Who Wrote Ronald Reagan's Radio Addresses ?

Edoardo M. Airoldi
eairoldi@stat.cmu.edu

supervised by:

Stephen E. Fienberg
fienberg@stat.cmu.edu

Abstract

In his campaign for the U.S. presidency from 1975 to 1979, Ronald Reagan delivered over 1000 radio broadcasts. For over 600 of these we have direct evidence of Reagan's authorship of the text of the speeches, in the form of yellow pads, with material written "in his own hand". The aim of this study was to determine the authorship of 314 of the broadcasts for which no direct evidence is available.

Peter Hannaford had been Reagan's main aide in drafting texts for the radio addresses during the years 1976-79, whereas the situation was less clear in 1975, thus we learned both how to discriminate between the writing styles of Reagan and Hannaford, and we focused on stylistic differences between Reagan and the undistinguished pool of his collaborators to properly address the prediction problem for speeches delivered in different epochs. We explored a wide range of off-the-shelf classification methods as well as fully Bayesian Poisson and Negative-Binomial models for word counts. Simple majority voting reinforced the cross-validated accuracies of our predictions on speeches of known authorship, that settled beyond 90% in most cases. We produced separate sets of predictions using the most accurate classification methods and the fully Bayesian models, for the 314 speeches whose author is uncertain. All the predictions agree on 135 of the "unknown" speeches, whereas the fully Bayesian models agree on 289 of them. We further approximated log-odds of authorship as a measure of the strength of our predictions.

Among the crucial issues we had to deal with were the bold difference in the number of "known" speeches available for each author, and the phase of word selection. In the original dataset there were 679 speeches drafted by Reagan "in his own hand" and only 39 drafted by few close collaborators. With the help of Prof. Kiron Skinner and Prof. Annelise Anderson we looked into the Reagan files and we found 30 newspaper columns originally drafted by Peter Hannaford, but published with Reagan's signature. We coded them to obtain a set of 69 texts drafted by Reagan's collaborators, on which we based our inferences. The process of word selection was critical in order to understand "the secrets" of Reagan's writing style. Word counts very much fit the Negative-Binomial profile, and we relied on this fact to compute p-values for a certain statistic (T_1) in order to capture structural elements of differential writing style. We considered other criteria to find words with discriminatory power as Information Gain scores, computed for Multinomial and multivariate Bernoulli models, as well as a semantic decomposition of the speeches using Docu-Scope software. Following Frederick Mosteller and David Wallace in their analysis of "the Federalist Papers", we aimed for non-contextual words with possibly a few exceptions, that occurred with high, medium and low frequency. In making the decisions about contextuality a prior idea of Reagan's style based on the text of the Presidential debate Reagan vs. Carter, several notes, comments, and books about Ronald Reagan played a role. As an example consider the word *Carter*: our prior idea about Reagan's style suggested that Reagan would seldom talk about his opponent, Carter, his line of attack being more subtle. He would mostly address the *government* or *capitol hill people* and similar figures instead. Thus when the word *Carter* passed severe testing to make sure that its differential use by Reagan and Hannaford was too marked to be the outcome of pure chance, and it was likely to capture some element of Hannaford writing style, we did not discard it as contextual. Some have argued that Reagan's writing style might be better captured by some idioms he used. Thus we extended our analysis to the study of successive words to discover that, for example, idioms like *if we*, *in our*, *I'd like to* or *in America* identify Reagan's writing style beyond reasonable doubts.

We concluded that, in 1975, Ronald Reagan drafted 77 speeches, and his collaborators drafted 71, whereas, over the years 1976-1979, Reagan drafted 90 speeches and Hannaford drafted 74. The cross-validated accuracy of our best fully Bayesian model based on the Negative-Binomial distribution for word counts was above 90% in all cases. Further our inferences were not sensible to "reasonable" variations in the sets of constants underlying the prior distributions, which we bracketed with a small study on 90 high-frequency, function words. Our predictions for the speeches whose author is uncertain are accurate and reliable, and the agreement of several methods in predicting the author of the "unknown" speeches in most cases reinforced our confidence.

Acknowledgments

I wish to thank Prof. Annelise Anderson of Stanford University and Prof. Kiron K. Skinner of the Department of Social and Decision Sciences at Carnegie Mellon who presented me with the problem of the uncertain authorship of the radio addresses Ronald Reagan delivered between 1975 and 1979, and who provided me with the electronic texts of 1032 of such addresses. They contributed during every stage of this study with comments and suggestions, and provided detailed explanations on every aspect of the problem.

I also wish to thank Prof. Pantelis Vlachos for providing me with Docu-Scope software, and for his suggestions and comments in preparing an earlier version of this manuscript, and Prof. Robyn Dawes for comments and suggestions at an early stage of this project.

I am grateful to my advisor Prof. Stephen E. Fienberg for introducing me to this exciting project, for valuable comments and suggestions, for always pointing me in the right direction, and for the many hours he devoted to reviewing, and often re-drafting, this manuscript.

Contents

1 Introduction

We accessed a computerized database containing the text of 1032 radio addresses Ronald Reagan delivered over the years from 1975 to 1979. President Reagan's original draft written "in his own hand" was found for 679 of them, but not for the remaining 353, so that their authorship is uncertain. For a small number of these latter speeches a different author has been established by direct search through archives.

Of the speeches we labeled here as "written by Reagan in his own hand", 220 are contained in the book edited by Kiron Skinner, Annelise Anderson and Martin Anderson titled *Reagan, in His Own Hand* [29], and more short stories written by Reagan are contained in another shorter book by the same editors and titled *Stories in His Own Hand* [30]. While reading the books, we tried hard to keep a scientific eye open to suggestions and ideas on how to better tell apart Reagan's style from those of others who worked for him. We often found ourselves deeply involved. Pages would go by very quickly. One speech after the other. An unspoken urge to read more, compelled by his arguments. Eventually from time to time the scientific eye would pop open, complaining, and we would go back and *analyze* those pages that by then were way behind. But that is a major key point! Reagan was direct, informal and he talked to people. He made his words our words. He did not keep a distance.

In this report we describe first how we explored and polished the dataset, and scanned more texts to limit the bold differences in the numbers of texts by Reagan and by his collaborators which we used as a training set for our models; then we identified some features that distinguish Reagan's literary style from that of his collaborators beyond differences we could expect to find in several writings of a same author; and finally we made good use of these features, in order to conclude who is the most likely author for each of the speeches whose author is uncertain, along with an assessment of the confidence we have in these results.

Briefly our best "machine learning" classification methods agree in predicting the author of 135 out of 312 speeches Reagan delivered over the years 1975-79, and the more reliable fully Bayesian models agree in predicting 289 of them. The cross validated accuracies of our methods on about 750 "known" speeches range between 95% and 85%, with standard deviations of about 3% and 9% on texts drafted by Reagan and others respectively. A fully Bayesian approach based on the Negative-Binomial model best captured the variability in the data, and yielded quite stable predictions across 21 sets of underlying constants.

1.1 Historical Notes

Ronald Reagan was elected Governor of the State of California in 1966, and re-elected for a second term running through the end of 1975. He originally prepared to run for the presidency during the 1976 presidential campaign, and anticipated that his opponent would be Nixon's vice-president Spiro Agnew, but Agnew resigned in 1973 in a cloud of scandals. Gerald Ford became Nixon's next vice-president and actually replaced him when he resigned over the Watergate scandal. Reagan did not find support during the 1976 campaign that saw Gerald Ford and Jimmy Carter as main actors and made Carter president. Reagan won the next presidential campaign, however, against Carter, in 1980.

A series of events caused the decline of Jimmy Carter's public approval ratings from 70% to a low 28% during his term at office. Among them were scandals involving some of his staff members, his brother's alcohol problems, record levels of inflation, the energy crisis, and finally, on November 4th 1979, the seizure of the American embassy in Teheran by Iranian students, who captured 63 American citizens and held 50 of them as hostages for 442 days. As the American captivity

continued throughout 1980 — the election year — the television networks opened the evenings' news with a count of the number of days the hostages had been held. That was devastating. Critics presented Carter as being better prepared, but he lost the election to Reagan.

Reagan started promoting his ideas and his image in 1975 by means of radio addresses. Former Hollywood actor through 1964, Reagan had a very good relationship with the TV cameras. He presented himself as the leader that would give America back to the Americans, who did not need Government intrusions in many matters because they could handle situations themselves "the American way". Reagan proposed the display of strength as a better defense, and promised to re-built a professional, well-payed army of one million soldiers that would be ready to intervene anytime. He promised to make America the "shining city on the hill" that all the world had admired as superior in the "good old days". And he won the election.

1.2 Problem Definition

The main goal of this project was to determine with some degree of confidence which of the 353 speeches whose authorship is uncertain Reagan wrote. Eventually we aimed to study the effects of time on the writing style of President Reagan, in order to extend the analysis to other early writings, whose authorship is also uncertain. An electronic version of the Reagan speeches was provided by Prof. Annelise Anderson and Prof. Kiron Skinner.

In the original data we could attribute 679 speeches to Ronald Reagan, 39 to one of his collaborators (12 to Peter Hannaford, 26 to John McClaughry and 1 to Martin Anderson) and there were 314 speeches remaining whose author is uncertain[1]. Hannaford had been Reagan's main aide in the preparation of the handwritten drafts for the radio addresses over the years from 1976 to 1979, whereas the situation was unclear during 1975. Thus we decided to code several of Reagan's newspaper columns, which we know were actually drafted by Peter Hannaford.

Author	1975	1976	1977	1978	1979	Total
R. Reagan	60	195	52	219	153	679
P. Hannaford (radio)	1	5	2	4	0	12
P. Hannaford (news)	5	0	7	18	0	30
J. McClaughry	0	3	1	15	7	26
M. Anderson	0	0	0	0	1	1
Author uncertain	149	80	4	25	56	314
Total (known author)	66	203	62	256	161	748
Total (all)	215	283	66	281	217	1062

Table 1: Breakdown of the available texts by author and year.

We began by learning how to discriminate between the writing style of Reagan and other collaborators in 1975, and then we focused on stylistic differences between Reagan and Hannaford over the years 1976-79.

1.3 Notation

The main random quantity in our analysis is the number of times a word appeared in a text; we used X to indicate it. We used ℓ to denote the integer word-length of a document, whereas we

[1] Two speeches that contain essentially only quotations which were excluded from the analysis.

used ω to express the word-length of the same document in thousands[2] of words, so that text (ij) containing $\ell_{ij} = 745$ words would have $\omega_{ij} = 0.745$ thousands of words. The counts corrected for document length as discussed in section 3.2.1 are denoted by $Y := \frac{X}{\ell} \times const.$

As far as probability distributions are concerned we denoted them all with the letter p, as in $p(X \mid \theta)$, where θ denotes a generic vector of parameters, as in $\theta = (\mu, \delta, \xi, \eta)$ to be introduced later. Occasionally we indicated a distribution by its name, as in $Beta(X \mid \theta)$ or $Gamma(\xi \mid \beta)$, where the first argument indicates the random quantity, and the second argument indicates the parameter vector, which is random itself. In order to avoid confusions we referred to the components of the vector $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ as *underlying constants*, they are parameters of the distributions of the various *parameters* $(\mu, \delta, \xi, \eta) = \theta$, primary objects of our inference.

