability to generalize than conventional classifiers.[5]

With that being said, we turn to features that have been shown in the literature to be successful in text analysis and more specifically authorship identification.

Formal approaches to authorship identification began with the use of stylometric features. Stylometry is quantitative and computational focusing on readily computable and countable features.

Stylometry originates from Augustus Morgan's suggestion in 1851 that Biblical authors can be identified by the fact that one might use different length of words than the other. [1] The classification of the Federalist papers by Mosteller and Wallace was a major breakthrough in modern stylometric studies, with the use of Bayesian classifiers to differentiate authors based on their relative use of a set of function words.

[10]

Various stylometric features have been proposed that attempt to determine authorship within a set of different authors. Some of these features, such as hapax legomena, are lexical, where a given differentiating words or bag of words may classify authors according to categories such as identity, affect, or gender. [6, 7, 9]

Vocabulary richness and repetition are among other prominent measures. Of particular interest is Zipf’s pioneering studies on word distribution frequencies. Richness means how frequently rarer words are used, and repetition examines how frequently commoner words used. However, these approaches are shown to be dependent on the size of text, thus they cannot be successfully applied to highly heterogeneous samples.

Syntactic analysis is another approach to syntactic stylometry that has been thoroughly investigated. In syntactic analysis the features can be part of speech and orthographic. Related to syntactic analysis is sentential complexity with the assumption that people have differing abilities to produce varying degrees of sentential complexity. [1]

Another technique is the use of first order Markov chains of characters used in the text. In this approach, there are transition matrices showing given a letter what is the probability of another letter following this letter as entries in the matrix. An average transition matrix is then constructed for each author (or any other classification measure). In order to predict the authorship of a new text, a probability for each author is computed. Final assignment to a particular author is made by ranking probabilities. This technique has produced impressive results in recent research. [3,4]

### 3 Differences in Time

To answer our questions 1 and 2, concerning changes in English over time, we sought out a single source of written English documents that lasted as long as possible. We set upon the judicial opinions of the United States Supreme Court, whose 110 justices is-

sued 25,765 signed plurality opinions<sup>1</sup> between 1789 and May 15, 2006, at a roughly steady rate.<sup>2</sup> The majority opinions average about 2,000 words each until about 1950, and about 4,000 words since.

Like Khmelev and Tweedie ([4]), we used a Markov chain to characterize the opinions. We grouped the opinions by decade and withheld every tenth opinion from each year from the training set. But while Khmelev and Tweedie used a first-order chain of letters and spaces, we tried Markov chains<sup>3</sup> with histories of order 1, order 2, and order 4.

For each of the three orders, we calculated the Markov coefficients separately for each decade of opinions, creating 20 Markov models (for the 1810’s through the 2000’s) per order.<sup>4</sup> Then we calculated the Markov coefficients for the withheld opinions and selected the model with the largest *a posteriori* probability of having produced the opinion.<sup>5</sup>

### Applied to Supreme Court

The Markov models turned out to perform well at identifying the correct decade of issue of a Supreme Court opinion. We believe that increasing the order of the Markov model will eventually lead to histories that are too large to accumulate enough counts to be meaningful (lowering the predictive power of our technique), but we did not actually see this effect

---

<sup>1</sup>By “signed plurality opinion,” we mean the first opinion in a decision reported in the *United States Reports*, which opinion must be signed by one of the justices. This excludes, for instance, *Bush v. Gore*, 531 U.S. 98 (2000), where the plurality opinion was “per curiam.” It also excludes concurrences and dissents. Finally, we excluded block-quoted material from our analyses.

<sup>2</sup>Through the 1840’s, the Supreme Court would publish about 350 signed majority opinions in each decade. Every decade since 1850 has had at least 750 signed majority opinions, with the 1880’s claiming the prize for the most, at 2,667.

<sup>3</sup>Of letters, spaces, and quotation marks.

<sup>4</sup>We excluded the 1780’s, 1790’s, and 1800’s because there were not enough opinions to produce reliable coefficients at order 4.

<sup>5</sup>There are some edge cases. If a history occurred in the test specimen but never in the document, we did not factor that state into our calculation of *a posteriori* probability. If a state transition occurred in the test specimen that never occurred in the model, we treated it as if the transition had instead occurred *once* given the same history. This may be similar to the “pseudo-counts” used by others to deal with this problem.