Indices appear in a few places: we used X_{nij} for word $n = 1, \ldots, N$ in document $j = 1, \ldots, J_i$ of author $i = 1, 2$; Reagan is always author $i = 1$, whereas we used $i = 2$ to denote Hannaford or the undistinguished group of Reagan's collaborators, depending on the occasion.

2 Literature Review

Augustus DeMorgan in his *Budget of Paradoxes* noticed that "some time somebody will institute a comparison among writers in regard to the average length of words used in composition, and that it may be found possible to identify the author of a book, a poem or a play in this way".

In a supplement of *Science* dated March 11 1887 [18], T.C. Mendenhall followed-up DeMorgan's idea and showed what he called the characteristics curves of composition. In the same fashion the spectroscope can be used to assess the presence of a certain element in a solid object, Mendenhall associated characteristic curves[3] to different authors, under the assumption that each writer makes use of a vocabulary peculiar to himself, the character of which does not change over his productive period. In the long run he expected that short words, long words, and words of medium length occur with definite relative frequencies. Mendenhall's assumptions were quite strong and soon his conjecture about one word-spectrum for each writer turned out to be wrong. But the fundamental idea that numerical summaries of texts could be used to some extent to extract relevant information about the authorship was born.

The early approaches to authorship attribution problems stemmed from the studies of G. Zipf (1932) on power laws [34], and of U. Yule (1944) on literary vocabulary [32]. Frederick Mosteller and David Wallace defined the problem of authorship attribution in its modern mathematical form in their book on the case of the 12 disputed "Federalist Papers" [24] [25], and George Miller ported into the linguistic community several mathematical solutions to the problem present in various communities in the 1950s [19].

Because of the work of Mosteller and Wallace (1964, 1984) [24] [25] a wide-spread strategy for authorship attribution problems is to consider hi-frequency function words (the *filler* words of the language, such as `a, an, by, to` and `that`, not related to the context). Burrows (1992) [5] and others suggest summarizing the information contained therein in terms of their principal components, or use some other method suitable for dimensionality reduction. Eventually natural clustering of the texts in the space spanned by these few highly descriptive features is investigated, and an attempt to classify the texts is made.

[2] For specific purposes we considered the word-length of a document in hundreds of words, rather than thousands, and we highlighted when that was the case. The same symbol ω was used.

[3] He also called these curves *word-spectra*.

3 Proposed Methodology

In order to make inferences for the speeches of uncertain authorship, we focused on the interpretability of our model along with a certain confidence on the validity of its underlying assumptions. Thus most of our effort was devoted to the study and fine tuning of the full Bayesian models in section 4.4.

3.1 Outline

In section 3.2 we briefly discuss the intuition behind the quantities and the statistics we used throughout the experiments in section 4. Mainly we corrected word counts that come from documents of different lengths, and we defined a statistic (T_1) that helped us distinguish stylistic patterns from variations in the use of words that could be the outcome of pure chance.

We present the exploratory data analysis in appendix 23 and discuss there how we parsed the text into what we considered *words* and *N-grams*, how we dealt with typos or words with different spelling, how we handled numbers, and other preliminary issues. In section 4.1 we describe the selection of good discriminating words and features. We made use of several heuristics and ended up with 9 pools of possibly good discriminating words, looking for differences in both the overall and average use of words in terms of their frequencies of occurrence, and made use of KL-distance related concepts in various ways. We also considered a semantic decomposition of the texts. In section 4.2 we describe features of the set of documents that we are going to analyze. We looked at natural clustering properties of the data, identified *difficult* documents, and assessed the potential for discrimination corresponding to each pool of words. In section 4.3 we summarize the accuracy of a wide spectrum of readily available off-the-shelf classifiers. Eventually we selected the most credible classifiers and combined them to increase their predictive power. In section 4.4 we present a fully Bayesian approach, using Poisson and Negative-Binomial models for the word counts; we checked assumptions, chose an appropriate parameterization and Prior distributions for the parameters of interest, and carried out the analysis for 21 sets of underlying constants. We also performed a brief empirical Bayes study. Eventually we made a final selection of the words to use for inference in terms of their *importance*. In section 4.5 we present our inferences regarding the speeches of uncertain authorship, discuss our findings, and perform sensitivity analysis.

3.2 Technical Issues

3.2.1 Word Length of Speeches

The different word length of the speeches may artificially increase or decrease the number of times a certain word appears, making the frequencies of occurrence not directly comparable across texts of different lengths[4]. Hence we needed to control for the word length. The solution we adopted was to measure the frequency of occurrence of words, relative to the length of the document, like so:

$$Y_{n\,i\,j} := \frac{X_{n\,i\,j}}{\ell_{i\,j}} = \frac{\#\text{ of times word } n \text{ occurs in text } j \text{ by author } i}{\#\text{ words in the text } j \text{ by author } i}. \tag{3.1}$$

It was convenient to re-scale the frequencies, a good choice seemed to be $(Y \times 1000)$, that was enough to keep the overall frequency of occurrence for the *function* words[5] greater than 1. Another

[4] All the speeches were planned for a radio program of constant duration. Much of the variability in the lengths of the addresses is due to the fact that we removed quotations in counting words, as they would otherwise mix with the author's stylistic patterns.

[5] These are the *filler* words of the language, such as `a, an, by, to` and `that`, not related to the context.

possibility was $(Y \times 500)$ since the speeches vary in length between 56 and 784 words, with a median length of 500 words.

3.2.2 Thresholds for Word Selection

We wanted to be able to distinguish random variations in the differential use of words by two authors from traits of their literary styles. In section 4.1 we use the statistic

$$T_1 := \frac{(\sum_j X_{i_1 j} - \sum_j X_{i_2 j})^2}{\sum_j X_{i_1 j} + \sum_j X_{i_2 j}} = \frac{(J_1 \bar{X}_{i_1} - J_2 \bar{X}_{i_2})^2}{J_1 \bar{X}_{i_1} + J_2 \bar{X}_{i_2}} \tag{3.2}$$

to select the words with good discriminating power. T_1 takes high values for both those words that author i_1 uses more often than author i_2 and vice-versa. In order to make good use of T_1 we needed a measurement scale to assess how much differential use[6] of a word is enough to select it as a *marker word*. In other words we wanted to filter all the excursions we would observe in T_1 if the same author wrote a series of documents, due to natural randomness in his writing style, from more severe excursions that may hardly be explained by mere chance. We concluded that, apart from necessary multiple testing considerations that we save for later, $t_1 > 3.85$ is a reasonably safe value to detect stylistic patterns that affect the use of a word.

We assumed Poisson and Negative-Binomial models for $p(X \mid \theta)$, and obtained the thresholds for T_1, via Monte-Carlo simulations and asymptotic calculations[7], for type 1 error values $\alpha = 0.05$ and $\alpha = 0.15$, and for values of the components of $\theta = (\mu, \delta)^T$ in the relevant ranges.

Model for $p(X\|\theta)$	Simulation			Asymptotic (2^{nd})	
	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$	$\alpha = 0.05$	$\alpha = 0.15$
Poisson	$t_1 > 3.80$	$t_1 > 2.69$	$t_1 > 2.07$	$t_1 > 4.65$	$t_1 > 2.93$
Negative-Binomial					
$\delta = 0.001$	$t_1 > 3.83$	$t_1 > 2.70$	$t_1 > 2.08$		
$\delta = 0.25$	$t_1 > 4.74$	$t_1 > 3.37$	$t_1 > 2.59$		
$\delta = 0.50$	$t_1 > 5.70$	$t_1 > 4.00$	$t_1 > 3.10$		
$\delta = 0.75$	$t_1 > 6.58$	$t_1 > 4.66$	$t_1 > 3.62$		
$\delta = 1$	$t_1 > 11.2$	$t_1 > 5.35$	$t_1 > 6.22$		
$\delta = 5$	$t_1 > 21.5$	$t_1 > 15.4$	$t_1 > 14.4$		

Table 2: Thresholds for T_1 in equation 3.2 at $\alpha = 0.05$, $\alpha = 0.10$ and $\alpha = 0.15$, obtained by simulation and asymptotic calculations, corresponding to Poisson and Negative-Binomial models for $p(X \mid \theta)$. The columns with the simulated values quote the median values of the 95%, 80% and 85% percentiles over 1000 simulations for each value of the parameters in the relevant ranges. The Poisson mean $\mu \in [1.28, 111]$ for a single speech, and the non-Poissonness parameter $\delta \in [0.001, 0.75]$, plus two more values $\delta = 1$ and $\delta = 5$ to check what happens for values of δ in the tails.

To get a feel for the thresholds, we can assume that X_{i_1}, $X_{i_2} \sim Poisson(\omega\, \mu)$, independent and with the same rate μ, and that all the documents we observe have the same length $\ell = 1000$. We fixed $T_1 > 3.80$ and simulated counts to obtain a p-value of ≈ 0.049545 for a minimum $\mu = \frac{1}{784} \cdot 1000 \approx 1.28$ for one of J_i speeches, and a p-value of ≈ 0.049969 for a maximum $\mu_{(the)} = \frac{39}{350} \cdot 1000 \approx 111$ for one of J_i speeches. $\mu \in [1.28, 111]$ is the range of rates we observe in the data[8].

[6] In terms of number of occurrences of a word in a set of texts, $\sum_j X_{n\,i\,j}$.

[7] We used the bi-variate delta method on an expansion of $T_1(\bar{X}_{i_1}, \bar{X}_{i_2})$ up to the second order.

[8] Note that in the simulation we considered $\mu \times \min(J_{i_1}, J_{i_2})$, consistent with formula 3.3 above.

3.2.3 Corrections for Unbalanced Sample Sizes

The fact that we have 679 speeches for Reagan and only 65 for other authors (pooling all their texts together) was a source of concern. In general the main impact of different sample sizes is on the standard deviations of various quantities we may want to estimate separately for the two authors, hence on the significance of related tests. We briefly discuss below the two points in the analysis where the marked difference in the number of texts available raised some technical concerns.

In section 4.1 we will be looking for words that the two authors use with a different average frequency $\hat{m}_i = \frac{1}{J_i}\sum_j \frac{X_{i\,j}}{\ell_{i\,j}}$. Formally we test $H_0 : m_{i_1} = m_{i_2}$ vs. $H_0 : m_{i_1} \neq m_{i_2}$ where i_1 =Reagan and i_2 =Hannaford. In this case $J_{i_1} = 679$ and $J_{i_2} = 38$, and at least for a high-frequency[9] word n we can believe the averages $\hat{m}$ are normally distributed via central limit theorem. This is a fairly standard testing problem even as the common variance is not known as long as J_{i_1}/J_{i_2} is close to $\frac{1}{2}$. As the ratio J_{i_1}/J_{i_2} moves from $\frac{1}{2}$ a more complex solution is needed, since the Normal test does not guarantee the desired significance level α in a non-controllable way. When the true variances are not necessarily equal, a sensible solution that is free of concerns about the different sample sizes was given by Welch (1938) [31], which we used in our analysis.