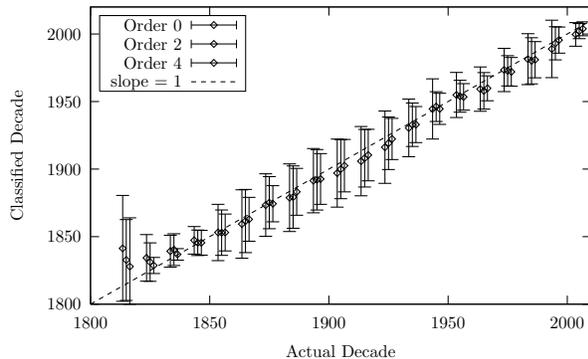


Figure 2: Mean and standard deviation of classifications, per decade, for each Markov-chain order. Note improvement in standard deviation as order increases (order 4 is rightmost within each decade).

with our order-1, order-2, or order-4 analyses. The order-4 Markov model consistently produced the best results:

Order	% within 15 yrs.	Error stddev
1	66 %	23.1 years
2	76 %	19.0 years
4	86 %	15.5 years

Because the order-4 chain was consistently superior in classifying unknown opinions to the correct decade, we focused the rest of our analysis on order-4 models.

We also plotted the “confusion matrix,” which shows the entire probability distribution for the order-4 model, that is, the probability that a document from one decade will be identified as coming from another (Figure 3). The matrix is strongly diagonal, again indicating the success of this model.

## Applied to District Court

Is the Markov model actually keyed to the decade of issuance, or are we just identifying features of the particular justices who served at various times?

One way to try to answer this question is to use our Supreme Court Markov model as a benchmark and evaluate other streams of text, written by different authors, against its metric. We downloaded the

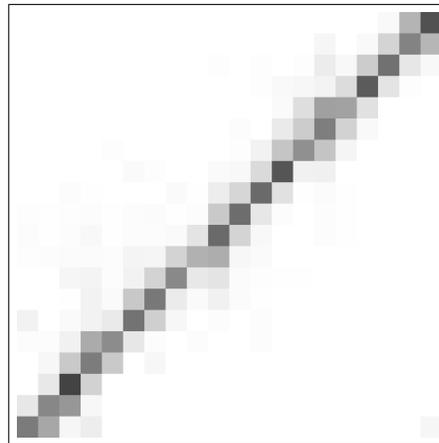


Figure 3: “Confusion matrix” for the order-4 Markov chain, indicating the probability that one decade’s document will be identified as another’s. The total ink density along any row is a constant. The strong diagonal indicates that there is not much uncertainty.

17,496 published decisions in the Lexis database from the United States District Court for the District of Massachusetts and attempted to classify their decade of issuance, using an order-4 Markov model that was trained on the Supreme Court’s plurality opinions.

This court is subordinate by two levels to the U.S. Supreme Court, and we hoped that looking at the trend in classifications might help us assess whether the federal trial court is “behind the times,” always playing catch-up to new modes of expression advanced by the Supreme Court, or perhaps the district court is “ahead of its time,” with younger judges writing in new ways that have yet to reach the Supreme Court.

Figure 4 shows the results. Surprisingly, since the Civil War, the younger judges on the District of Massachusetts have been behind the curve of the old foegys on the Supreme Court, by about 10 to 20 years!

As a caution, we agree with [4] that a first-order Markov chain is “linguistically microscopic,” and we would further concede that a fourth-order Markov chain is still “jurisprudentially microscopic.” So it’s not clear how much can be drawn from the Massachusetts court’s lagging behavior.

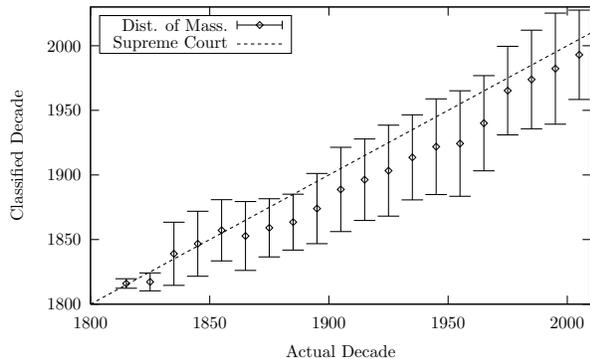


Figure 4: Classifications per decade of Massachusetts federal-court decisions, using an order-4 Markov model trained on the Supreme Court.