In using the statistic T_1 defined by equation 3.2 above, we needed to correct the word counts in the two sums $\sum_{j=1}^{J_{i_1}} X_{i_1\,j}$, and $\sum_{j=1}^{J_{i_2}} X_{i_2\,j}$, since they involved different numbers of documents in the two samples, being $J_{i_1} = 679$ and $J_{i_2} = 38$[10] in the example above. The correction is simple, we will consider the counts as if observed on a set of $\min(J_{i_1},\, J_{i_2})$ documents, which in the example we are discussing means considering the word counts we would observe in 38 documents. Using the numbers in the example we correct equation 3.2 like so:

$$T_1 := \frac{(\frac{\sum_{j=1}^{679} X_{i_1\,j}}{679} \times 38 - \sum_{j=1}^{38} X_{i_2\,j})^2}{\frac{\sum_{j=1}^{679} X_{i_1\,j}}{679} \times 38 + \sum_{j=1}^{38} X_{i_2\,j}}. \tag{3.3}$$

Note that we used the minimum number of documents to avoid implicit inference; i.e. in the formula above we *summarize* the word counts in 679 documents into word counts over 38 summary documents, instead of *projecting* the word counts in 38 documents into word counts over say 100 representative documents, thus using word counts that we did not actually see. This was an important point since if we adjusted T_1 to represent the counts we would observe in 100 documents, thus making some inference on the counts in the 38 texts, we would obtain a different and bigger set of marker words, implicitly reducing the α level of our selection, per word.

3.2.4 Estimating the Prediction Error

Training and testing a classifier on the same data yields a biased estimate of the prediction error, usually called the *apparent* error. Cross-validation reduces the bias by basically making sure that training and test sets do not share data points. It is still possible that, depending on the slope of the learning curve at the chosen size of the training set, cross-validation yields an estimate of the prediction error biased upward. Cross-validation is good enough for model selection, though, under the assumption that bias does not change the relative performance of the methods we are comparing — see Efron (1983) [9], Davison and Hinkley (1999) [7], and Hastie et al. (2001) [11]

[9] In these cases the corresponding Poisson or Negative-Binomial distributions are almost symmetric to start with, and the number of independent component in the average is high enough to believe the central limit theorem.

[10] In section 4.1 we adopted a two stage filtering procedure, and J_{i_1} and J_{i_2} will take different values from the ones mentioned here for explanatory purposes. See also appendix 23 for more details.

for more details. Model selection was a primary focus in section 4.3 and we used a hybrid cross-validation scheme, namely:

1. Repeat $b = 1, ..., B$ times:
   1.1. randomly permute the data points,
   1.2. train on the first 80% of the data,
   1.3. compute error err_b on the remaining 20% of the data.
2. Estimate the prediction error and its standard deviation using

$$\widehat{err} = \frac{1}{B}\sum_{b=1}^{B} err_b, \quad \text{and} \quad \widehat{sd}_{err} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(err_b - \widehat{err}_b)^2}.$$

4 Experiments

4.1 Word Selection

We considered several lists of candidate words, and we tried to distinguish random variations in the differential use of words by two authors from traits of their literary styles. In other words we wanted to filter the variation in the use of words we would observe in a series of documents by a same author, due to natural randomness in his writing style, from more severe variations that may hardly be explained by mere chance.

In order to measure how variable the use of a word in two texts had to be in order to mark the word as good discriminator, we made use of the statistic T_1 and simple probability arguments explained in detail in section 3.2.2, and we performed a series of two-sample Kolmogorov-Smirnov tests or Welch approximate t-tests, along with False Discovery Rate correction to control the overall α level in multiple tests. We freely discarded the words that were possibly related to the context of a radio address, by simply looking at them one by one. Some prior information about Reagan's style entered at this stage, since whenever a word was on the borderline between being contextual and not we made an arbitrary decision. e.g. `carter` was retained whereas `republicans` and `congress` were discarded.

This way of proceeding gave us reason to believe that the words we found had some real discriminative power related to literary styles that affect the frequency of use of words. Thus we would be surprised to be the result of mere chance.

4.1.1 Pools of Words

Below we provide a short summary of the pools of words we considered, along with brief comments. Appendix 27 contains the lists of marker words we retained.

1. **High-Frequency words:** We selected 267 function words among the most frequent 3000 words in Reagan and Hannaford vocabularies, and we filtered them using both the Kolmogorov-Smirnov statistic and the Welch approximate t-statistic along with the False Discovery Rate (FDR) correction, proposed by Benjamini and Hochberg (1995) [2], for the many tests involved[11]. We came up with a list of 55 markers.

[11] The False discovery rate correction proposed by Benjamini and Hochberg is derived under the hypothesis of independent tests. Independence of words statistically holds, as we show in section 4.4.1; however, for some words, we do not expect independence to hold in general, and our results may be taken as practical approximation.

2. **SMART stop words:** We considered the words in the list of 523 *stop words* using by the SMART system by Salton and Buckley, and we tested for differences in distributions and means correcting for multiple testing with the False Discovery Rate, to come up with a list of 62 markers.

3. **DOCU-SCOPE semantic features:** We retrieved a total of 21 stylistic features using Docu-Scope — a text tagging and visualization software courtesy of Professor Pantelis Vlachos (Statistics Department) and Professor David S. Kaufer (English Department) [6] — tested for differences in distributions and means correcting for multiple testing with False Discovery Rate, and we explored their relevance further using a jackknife procedure on a linear discriminant function to come up with 6 weakly discriminating features. We provide further details in appendix 27.

4. **Two-stage selection on all 4-grams:** We considered all the unique 4-grams, $69,000$ in the Hannaford dictionary and $729,000$ in the Reagan dictionary. The first-stage consisted of filtering all the 4-grams in the dictionaries using two groups of documents in sequence, texts 1975-1977 and texts 1978-1979, to mitigate selection effects since we did not correct for multiple testing in this first stage. The second stage consisted in testing for differences in distributions and means of all the words that passed through the first stage, and correcting for the many tests with the False Discovery Rate, to come up with a list of 50 markers.

   We also adopted a different scheme; we filtered all the 4-grams on all the texts in a first stage, we removed the contextual words, we approximated the distributions of T_1 in order to compute p-values, simulating the counts for each word over 1 million pairs of texts, and eventually we performed the tests using $\alpha = 0.01$ and the False Discovery Rate. See the lists of words in section 22.

5. **Common words:** We tried working with the set of common words {`and in the of or`} used by Mosteller and Tukey (1968) [23] to demonstrate their double jackknife leave-one-out procedure.

6-9. **Information gain:** Pool no.6 contains words with high Information Gain (IG) with respect to a Bernoulli model, and Reagan and Hannaford as authors; pool no.7 contains words with high IG with respect to a Multinomial model, and Reagan and Hannaford as authors; in pool no.8 the IG is computed with respect to a Bernoulli model, and Reagan, Hannaford and McClaughry as authors; and finally in pool no.9 we use a Multinomial model to compute IG and Reagan and others as authors.

The most promising pools are no.4 and no.6-9, because they provide several markers for both Reagan and Hannaford. In pools no. 1 and no. 2 the correction for multiple testing might have been particularly severe because of the high number of words at the onset (267 and 469 respectively) and we did not correct for multiple testing during the first-stage filtering that produced pool no. 4.

Different search criteria yield different and yet useful sets of markers. Here we focused on markers with a different distribution of frequencies of occurrence, and with different average frequencies of occurrence because some of the classification techniques the we are going to use below base their statistical strategies on comparing the texts of different authors using differences in the means and in the distributional properties of the frequencies of occurrence of words.

4.2 Exploring Feature Spaces

In this section we investigate some characteristics of the texts of the "known" speeches, in terms of the relevant dimensions (words) we identified in the preceding section.

4.2.1 Principal Components

We applied the standard principal components method suggested by Burrows (1992) [5] to describe the texts by various authors in terms of few principal components [15] [3]. Usually dimensionality reduction is applied to a list of 75 hi-frequency words, and we used all the 267 hi-frequency candidate words of pool no.1. The goal was to assess the value of such an approach in discriminating authors.

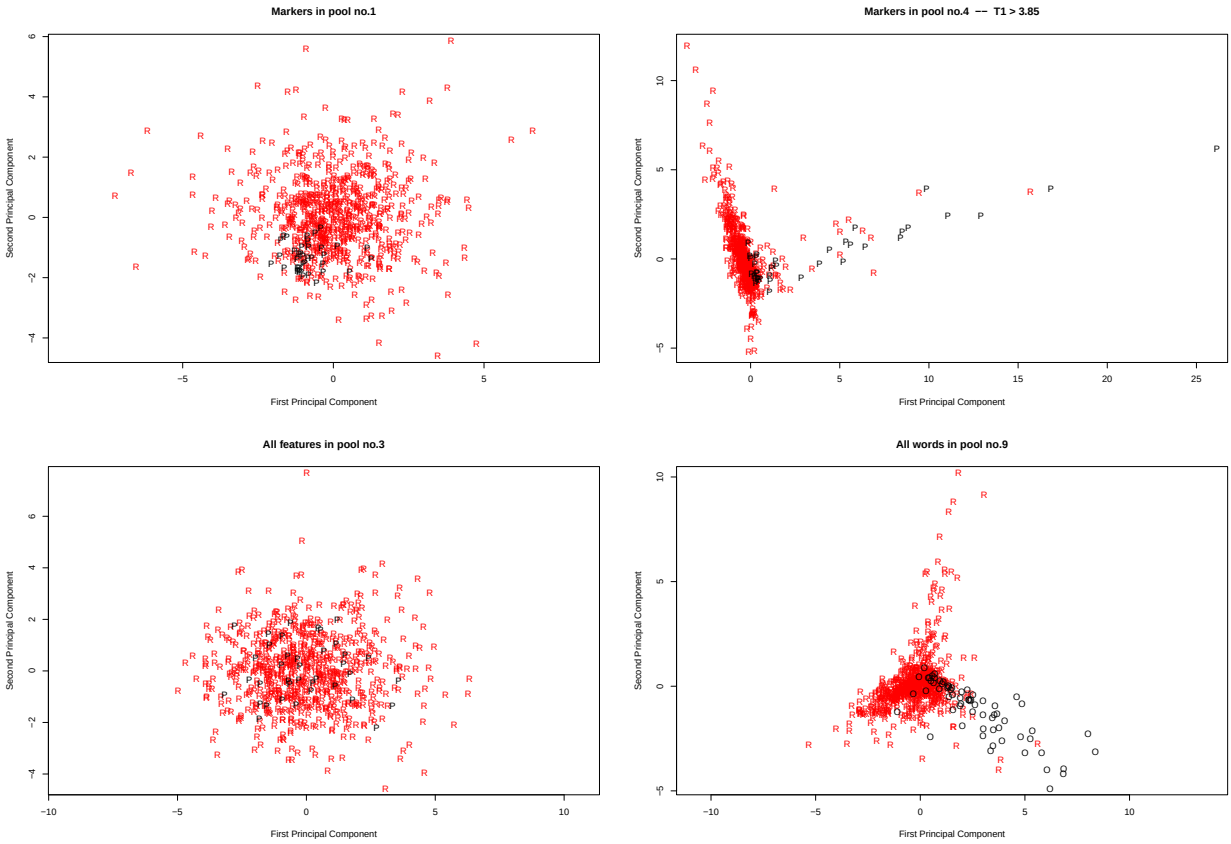

Figure 1: The information contained in the hi-frequency words (top-left panel) is compared to the information contained in the 41 markers we found using the statistic T_1 (top-right panel), to the information contained in the 18 second-level semantic features (bottom-left) and in the 30 words selected using Information Gain (bottom-right), as captured by the first two principal components. The benefits of an explicit search for good discriminating words to the purpose of authorship attribution is evident.