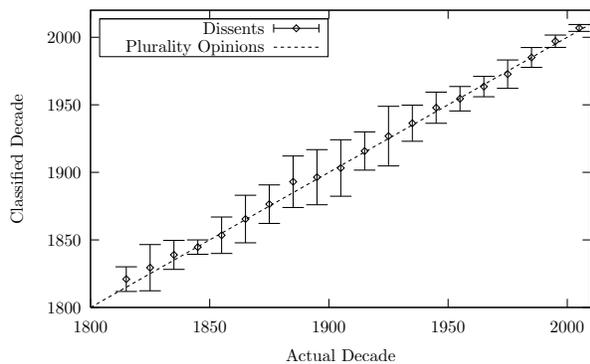


Figure 5: Classifications per decade of Supreme Court dissents, using an order-4 Markov model trained on the plurality opinions.

### Applied to Dissents

It is a legal aphorism that “yesterday’s dissent can become tomorrow’s Supreme Court majority opinion.”<sup>6</sup>

We attempted to confirm this adage by classifying the Supreme Court’s 5,488 decisions with dissents against our benchmark order-4 Markov chain trained on the Court’s plurality opinions. As Figure 5 indicates, we found no evidence that the Supreme Court’s dissents are “ahead of the curve,” or that yesterday’s

<sup>6</sup>See, e.g., Roger Goldman, Thurgood Marshall: Justice for All (1992) (quoted in Nadine Cohodas, A Pioneer In the Halls of Justice, Wash. Post, Jan. 24, 1993, at X1).

dissents tend to become tomorrow’s pluralities.

### Analysis

Clearly, the extent to which our analysis can speak to the flow of actual ideas in the judicial system is far from clear. In the future, we think it would be interesting to quantify the relationship between stylistic similarity and judicial accord — e.g., do justices who frequently concur with each other’s opinions also express their opinions similarly? Do judges on lower courts who write similarly to Supreme Court majorities have a lower probability of being reversed on appeal?

## 4 Differences Between Authors

Now we shift gears toward addressing the third and fourth questions posed in the introduction. How can we characterize the differences between contemporary authors’ use of the English language? A Markov chain might also do well in this arena, but in order to explore other kinds of features, we restricted ourselves to metrics that, this time, would hardly be related at all to an author’s word choices.

One of these “structural” features we examined was the tree-shape; a measure of how “straggly” or “bushy” the sentences in each document were. A straggly sentence is defined as one whose parse tree is deep as opposed to wide, while a bushy sentence is wide as opposed to deep. Thus, very straggly sentences are “recursive” in that they contain several levels of nested sub-sentences (complete phrases) within the sentence. In contrast, bushy sentences contain phrases made up of simple constructs chained together without many nested phrases.

We calculate the straggleness of a sentence to be the maximum depth of the tree (calculated by counting the maximum stack depth of the parse trees) divided by the number of phrases in the sentence (as determined by the number of lines in the parse tree). In addition, we took the average and standard deviation of the bushiness and straggleness of each sentence and used those as features as well.



two files of output: *filename.out* (which contains the input split into sentences and formatted as required by the parser) and *filename.parsed* (which contains the parse tree in pretty-printed format). The next phase, which actually generates the bushy vs. straggly statistics, is implemented by the script named “BvsSfeature.” This script runs the calculations to determine the max depth of the parse tree and its total width and calculates the straggliness from there.

## Results

We ran the tree-shape features through an “Svm-light” support vector machine in order to see how well it could distinguish between different authors.

Using only tree shape (straggliness percentage, and mean and standard deviation for the breadth and depth of the trees in the document), we found that the SVM was able to distinguish between authors with the following success levels:

Tree-shape features alone			
Success %	Friedman	Krugman	Safire
Dowd	76 %	51 %	62 %
Friedman		50 %	61 %
Krugman			25 %

For these six entries, we calculate the amount of useful information supplied by the discriminator. Even a discriminator that is mostly wrong (such as the Safire-Krugman discriminator, at 25 percent accuracy) can supply useful information! The amount of information supplied by a binary discriminator with an accuracy of  $x$  is:

$$-(x \lg x + (1 - x) \lg(1 - x))$$

Thus, the amount of information supplied by the tree-shape features alone toward the binary discrimination task is:

Tree-shape features alone			
Bits	Friedman	Krugman	Safire
Dowd	0.21	0.00	0.04
Friedman		0.00	0.03
Krugman			0.19

These figures are not so great. The most distinct pair of authors is Maureen Dowd and Thomas L. Friedman, and even with them the discriminator only supplies 0.21 bits of useful information.<sup>9</sup>