Burrows suggested that the information contained in the principal components possibly captures features of literary style, and if, usually a big if, the principal components could be interpreted they would explain how different authors write. The main objection to this method is that principal components do not help us decide whether the differential information they capture is simply the result of chance or not. We do not exclude the possibility that describing the space of hi-frequency words in terms of its principal components may be a useful tool in assessing authorship, as testified

by the several recent studies that employ this technique, but its place should be among the tools for a first stage exploratory data analysis. A final decision about the author of a text should rely on methods able to capture structural elements of style.

We conclude that, in this problem, more refined criteria for the search of marker words, such as information gain and the statistic T_1, along with few medium and low frequency non-contextual words yielded feature spaces that better separated the texts in terms of their respective principal components. Further we trust that the information these feature spaces contain is related to differences in the usage of words that were not likely to be due to pure chance!

4.2.2 Unsupervised Clustering

We performed some unsupervised classification [12] to see how close the documents authored by Reagan, Hannaford, McClaughry and Anderson position themselves in the various spaces of features, according to various definitions of distance and similarity.

The raw, scaled or transformed data did not cluster naturally. We tried logarithm and N-root transformations on the raw and scaled data, in combination with various distance measures (Euclidean, city-block, Mahalanobis, cosine, correlation and Minkowski) and various aggregation rules during the clustering (single-linkage, average-linkage, complete-linkage, median, Ward and centroid[12]) without exciting results. A second series of attempts, more promising at the onset, was aimed to cluster the first few, say 3, principal components that emerged from the experiments above, using all the combinations of transformations, distance measures and linkage methods we used on the row data. The principal components for the words in pools no.4 and no.9 separate the texts into 4 reasonably homogeneous groups, but the groups themselves are not close to one another.

We conclude that lack of strong natural clustering is one of the difficulties of the problem. Purely geometrical methods did not work well; see for example the poor performance of nearest neighbor classifiers in section 4.3.

4.3 Off-the-Shelf Classifiers

In this section we explore some simple techniques to perform text classification in order to get a feel for which classification technique works best on the set of data that we had. We briefly discuss assumptions underlying each technique, and conclude with an application to speeches whose authorship is unknown in section 4.3.1.

We provide possibly optimistic estimates of the mis-classification error, as follows. Given data $\{(X_1, Y_1), ..., (X_N, Y_N)\}$ and a classifier $H(\cdot)$

1 repeat $b = 1, ..., B$ times:

1.1 permute the data $\{(X_1, Y_1)^*, ..., (X_N, Y_N)^*\}_b$,

1.2 train H on 80% of the * sample,

1.3 compute $error_b$ on the remaining 20% of the * sample.

2.1 compute total mis-classification error and its variability on the B estimates ($error_b$) available.

[12] If the resulting cluster tree was not monotonic, a phenomenon that occurred when the distance from the union of two clusters to a third cluster is less than the distance from either individual cluster to that third cluster we would discard the results.

Whenever it is not specified otherwise, we set $B = 1000$. Specifically we implemented a balanced version of the algorithm above, where in each * sample we set aside 20% of the Reagan and 20% of the Hannaford speeches for testing, as opposed to 20% of the speeches in general.

Naïve Bayes

The naive Bayes classifiers we considered model the probability of a document j from author i as both multivariate Bernoulli and multinomial. We assumed that words occur independently across texts, and that their occurrences are independent from one another, so that the probability $p(d_{ij}|\theta_i)$ of document (i, j) can be expressed in terms of the probabilities of observed counts X_{nij} for word n like so:

$$\prod_{n=1}^{N} p(X_{nij}|\theta_i) = \begin{cases} \prod_{n=1}^{N} Bernoulli_n\left(X_{nij}|\theta_i = p_{ni}\right) & X_{nij} \in \{0, 1\} \\ \prod_{n=1}^{N} Multinomial_n\left(X_{nij}|\theta_i = (p_{1\,ni}, \ldots, p_{K\,ni})\right) & X_{nij} \in \{0, ..., K\} \end{cases}$$

Our experiments included variations in the prior for the unseen words (Dirichlet, M-estimate, Good-Turing, Witten-Bell), in the prior about the author of each text, and in the further pruning of the dictionary on the training texts (by information gain on word counts, or on presence/absence of a word in a document). The highest accuracy corresponded to a multinomial model with a Dirichlet prior, using the single words we obtained using Information Gain and the statistics T_1.

	Predicted Author	
True Author	Hannaford	Reagan
Hannaford (42)	40 (2.9)	2
Reagan (679)	60	619 (16.6)

	Predicted Author	
True Author	Others	Reagan
Others (69)	62 (5.1)	7
Reagan (679)	45	634 (13.9)

Table 3: Confusion matrices for Naïve Bayes classifier with multinomial likelihood and Dirichlet prior for unseen words over 1000 experiments, standard deviations in terms of number of texts in brackets. Left: Reagan vs. Hannaford using single words obtained by the Information Gain. Right: Reagan vs. other collaborators using single words obtained with the statistics T_1. The accuracies are about 90% or more in all cases, and the standard deviations about 2% and 7%, on Reagan and the alternative author(s) respectively.

Useful references for naïve Bayes can be found on the web site of the Text Learning Group at Carnegie Mellon [22] directed by Tom Mitchell, in Mitchell (1997) [21], and in the paper by Zhai and Lafferty (2001) [33] who compare smoothing methods for language models. Smoothing is done to adjust the maximum likelihood estimators of certain quantities to correct the inaccuracies due to the sparseness of the data.

Majority voting

This classification method involves the non-parametric estimation of the distributions $p(Y_{ni}|i)$ of the adjusted frequency of usage of word n by author i. We obtained a density estimate for each word and author (n, i); a single-word classifier then assigned document j to the author i that was more likely to use word n with the observed frequency Y_{nij} as in $\hat{i} = \arg\max \hat{p}(Y_{ni}|i)$, and a multi-word classifier aggregated single ones via simple majority rule. This method did not achieve performances above 80% on both authors.

Maximum likelihood

As above, we estimated the distributions $p(Y_{ni}|i)$ in a non-parametric fashion to end up with single-word classifiers, and eventually we obtained the multi-word classifier by maximizing the empirical likelihood, product of the estimated $\hat{p}$. Our experiments included variations in the estimation of the joint density of the frequencies of occurrence of words (histogram and kernel, both with optimal bandwidth via cross-validation and/or normal reference rule), and in the selection of words that voted on each text (all, or only the words that occurred). This method did not achieve performances above 80% on both authors.

Unit-weight models

Dawes and Corrigan (1976) [8] proposed a unit-weight model to make decisions about the author of the twelve disputed Federalist Papers. In that study this simple method gave the same predictions as the more elaborate "double Jackknife" procedure proposed by Mosteller and Tukey (1968) [23].

Predictions of unit-weight models were obtained by first rescaling all the variables (word counts in our case) to force an equal unit variance, then computing the correlations between each of the variables and the author variable to assign positive or negative unit-weights accordingly, and finally adding up the products of word counts and corresponding unit-weights to obtain the final score for each of the texts. This method performed poorly on the known Reagan speeches. Its cross-validated accuracy was at $\approx 40\%$ for each author.

LDA and QDA

Our experiments included three versions of LDA; Fisher (or canonical) LDA with threshold via cross-validation, which assumes no statistical model and finds the best separating direction using mean and variance arguments, and two variants of the population version of LDA, which on the contrary assume multivariate normality of the data. Specifically `Pop.v1` predicted the authorship using a threshold that corresponds to an even prior ($\pi_{RR} = 0.5$, $\pi_{PH} = 0.5$), whereas `Pop.v2` further corrected the estimates of the grand mean, the within-group and the between-group variance-covariance matrices by weighting the observation in group j by $\frac{\pi_j\, n}{n_j}$. Note that the MLE estimates are not weighted in this case, but Ripley (1996) [28] suggests that the weighted estimates are more accurate[13]. We considered variance-stabilizing transformation, mainly $\log(X + 0.75)$ because of the many zero counts, we removed variables that were collinear from a pool, and attempted variable selection using AIC and t-tests.

	Predicted Author	
True Author	Hannaford	Reagan
Hannaford (42)	38 (4.2)	4
Reagan (679)	81	598 (20.4)

	Predicted Author	
True Author	Hannaford	Reagan
Hannaford (42)	29 (7.1)	9
Reagan (679)	170	509 (34.0)

Table 4: Confusion matrices for sample LDA accuracies. Left: accuracies are at about 88% and 90%, with standard deviations at about 3% and 10% on Reagan and Hannaford texts respectively. Right: accuracies are at about 75% and 70%, with standard deviations at about 5% and 17% on Reagan and Hannaford texts respectively.

[13] This is particularly effective when the sample is not totally random, but there is a fixed number of examples that comes from rare groups, as it is the case in our permuted samples.

Fisher LDA does not assume Normality, but was less then 65% accurate on Hannaford texts in all cases, whereas the population versions of LDA assume multivariate Normality on data that are at most a mixture of a Normal and a point mass at zero in each dimension, but yield accuracies as high as 88% on Reagan texts and 90% on Hannaford texts with standard deviations at about 3% and 10% respectively. Docu-Scope semantic features could be safely regarded as Normal, because of a robust construction, but their accuracy in the population versions of LDA was at most 75% and 70%, with a standard deviation of about 5% and 17% on Reagan and Hannaford texts respectively.

It was not possible to use QDA in any pool but with Docu-Scope features because of the way we picked marker words; in fact some good Reagan markers occurred often in Reagan speeches, but not at all in Hannaford speeches, and vice-versa, so that as soon as we attempted to estimate different variance covariance matrices for the two authors separately, we would obtain rank deficient estimates because of the lack of variability for some pairs (word, author).

Fisher Linear Discriminant Analysis was originally introduced by R.A. Fisher (1936) [10] and extended by C.R. Rao (1948) [27]. For more references see Ripley (1996) [28].

Logistic regression

Good performance, more accurate on Reagan texts. Our experiments included plain logistic regression, variance-stabilizing transformations such as $\log(X + 0.5)$, and a weighted version of logistic regression where we artificially augment Hannaford texts via Bootstrap to compute weights to assign to each text drafted by Hannaford. Notably we achieve an accuracies as high as 97.4% on Reagan texts and 81.2% on Hannaford texts, with standard deviations at about 2% and 15% using the words we found using the statistic T_1 and Welch tests, and plain logistic regression with transformation $\log(X + 0.5)$.

	Predicted Author	
True Author	Hannaford	Reagan
Hannaford (42)	35 (5.0)	7
Reagan (679)	16	663 (10.4)

Table 5: Confusion matrix for logistic regression and transformation $\log(X + 0.5)$ over 1000 experiments, standard deviations in terms of number of texts in brackets. The accuracies are at about 98% and 83%, with standard deviations at about 2% and 12% on Reagan and Hannaford texts respectively.

Support Vector Machines

Joachims (1998) [13] first introduced the SVM methods of Vapnik into disputed authorship problems. Our experiments included linear, polynomial, Gaussian, sigmoid and Fisher kernels, on both raw and transformed data. The accuracy on Hannaford texts was constantly low, and in any case below 61% with a standard deviation about 18%.

κ-Nearest Neighbor

This method performed poorly on Hannaford speeches. The reason is that the documents authored by Hannaford are few, and not close enough in space to one another. See section 4.2 for more details. See Hastie et al. (2001) [11] and Kleinberg (1997) [16] for references.