However, we next ran the SVM on a different set of features: the average word length, sentence length, and paragraph length, and the standard deviations of each within each article. The SVM was able to distinguish between pairs of authors and supply, based on its success rates, the following amounts of information:

Length features alone			
Bits	Friedman	Krugman	Safire
Dowd	0.33	0.02	0.00
Friedman		0.29	0.02
Krugman			0.19

Finally, we gave the SVM *both* sets of features. With the combined feature set, the amount of information supplied, based on the success rates, was:

Tree-shape and length features together			
Bits	Friedman	Krugman	Safire
Dowd	0.62	0.15	0.03
Friedman		0.23	0.60
Krugman			0.11

## Analysis

These results are perplexing, and we do not fully understand what is happening. Somehow, when supplied alone, one set of features (tree-shape) is only worth 0.03 bits of information to a support-vector machine trying to discriminate between Thomas L. Friedman and William Safire, and another set of features (word-, sentence-, and paragraph-lengths) is only worth 0.02 bits by itself. And yet, when these two sets are combined and given to the same SVM, the discriminator is now worth 0.60 bits!

<sup>9</sup>This is equivalent to saying that, armed with the tree-shape discriminator, an oracle would still have to send us 0.79 bits of information per article in order for us to be able to discriminate perfectly between Dowd and Friedman.

How can  $0.02 + 0.03$  equal  $0.60$ ? Are there hidden clues in the interaction between, e.g., word-length and sentence bushiness, that makes the separate features worth much more when supplied separately than when carelessly aggregated? Maybe. But also perplexing is that some of the discrimination tasks became *harder* when more features were supplied to the SVM, or that the tree-shape features and length features each produced a success rate lower than 50% on the Safire-Krugman discrimination task. These issues deserve further study.

## 5 Future Work

Our four initial questions have left us with many new issues to ponder.

- Is there any link between the stylistic or computer-inferable features of judicial opinions and the substantive ideas of the jurists?
- How well would we be able to identify the authorship of the Supreme Court's occasional unsigned opinions?
- In *The New York Times*, why did our two feature sets produce such perplexing results?
- How well could we discriminate among every op-ed columnist who works for the newspaper? What about every reporter?
- How well would a Markov chain of words perform against a Markov chain of letters?
- *The New York Times* employs different front-page editors for each day of the week — do they leave linguistic or stylistic fingerprints on the articles they edit strongly enough that a computer could guess an articles day of publication based on its contents?
- How do the features that can be used to date newspaper articles, or identify their authors, compare with the features one would want to use on the spoken word, such as in a radio transcript? What about blog articles?

## References

- [1] C. E. Chaski. Empirical evaluations of language-based author identification techniques. University of Birmingham Press 2001 1350–1771 Forensic Linguistics 8(1) 2001.
- [2] T. Grant and K. Baker. Identifying reliable, valid markers of authorship: a response to Chaski. Forensic Linguistics 2001, VOL 8; ISSU 1, pages 66–79.
- [3] D. Khmelev and W. Teahan. A repetition based measure for verification of text collections and for text categorization. Annual ACM Conference on Research and Development in Information Retrieval Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, 2003.
- [4] D. Khmelev and W. Tweedie. Using Markov Chains for Identification of Writer. Literary and Linguistic Computing, 2001, vol.16, no.4, pp.299-307.
- [5] O. de Vel. Mining E-mail Authorship. KDD-2000 Workshop on Text Mining, August 20, 2000.
- [6] C. Chaski. Who's At The Keyboard? : Authorship Attribution in Digital Evidence Investigations, International Journal of Digital Evidence, Spring 2005, Volume 4, Issue 1.
- [7] A. Anderson, M. Corney and G. Mohay. Multi-Topic E-mail Authorship Attribution Forensics. ACM Conference on Computer Security - Workshop on Data Mining for Security Applications, November 8, 2001, Philadelphia, PA.
- [8] S. Argamon, M. Saric, S. S. Stein. Style Mining of Electronic Messages for Multiple Authorship Discrimination: First Results. KDD-2003 Washington, DC.
- [9] M. Corney, O. de Vel, A. Anderson, G. Mohay. Gender-Preferential Text Mining of E-mail Discourse. Computer Security Applications Conference, 2002. Proceedings. 18th Annual.
- [10] D. Madigan, A. Genkin, D. D. Lewis, S. Argamon, D. Fradkin, L. Ye Author Identification on the Large Scale.