CART and Random Forests

Classification trees performed poorly on Hannaford speeches, with an accuracy constantly below 50%. The standard deviation of the predictions was very high at about 25%, since for several training/test set combinations the accuracy got as high as 85%. Classification trees entail a partition of the space of features (word counts in our case) with cuts parallel to the coordinate axes; there were no evident reasons that hinted trees would under-perform in the classification task, with respect to the other methods. See Hastie et al. (2001) [11] and Mitchell (1997) [21] for references.

Random forests improved on CART, as expected, but did not get as accurate as we needed. Its accuracies settled at about 95% and 75% on Reagan and Hannaford texts respectively, with standard deviations of about 5% and 15% respectively. It always performed poorly on Docu-Scope semantic features. See Breiman (2001) [4] for references.

Combining Classifiers

We attempted combining accurate classifiers by simple majority voting. Few experiments included linear discriminant analysis, and logistic regression over different pools of words. In the cases where the classifiers had different cross-validated accuracies, voting would yield a sort of average accuracy. In several cases where the cross-validated accuracy of all the classifiers involved was similar, and about 90% or more on both texts of Reagan and the alternative author, the cross-validated accuracy of the combined classifier would settle at the level of the best classifier in the pool, or would slightly increase of about 0.5 to 1 percentage point. In section 4.5 we show that combining accurate predictions of the fully Bayesian models also increased the cross-validated accuracy of about 1 percentage point, in several cases. We conclude that combining high quality predictions mostly reinforces, and some times improves their accuracy.

4.3.1 A Glance at the Unknown Speeches

Here we present a first attempt to classify the speeches whose author is uncertain, using combinations of the classifiers tested so far.

We considered all the 312 "unknown" speeches, and we produced separate predictions for the speeches Reagan delivered in 1975, and over the years 1976-79, since we believe that Hannaford had been Reagan's main aide in the preparation of the radio addresses from 1976 on, whereas in 1975 several other collaborators may have drafted the speeches. The most accurate classifiers in our experiments were naïve Bayes (multinomial), and logistic regression, whose cross-validated accuracies were above 85% on both classes, with standard deviations that vary between 1% and 5% on Reagan speeches, and between 10% and 15% on Hannaford speeches[14]. The table below summarizes the agreement of the two best classifiers in classifying the 312 unknown speeches.

	Logistic Regression	
Naïve Bayes	Hannaford	Reagan
Hannaford	68	6
Reagan	80	158

Table 6: Agreement of logistic regression on words in pool no.4 ($T_1 > 3.85$) and naïve Bayes classifier with uniform prior of authorship and Dirichlet smoothing for unseen words.

[14]The difference is due to uneven sample sizes, 679 training speeches for Reagan and 42 only for Hannaford.

Making use of the relative accuracies in order to break the ties in table 6, we were able to assign an author to all of the unknown speeches. Considering the 164 speeches between 1976-79, starting with a prior that assigned equal probability for each speech to Reagan and to Hannaford we would assign 65 speeches (39.6%) to Hannaford and 99 (60.4%) to Reagan. Staring with a 95%, 5% prior we would assign 50 speeches (30.5%) to Hannaford and 114 speeches (69.5%) to Reagan. Considering all the 312 speeches, again starting with an even prior we would assign 93 speeches (31.7%) to Hannaford and 213 (68.3%) to Reagan. Staring with a 95%, 5% prior we would assign 76 speeches (24.4%) to Hannaford and 236 speeches (75.6%) to Reagan.

4.4 Full Bayesian Modeling

The fully Bayesian analysis involved several steps. Briefly we (1) checked the assumption that words occur independently across the text in section 4.4.1, (2) estimated parameters of Poisson and Negative-Binomial models for a handful of words in section 4.4.3 in order to get an idea of their distributions, and (3) chose a meaningful parameterization and sensible priors for the new parameters in section 4.4.4. Next we bracketed the constants underlying the prior distributions and estimated "likely" values for these constants in section 4.4.6, and eventually we carried out posterior computations for the parameters. We further selected words by some measure of their importance in section 4.4.7, and finally, in section 4.5, we assessed the accuracy of the fully Bayesian classifiers using cross-validation, compared the accuracy of the predictions obtained using the posterior modes to those obtained using the posterior means, explored ad-hoc models for the bi-grams, and produced final predictions for the authors of the "unknown" speeches. The quantities we used to make decisions about the authors of the texts have an interpretation in terms of odds of authorship, which added to valid and reasonable assumptions, and to an extremely good fit of the Negative-Binomial model, makes the results of our analysis relatively trustworthy.

4.4.1 Independence of Words

Here we discuss the assumption that words are independent of one another. Briefly, since we considered function words like *thus*, *that*, *till*, not related to the context it seemed reasonable to consider such words as independent at the onset since they tend to be separated by multiple other words in the text; we then considered pairs of function words and explored their independence. In most of the cases independence held. For the most dependent pairs like *if we* and *that it* we investigated the assumption more closely.

we	*if*			
	0	1	2	3+
0	55 (37)	43 (39)	30 (27)	72 (97)
1	36 (38)	47 (39)	28 (28)	92 (98)
2	18 (25)	21 (26)	14 (18)	81 (65)
3+	18 (27)	21 (28)	20 (19)	83 (69)

Table 7: The table quotes the observed number of speeches that contain {0,1,2,3+} counts for the function words *if* and *we* out of the 679 speeches drafted by Reagan; the expected number of speeches is quoted in brackets. The p-value corresponding to a Pearson's χ^2 test for independence (≈ 0.0002) did not provide strong evidence against independence when a correction for multiple testing was used.

Consider, for example, the pair of function words *if we*: independence held on the speeches drafted by Reagan's collaborators and on the "unknown" speeches, whereas on the speeches drafted

by Reagan it did not, mainly as a result of the high number of speeches. As we considered subsamples of 250 speeches, though, independence held on average. Further word independence was one of the assumptions underlying both the seminal work of Mosteller and Wallace, who considered function words too, and characterize the basic model now widely used in the Computer Science literature[15]. Finally our aim was to produce reliable predictions, and considerations about the stability of our predictions as much as about their cross-validated accuracies supported our modeling choices. We concluded that independence of words of one another was a reasonable assumption.

Stationarity

Here we want to assess whether the independence of occurrence of words across text (a Bernoulli process) is a reasonable assumption to make, in order to justify models for words independent of positions in the text. We considered the hi-frequency words in pool no.1, the semantic features in pool no.3, the markers we found with the statistics T_1 in pool no.4, and the words we identified as having high information gain in discriminating Reagan as opposed to Hannaford in pools no.6 and no.7 (Bernoulli and multinomial statistical models respectively), and as opposed to other authors in general in pool no.9.

We concluded that the independence assumption generally held. It held for most of the words across Hannaford's and other authors' texts, whereas at a first glance it seemed not to hold across Reagan's texts. The main cause of non-independence was large sample sizes: we considered blocks of 4 adjacent sets of 200 words each, and Reagan texts provided about 1600 blocks. Such a big sample size was enough to make significant even relatively small differences between observed and expected counts. Sampling subsets of about 250 blocks[16] yielded independence for most of the words across Reagan's texts as well. The table above summarizes the situation.

Pool of words	Hannaford (111 blocks)	Reagan (1612 blocks)	Reagan (250 blocks)
50 highest frequency words	46 (46)	34 (48)	48 (48)
54 markers in pool no.1	30 (30)	45 (54)	48 (51)
21 features in pool no.3	13 (13)	18 (18)	16 (16)
49 markers in pool no.4	36 (36)	42 (49)	39 (41)
27 markers in pool no.6	24 (27)	19 (27)	24 (24)
31 markers in pool no.7	31 (31)	23 (25)	21 (21)
27 markers in pool no.9	26 (26)	18 (25)	24 (26)

Table 8: Results of independence tables for various pools of words. In brackets we quote the actual number of words compared using the corresponding χ^2-values, in fact a χ^2-value cannot be computed for extremely high or low frequency words. The rightmost column titled Reagan (250 blocks) contains the average number of independent words over 100 samples of 250 blocks each. Results include False Discovery Rate correction.

4.4.2 First Parameterization

Here we introduce the functional forms which we are going to use for the Poisson and the Negative-Binomial word counts models for $p(X_{nij} \,|\, \theta)$ — see Johnson et al. (1992) [14] for details. Recall that X_{nij} stands for the counts for word n in text j of author i. Dropping the indices on the right

[15] See the book by Tom Mitchell (1997), pgg. 180-.

[16] Mosteller and Wallace used 247 blocks of 200 words each in their analysis of the "Federalist Papers".

hand sides to improve readability, for the Poisson density, we write

$$\begin{array}{rcl} f_p(x_{nij}|\omega_{ij},\,\mu_{ni}) & = & \frac{e^{-\omega\mu}\,(\omega\mu)^x}{x!}, \quad x=0,1,2,\ldots\,. \\ & & \text{s.t.} \quad \omega>0, \quad \mu>0. \end{array} \tag{4.1}$$

For the Negative-Binomial density, we write

$$\begin{array}{rcl} f_{nb}(x_{nij}|\omega_{ij}\mu_{ni},\,\kappa_{ni},\,\omega_{ij}\delta_{ni}) & = & \frac{\Gamma(x+\kappa)}{x!\,\Gamma(\kappa)}\,(\omega\delta)^x\,(1+\omega\delta)^{-(x+\kappa)}, \quad x=0,1,2,\ldots \\ & & \text{s.t.} \quad \omega>0, \quad \mu>0, \quad \kappa>0, \quad \delta>0, \quad \kappa\delta=\mu. \end{array} \tag{4.2}$$

We index the parameters consistently, so that μ_{ni} is the Poisson rate for word n and author i, that is the number of such words we would expect to see in any thousand[17] consecutive words of text, δ_{ni} is the non-Poissonness rate, $\kappa_{ni} := \frac{\mu_{ni}}{\delta_{ni}}$ is a redundant parameter that will be useful for some derivations, and ω_{ij} is the word-length of a document expressed in thousands of words.

4.4.3 Parameter Estimation

Here we discuss the estimation of parameters of Poisson and Negative-Binomial models for word counts.

Counts		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Ahead	(obs)	6	21	36	57	92	98	95	57	57	54	34	20	21	11	11	4	1	0
Ahead	(Neg-Bin)	3	15	35	61	82	93	93	83	67	51	36	24	15	9	5	3	2	1
Ahead	(Poisson)	1	8	24	51	81	102	108	98	78	55	35	20	11	5	2	1	0	
they	(obs)	112	127	113	101	59	52	39	24	17	14	4	9	3	1	0	1	0	0
they	(Neg-Bin)	103	128	118	95	72	52	37	25	17	11	7	5	3	2	1	1	1	0
they	(Poisson)	32	98	149	152	116	71	36	16	6	2	1	0						
our	(obs)	146	171	124	81	55	42	20	13	9	3	8	3	1	0	0	1	0	1
our	(Neg-Bin)	167	152	116	82	56	37	25	16	10	7	4	3	2	1	1	0		
our	(Poisson)	67	155	180	139	81	37	15	5	1									
you	(obs)	255	170	103	51	36	26	10	9	4	3	3	0	3	0	1	0	2	0
you	(Neg-Bin)	298	135	80	52	35	24	16	11	8	6	4	3	2	1	1	1	1	0
you	(Poisson)	120	208	180	104	45	16	5	1	0									
us	(obs)	322	195	92	37	18	9	3	1	2	0								
us	(Neg-Bin)	325	189	92	42	18	8	3	1	1	0								
us	(Poisson)	261	250	119	38	9	2	0											
its	(obs)	399	177	73	15	11	2	2	0										
its	(Neg-Bin)	397	181	67	23	8	2	1	0										
its	(Poisson)	358	229	73	16	2	0												

Table 9: Expected versus observed counts for Poisson and Negative-Binomial models, for various words. The Negative-Binomial model is able to capture the variability in the occurrence of words.

For the parameters of the Poisson model we computed maximum likelihood estimates. For the Negative-Binomial model the situation is more messy. The main results about estimators for both parameters of a Negative-Binomial distribution are not useful in our case, as the texts all have different lengths (ω_{ij}). Following Mosteller and Wallace (1964, 1984) we used method of moment estimators that make use of weights to deal with different word length of texts, and their choice of weights is "optimal" at the Poisson limit[18]. The estimators we used are:

$$\left\{ \begin{array}{rcl} \hat{\mu} & = & m, \\ \hat{\delta} & = & d = \max\left\{0, \frac{v-m}{mr}\right\}, \end{array} \right. \tag{4.3}$$

[17] Or hundred when we consider Docu-Scope features.

[18] The Negative-Binomial distribution equals the Poisson distribution in the limit, as $\delta \to 0$ (for fixed μ).

where

$$m = \frac{\sum x_j}{\sum \omega_j}, \qquad v = \frac{1}{J-1} \sum \omega_j \left(\frac{x_j}{\omega_j} - m \right)^2, \quad \text{and} \quad r = \frac{1}{J-1} \left(\sum \omega_j - \frac{\sum \omega_j^2}{\sum \omega_j} \right). \tag{4.4}$$

Overall the Negative-Binomial model captures the variability in the data better than the Poisson model. Docu-Scope features also fit the Negative-Binomial profile. Some examples are shown in table 9 above. We explored further the goodness of fit of the two distributions, and we summarize the results[19] in the table 10 below.

Pool of words	Poisson Model			Negative-Binomial Model		
	Hannaford (38 texts)	Reagan (679 texts)	Reagan (75 texts)	Hannaford (38 texts)	Reagan (679 texts)	Reagan (75 texts)
50 highest frequency words	12 (50)	0 (50)	3 (50)	31 (50)	0 (50)	49 (50)
54 markers in pool no.1	4 (15)	0 (17)	0 (17)	14 (15)	2 (17)	13 (17)
21 features in pool no.3	3 (21)	0 (21)	1 (21)	21 (21)	0 (21)	20 (21)
49 markers in pool no.4	1 (12)	0 (14)	0 (14)	12 (12)	2 (14)	14 (14)
27 markers in pool no.6	1 (11)	0 (11)	0 (11)	10 (11)	1 (11)	11 (11)
31 markers in pool no.7	1 (5)	0 (3)	0 (3)	5 (5)	0 (3)	1 (3)
27 markers in pool no.9	0 (7)	0 (8)	0 (8)	7 (7)	2 (8)	8 (8)

Table 10: Goodness of fit of Poisson and Negative-Binomial models for various pools of words. In brackets we quote the actual number of words compared using the corresponding p-values obtained using a two-sample Kolmogorov Smirnov test. The rightmost columns of each distribution titled Reagan (75 texts) contain the results of our tests over 100 samples of 75 texts each. We freely discarded low frequency words — less than 8 per ten-thousand words.

In the parameterization in terms of (μ, δ, ω, κ) we used for the Negative-Binomial model, δ seemed stable across words and authors (see table 10 below); mostly $\delta_i \in [0, 0.75]$ with some heavy tails. Such heavy tails in the non-Poissonness parameter δ are mostly due to personal pronouns, but we included them in the analysis nonetheless since they make good discriminators. In order to use a simple prior for δ_i we used a variance stabilizing transformation to reduce the heavy tails as in $\zeta = \log(1 + a\delta)$, with $a = 1$ for length measured in thousands of words.

Assume that $\delta_1 = \delta_2$ is satisfactory for most function words, but not for low frequency markers. Even though differential non-Poissonness is potentially discriminating our actual motivation for the choice of modeling possibly distinct δ_i was not to upset the analysis.

4.4.4 Reparameterization and Prior Distributions

Here we introduce a different parameterization and we motivate it in terms of the simple priors we are able to find for the new parameters. From $\theta = (\mu_1, \mu_2, \delta_1, \delta_2)$ we switch to $\theta = (\sigma, \tau, \xi, \eta)$.

In order to separate the average rate of use of a word n from a comparison between the rates themselves for Reagan and the alternative author, we introduced the parameters (σ_n, τ_n), where

$$\sigma_n = \mu_{n,1} + \mu_{n,2}, \quad \text{and} \quad \tau_n = \frac{\mu_{n,1}}{\mu_{n,1} + \mu_{n,2}}.$$

[19] Two notes: (1) we need to check the appropriateness of the models whenever the sample is under-dispersed; and (2) we are bound to observe ties since we compare two discrete distributions, whereas the Kolmogorov-Smirnov test is for continuous distributions. A simple practical solution to avoid ties is to add a small perturbation to word counts, say $\epsilon \sim Unif[-10^{-5}, +10^{-5}]$, before performing the test. A different solution is to fit $Gamma(\theta)$ models to the two sets of word counts to come up with estimates $\hat{\theta}_1$ and $\hat{\theta}_2$, and then perform a parametric test to check whether $H_0 : \theta_1 = \theta_2$ or not.

Parameter Estimator Word	μ_1 m_1	μ_2 m_2	δ_1 d_1	δ_2 d_2	$1/\kappa_1$ d_1/m_1	$1/\kappa_2$ d_2/m_2	ξ	η
the	61.4922	66.6606	1.5532	1.1914	0.0253	0.0179	1.7219	0.5444
of	32.7620	30.8648	0.5328	0.1942	0.0163	0.0063	0.6046	0.7064
to	26.5245	28.9015	0.3265	0.3404	0.0123	0.0118	0.5756	0.4909
a	24.4899	21.8701	0.9810	0.0000	0.0401	0.0000	0.6836	1.0000
and	23.9112	21.9158	0.1921	0.0000	0.0080	0.0000	0.1757	1.0000
in	21.1984	20.5004	0.6303	0.1524	0.0297	0.0074	0.6306	0.7751
that	14.0524	14.4736	0.8491	0.6967	0.0604	0.0481	1.1434	0.5376
it	10.2632	12.1450	1.4490	0.7924	0.1412	0.0652	1.4792	0.6055
for	9.9210	8.6750	0.7128	0.3713	0.0718	0.0428	0.8539	0.6302
on	7.4135	8.5837	0.7701	0.6384	0.1039	0.0744	1.0648	0.5363
as	5.8736	6.2095	0.6809	1.3562	0.1159	0.2184	1.3764	0.3773
be	5.8487	5.9355	1.0907	0.4654	0.1865	0.0784	1.1196	0.6587
with	5.8114	6.1638	0.4295	0.3151	0.0739	0.0511	0.6312	0.5661
are	5.6900	7.3053	1.5820	1.0802	0.2780	0.1479	1.6810	0.5643
by	5.6372	5.0224	0.6655	0.7754	0.1181	0.1544	1.0842	0.4705
have	5.4256	5.2050	0.5712	0.8065	0.1053	0.1549	1.0432	0.4331
our	4.9061	1.6893	3.2639	4.9695	0.6653	2.9417	3.2368	0.4480
has	4.7505	3.6983	0.8945	2.3803	0.1883	0.6436	1.8569	0.3441
this	4.3430	4.7028	0.2368	1.1914	0.0545	0.2533	0.9971	0.2132
an	4.1221	3.2874	0.3718	1.0801	0.0902	0.3286	1.0486	0.3015
no	2.4546	1.4154	0.4377	0.4193	0.1783	0.2962	0.7132	0.5090
which	2.8684	2.1916	0.3397	0.9107	0.1184	0.4156	0.9399	0.3111
its	1.3502	3.1504	0.8434	3.9614	0.6246	1.2574	2.2133	0.2763
carter	0.0996	2.3286	0.5011	2.7037	5.0339	1.1611	1.7155	0.2368
may	0.5040	1.9176	0.3361	0.0000	0.6668	0.0000	0.2897	1.0000
if	3.1701	4.0636	0.7664	0.0000	0.2418	0.0000	0.5689	1.0000

Table 11: Method-of-Moments estimates of Negative-Binomial parameters for 26 words.

Recall that we defined $\zeta_i = \log(1 + a\delta_i)$ to reduce the heavy tails of the non-Poissonness parameters δ_i. We eventually transformed ζ_1, ζ_2 into (ξ_n, η_n), like so

$$\xi_n = \zeta_{n,1} + \zeta_{n,2}, \quad \text{and} \quad \eta_n = \frac{\zeta_{n,1}}{\zeta_{n,1} + \zeta_{n,2}},$$

where ξ_n and η_n measure combined and differential non-Poissonness respectively.

We used 90 words to get a rough idea of what sensible prior distributions for $(\sigma, \tau, \xi, \eta)$ should look like. The 90 words we considered were ranged from high, to medium and low frequency, and some of them were weakly discriminating. The words we used in the experiments to study possible priors were then set aside, and never used again.

The panels in figure 2 show that both τ and η appear approximately symmetric about 0.5, which is the value for no differential use of words, and more analysis yielded a set of Beta distributions that brackets reasonable priors for both. Due to the small variability of ξ we assumed that the prior on ξ, for which a gamma distribution turned out to be a reasonable choice, is independent of

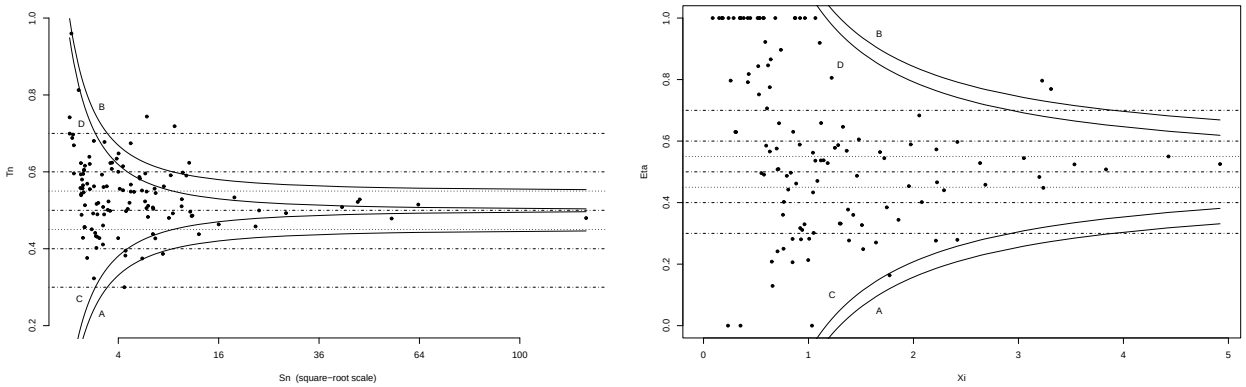

Figure 2: **Left:** Sample estimates of the parameters (σ_n, τ_n) for 90 function words, with high and low frequency. Curves C and D show two-standard-error bands for t_n when $\tau_n = 0.5$. Curve A shows a two-standard-error band below $\tau_n = 0.45$. Curve B shows a two-standard-error band above $\tau_n = 0.55$. **Right:** Sample estimates of the parameters (ξ_n, η_n) for 90 function words, with high and low frequency. Curves C and D show two-standard-error bands for t_n when $\eta_n = 0.5$. Curve A shows a two-standard-error band below $\eta_n = 0.45$. Curve B shows a two-standard-error band above $\eta_n = 0.55$.

the prior from η. It was not safe to make the same assumption about (σ, τ) because of the wide range of σ, and we assumed that the variability of τ decreases as σ increases, as the left panel of figure 2 suggests. We then assumed a constant prior for σ.

4.4.5 Full Model Specifications

For each word n we introduced the parameters:

$$\begin{array}{rclcrcl} \sigma_n & = & \mu_{n,1} + \mu_{n,2}, & & \tau_n & = & \frac{\mu_{n,1}}{\mu_{n,1}+\mu_{n,2}} \\ \zeta_i & = & \log(1 + a\delta_i), & & i & = & 1, 2 \\ \xi_n & = & \zeta_{n,1} + \zeta_{n,2}, & & \eta_n & = & \frac{\zeta_{n,1}}{\zeta_{n,1}+\zeta_{n,2}}. \end{array} \tag{4.5}$$

For each set of underlying constants $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ we assume that for all words in the pool from which the N words were selected:

(A1) the vectors $(\sigma_n, \tau_n, \xi_n, \eta_n)$ are independent across words,

(A2) ξ_n, η_n and the pair (σ_n, τ_n) are independent from each other for each word n,

(A3) σ_n has a χ^2 density that can be approximated by a constant,

(A4) conditional on $\tau_n|\sigma_n$ has symmetric Beta density with parameter $(\beta_1 + \beta_2\sigma)$,

(A5) η_n has symmetric Beta density with parameter (β_3),

(A6) ξ_n has Gamma density with parameters $(\beta_5, \frac{\beta_4}{\beta_5})$.

We mainly used 21 sets of underlying constants β in the posterior computations for all markers in all pools, for both Poisson and Negative-Binomial models. We used these different β to perform sensitivity analysis, and explored more sets in several occasions.

The figure below displays the inference scheme we followed. In a first stage we used the word counts in the "known" speeches along with the priors to compute values for the parameters $(\sigma_n, \tau_n, \xi_n, \eta_n)$ that maximized their joint posterior distribution, for each word in a final selection of about 170 marker words. In a second stage we used these modal parameter values to fine tune the selection of words, and make inference on the "unknown" speeches.

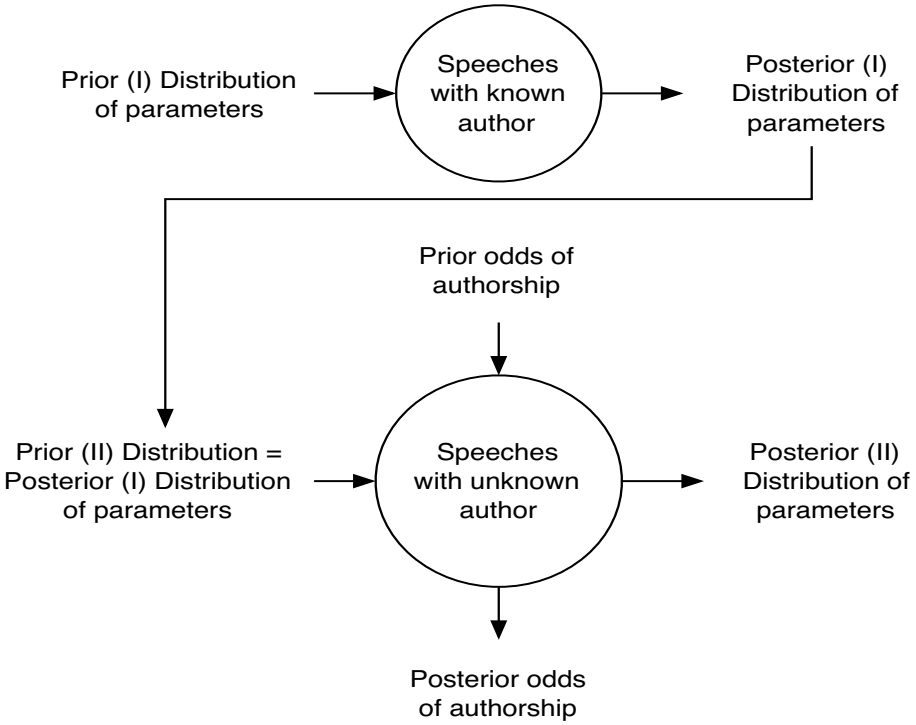

Figure 3: Our two-stage inference scheme at a glance: we learned the characteristics of the different literary styles using the "known" speeches, and eventually we used this knowledge to compute the odds of authorship for the "unknown" speeches.

4.4.6 Empirical Bayes for Non-Believers

We computed 21 sets of posterior modal values for the parameters, corresponding to 21 sets of underlying constants, for different pools of words. Each of these sets of constants β contained information about the combined and differential "average" use of a word, and the combined and differential "variability" in the use of such word, with respect to two authors. The prior distributions we devised would entail the beliefs that the two authors write in the same way, but some sets would entail beliefs "more extreme" than others[20], and the more extreme the beliefs about equal writing styles, the more difficult the classification task using the odds.

We began our posterior computations with more than 40 sets of underlying constants, containing both the 21 "more extreme" sets of constants used by Mosteller and Wallace, and 20 more we chose. Eventually we dropped the sets that were too favorable to differential writing style, to end up with the sets of constants below, that we used in most of the analysis.

[20] In terms the strength of the evidence supporting differential writing styles needed to change such prior beliefs of a certain amount.

Set no.	β_1	β_2	β_3	β_4	β_5
1	5	1	6.0	1.25	2.0
2	2	1	6.0	1.25	2.0
3	10	1	6.0	1.25	2.0
4	20	1	6.0	1.25	2.0
5	5	5	6.0	1.25	2.0
6	2	10	6.0	1.25	2.0
7	20	10	6.0	1.25	2.0
8	5	1	6.0	0.91	2.0
9	5	1	6.0	1.54	2.0
10	5	1	18.0	1.25	2.0
11	5	1	1.5	1.25	2.0
12	10	0	12.0	1.25	2.0
13	10	0	6.0	1.25	2.0
14	10	0	12.0	0.83	1.2
15	10	0	6.0	0.83	1.2
16	15	0	12.0	0.83	1.2
17	10	0	18.0	0.83	1.2
18	10	0	12.0	0.72	1.2
19	10	0	30.0	0.83	1.2
20	5	0	12.0	0.83	1.2
21	5	5	6.0	0.83	1.2

Table 12: The 21 sets of underlying constants used in the posterior computations.

We then looked for the set of constant β "most likely" in the sense of Mosteller and Wallace (1984). Preliminary calculations indicate the values $\beta_1 = 7$, $\beta_2 = .2$, and for the Negative-Binomial model the additional values $\beta_3 = 7$, $\beta_4 = .9$ and $\beta_5 = 1.1$. We note, however, that the results we obtained with the Negative-Binomial model, in terms of accuracy, are very stable across all sets of underlying constants. Further simple majority voting among the predictions of 21 sets yields an accuracy of about 92% on Reagan's texts and 99% on the texts drafted by his collaborators. These considerations are the best guarantee of reliable predictions for the unknown speeches.

4.4.7 Final Word Selection

Here we discuss how we computed the importance of each of the 168 words that made it up to here, and why we chose to discard some of them. The goal was to avoid performing an expensive computation for words without *consistently* strong discriminatory power. A different road to final word selection, which favored combination of words with high accuracy and disregarded concerns about computational costs, is explored in appendix 22. In any case the predictions on the "known" texts identify the correct author in more than 90% of the cases.

In order to attribute a text to Reagan or to the alternative author using word n we compute the odds of authorship as:

$$\begin{aligned} Odds(i_1, i_2|X_n) &= \frac{p(X_{n\,i_1}\,|\,\hat{\theta}_{i_1})}{p(X_{n\,i_2}\,|\,\hat{\theta}_{i_2})} \times \frac{\pi_{i_1}}{\pi_{i_2}} \\ \text{(final odds)} &= \text{(likelihood ratio)} \times \text{(initial odds)} \end{aligned} \tag{4.6}$$

or equivalently in terms of log odds as:

$$\text{(final log odds)} = \text{(log likelihood ratio)} + \text{(initial log odds)} \tag{4.7}$$

A reasonable measure of importance of a word is the difference between the expected log likelihood ratio given Reagan's authorship and that given, say, Hannaford authorship, as in $\omega\,(\mu_{i_1} - \mu_{i_2})\,\log(\mu_{i_1}/\mu_{i_2})$. As an example we computed the posterior rates according to the Poisson model for three words, and their importance.

Word	Reagan rate	Hannaford rate	Importance ($\omega = 2$)
also	.97	1.09	.02
an	4.10	3.17	.49
because	1.28	.79	.47

Table 13: Rates per thousand words for *also, an* and *because*, and their importance for discrimination.

Our results were obtained discarding words with importance less than 0.1 for words whose combined rate was less than 1.5, and discarding words with importance less than 0.48 for words whose combined rate was more than 1.5, plus we selected in each pool only those words whose importance would exceed the threshold for all the 21 sets of underlying constants. More results were obtained by selecting all the 1-Gram markers, and discarding the 2-Grams, 3-Grams and 4-Grams freely.

4.5 The Unknown Speeches

Here we present some results on the cross-validated accuracy of Poisson and Negative-Binomial models, that we obtained using subsets of the final selection of words in appendix 22. These results constitute a lower bound for the accuracies we can get using combinations of such words; for example, the number of combinations of 30 words out of 117 is the same order of magnitude as 10^{27}. Eventually we predicted the author for the "unknown" speeches and we looked at degree of agreement of these predictions with the predictions of multinomial naïve Bayes and logistic regression, best among the off-the-shelf classifiers we studied in section 4.3.

Poisson Predictions

Using the words we obtained with the statistic T_1, we predicted the author for the speeches whose author is known for several sets of underlying constants[21]. The accuracy of the Poisson model very much depended on the set of underlying constants that we used. Using two separate pools of words for the speeches given in 1975 and over the years 1976-79 improved the accuracies we would obtain using the same pool of about 1%. The accuracy dropped as we used simple majority voting among predictions obtained with different sets of constants.

In order to check the accuracy of the modal approximations of the likelihood ratios, we simulated the joint posterior distribution of the parameters $(\tau, \sigma, \eta, \xi)$ for each word in the pool, using a Gibbs sampler with Metropolis steps, and computed the values of the likelihood ratios at the posterior means for several sets of underlying constants. The apparent accuracies do not change in the Poisson case. The process of cross validation for the approximations of the likelihood ratios at the posterior means is computationally expensive.

[21] In the Poisson model β_1 and β_2 are the only relevant constants so that we count 10 different sets.

True Author	Set of underlying constants (β_1, β_2) used										Voting
	no.1	no.2	no.3	no.4	no.5	no.6	no.7	no.12	no.16	no.20	(all sets)
Reagan (136)	117.0	120.8	112.8	109.6	112.5	111.5	104.4	114.1	111.6	118.6	112.5
Std. Dev.	3.6	2.9	4.1	4.5	3.8	4.3	5.6	3.9	4.2	3.3	3.9
Others (14)	12.2	12.4	11.9	11.5	12.5	12.5	11.7	11.8	11.6	12.0	12.5
Std. Dev.	1.6	1.5	1.6	1.8	1.4	1.5	1.7	1.6	1.7	1.8	1.4

Table 14: Average accuracies and standard deviations on the "known" speeches in 1975: we quoted the cross-validated number of texts correctly predicted using the Poisson model, for the words obtained by T_1. Refer to table 12 for the specific values of the underlying constants.

True Author	Set of underlying constants (β_1, β_2) used										Voting
	no.1	no.2	no.3	no.4	no.5	no.6	no.7	no.12	no.16	no.20	(all sets)
R. Reagan (136)	117.8	122.0	114.2	107.4	112.7	112.5	100.9	115.0	111.8	119.3	115.0
Std. Dev.	2.79	2.90	2.75	4.41	2.87	2.16	5.94	2.49	3.66	2.49	2.49
P. Hannaford (8)	6.9	6.9	6.5	6.6	7.0	7.1	6.8	6.4	6.5	6.6	6.4
Std. Dev.	0.70	0.54	0.92	0.80	0.63	0.54	0.60	0.92	0.92	0.66	0.92

Table 15: Average accuracies and standard deviations on the "known" speeches in 1976-79: we quoted the cross-validated number of texts correctly predicted using the Poisson model, for the words obtained by Information Gain. Refer to table 12 for the specific values of the underlying constants.

Negative-Binomial Predictions

The Negative-Binomial model is more accurate than the Poisson model. The apparent accuracies on both the speeches delivered in 1975 and the speeches delivered over the years 1976-79 jump beyond 92%, and up to 99%. Cross validated accuracies were computed for the two sets of underlying constants no.2 and no.20, to discover that the average cross-validated accuracies remained as high as 91% and 95%, on Reagan and the alternative author respectively; in most of the experiments the accuracies were at 100% on both authors, and dropped down to 50% on a few experiments, hence increasing the variability. More experiments are needed here.

True Author	Type	Set of underlying constants (β_1, β_2, β_3, β_4, β_5) used	
		no.2	no.20
Reagan (136)	apparent	125	130
	cross-val.	121.1	129.3
Std. Dev.		7	9.4
Others (14)	apparent	14	13
	cross-val.	12.5	12.8
Std. Dev.		1.9	1.3

Table 16: Average accuracies and standard deviations on the "known" speeches in 1976-79: we quoted the apparent and cross-validated number of texts correctly predicted using the Negative-Binomial model. Refer to table 12 for the specific values of the underlying constants.

The posterior modes could not be computed for some combinations of words and underlying constants, because of rank deficiencies of the Hessian matrix. More desirable approximations of the likelihood ratios at the posterior means, obtained with a Gibbs sampler with Metropolis steps, were always available and stable across the sets of underlying constants we tested. Their corresponding

apparent accuracies increased as well by 1 to 4 percentage points. We show the apparent accuracies for the modal approximations in tables 17 and 18.

True Author	Set of underlying constants (β_1, β_2, β_3, β_4, β_5) used										Voting
	no.5	no.6	no.7	no.8	no.9	no.10	no.11	no.12	no.13	no.14	(all sets)
Reagan (679)	624	631	576	633	630	630	648	613	602	621	628
Others (69)	68	68	68	66	67	66	56	67	67	66	68

Table 17: Average accuracies and standard deviations on the "known" speeches in 1975: we quoted the apparent number of texts correctly predicted obtained evaluating Negative-Binomial likelihood ratios at the posterior modes, for the words obtained with the statistics T_1. Refer to table 12 for the specific values of the underlying constants.

True Author	Set of underlying constants (β_1, β_2, β_3, β_4, β_5) used										Voting
	no.1	no.2	no.3	no.6	no.7	no.8	no.11	no.12	no.17	no.20	(all sets)
R. Reagan (679)	640	663*	618	633	569	648*	649	624	630*	651*	633
P. Hannaford (42)	39	39*	39	41	41	40*	40	40	40*	40*	40

Table 18: Average accuracies and standard deviations on the "known" speeches in 1976-79: we quoted the apparent number of texts correctly predicted obtained evaluating Negative-Binomial likelihood ratios at the posterior modes, for the words obtained with the statistics T_1. An asterisk indicates that the likelihood ratios were evaluated at the posterior means. Refer to table 12 for the specific values of the underlying constants.

We conclude by providing the apparent accuracies we obtained by using the Negative-Binomial model for counts of the semantic features of Docu-Scope software, which is stable about 76% on Reagan speeches, and about 90% on Hannaford speeches.

True Author	Set of underlying constants (β_1, β_2, β_3, β_4, β_5) used										Voting
	no.1	no.2	no.3	no.6	no.7	no.8	no.9	no.12	no.20	no.21	(all sets)
R. Reagan (679)	511	513	507	503	502	510	512	515	517	508	508
P. Hannaford (42)	34	34	34	34	34	34	34	35	34	35	35

Table 19: Average accuracies and standard deviations on the "known" speeches in 1976-79: we quoted the apparent number of texts correctly predicted obtained evaluating Negative-Binomial likelihood ratios at the posterior modes, for the semantic features of Docu-Scope software. Refer to table 12 for the specific values of the underlying constants.

4.5.1 Multiple Predictions

We conclude by presenting two three-way tables that display the degree of agreement of our classifiers on the speeches whose authorship is uncertain. In the first table, we combined the predictions obtained from the fully Bayesian models, using several sets of underlying constants, with the predictions of the multinomial naïve Bayes and the logistic regression classifiers. In the second table, we combined the predictions of the fully Bayesian Negative-Binomial model for the sets of underlying constants no.2 and no.20, which on the "known" speeches had were most accurate, with the aggregate predictions of all the sets of underlying constants, composed by majority voting. In table 20, we notice that the logistic regression classifier assigns more speeches to Ronald Reagan than the

	Poisson Bayesian model			
	Hannaford		Reagan	
Multinomial naïve Bayes	Logistic Regression		Logistic Regression	
	Hannaford	Reagan	Hannaford	Reagan
Hannaford	53	31	26	8
Reagan	8	10	21	154

Table 20: Agreement of un-weighted logistic regression, naïve Bayes classifier with uniform prior for authorship and Dirichlet smoothing for un-seen words, and Poisson Bayesian model using sets of underlying constants β no.1 to no.4, and no.8. The predictions were obtained using the words obtained by the statistic T_1 for the all the "unknown" speeches.

	Neg-Bin: Simple Majority Voting (21 sets β)			
	Hannaford		Reagan	
Neg-Bin: $\beta^{(2)}$	Neg-Bin: $\beta^{(20)}$		Neg-Bin: $\beta^{(20)}$	
	Hannaford	Reagan	Hannaford	Reagan
Hannaford	172	12	0	8
Reagan	12	3	1	104

Table 21: Agreement of the predictions obtained with the Negative-Binomial fully Bayesian model using sets of underlying constants β no.2, no.20, and no.1 to no.21 composed by simple majority voting. The predictions were computed using the words obtained by the statistic T_1 for the all the "unknown" speeches.

Bayesian models. Nonetheless the three classifiers all agree on 207 out of 312 speeches (66.3%). In table 21, we notice the bold agreement between the predictions obtained with the sets of constants $\beta^{(2)}$ and $\beta^{(20)}$, which gave the most accurate results during cross-validation, and the predictions obtained aggregating by simple majority voting all the 21 sets of constants. Such an agreement suggests that the predictions we got with the Negative-Binomial model are consistent, and that different prior beliefs about the differential use of words by Reagan and others is not crucial in attributing authorship.

We provide in appendix A details about the predictions for 312 out of the 314 speeches whose author is uncertain; we do not provide predictions for the reprints of the radio addresses titled "Double Standard", and "The Superintendent's Dilemma" because they contain mainly quoted paragraphs.

5 Conclusions

The aims of this study were to determine the authorship of 314 of Ronald Reagan's 1970s radio broadcasts for which no direct evidence of authorship is available, and to provide an assessment of the confidence we have in the predictions of authorship. We used the study of "The Federalist" papers by Frederick Mosteller and David Wallace (1964, 1984) [24] [25] as a starting point for modeling word count data. From them we learned about the statistics T_1, about possible parameterizations and related estimation issues for Negative-Binomial counts when the sampling units (the texts) have different lengths, and we learned how to "bracket" the prior distributions. Then we fully explored the distributions of T_1 based on the Poisson and Negative-Binomial models to properly address the selection of features as a multiple testing problem, and we used both an ad-hoc word counts analysis and a semantic decomposition of the speeches to create features able to capture

elements of literary style beyond those affected by the frequency of function words, thus adding robustness to our predictions. Finally we cross-validated the accuracies of the fully Bayesian models, and assessed the goodness of the approximations of the log-odds at the posterior mode and at the posterior mean. We also compared our results with standard solutions to authorship attribution problems both in the Linguistic and Computer Science communities, and we concluded that in 1975 Ronald Reagan drafted 77 speeches, and his collaborators drafted 71, whereas over the years 1976-79 Reagan drafted 90 speeches whereas Hannaford drafted 74.

Some highlights of our analyses and assessments were:

1. The goodness-of-fit study indicated that the Negative-Binomial model was appropriate for word counts and semantic features counts data, and we based both our best word selection scheme, through thresholds for the statistic T_1, and the likelihood of the data upon it.

2. We chose the constants underlying the prior distributions with the aim of mitigating the variations in the use of words that would play a role in the attribution of authorship. We ran our experiments for 21 sets of constants, entailing possible scenarios, that we identified as "reasonable" with two small studies on 90 and 120 words, on speeches drafted by Ronald Reagan and other collaborators.

3. The remarkable descriptive power of the Negative-Binomial model fully translated into predictive power. The predictions we obtained with the fully Bayesian Negative-Binomial model were very much stable, both in terms of cross-validated accuracy across 21 sets of constants, and in terms of predicted author for the 312 "unknown" speeches.

4. We provided separate models for the speeches in 1975 and those in 1976-1979, and obtained stable and accurate predictions on speeches given in different years, about various topics.

5. The magnitude of the log-odds of authorship entailed clear-cut predictions for the authorship of many of the "unknown" speeches. Further the bold agreement of several accurate classification methods, based both on the analysis of words and on a semantic decomposition of the speeches, reinforced our confidence.

A major shortcut that we used in our models was the assumption of independence; (A1) independence of words from one another, and (A2) independence of words from positions in the text of the speeches. Even though in the independence study we examined (A1) briefly and (A2) thoroughly to discover that they statistically held in our data, and in our analysis we mostly focused on function words, we would expect (A1) not to be true in general. In particular a more desirable model would account for some functional form of dependence, for example "attraction and repulsion" among words along the lines of Beeferman, Berger and Lafferty (1997) [1]. Further, the extent to which (A2) holds is questionable for the positions at the end of sentences, and somewhat questionable for the positions at the beginning of sentences. Assumptions A1 and A2 are crucial in that they enabled us to cut down to a feasible number the probabilities to be estimated in several cases. But more importantly, because of our reliance on out-of-sample cross-validation the results of their application are not overstatements or misrepresentions. Rather, the assumptions relating to independence at worst only result in poorer accuracy than that that we might achieve if we captured dependence appropriately; cross-validated accuracies above 90% in all cases, and predicted authors for the "unknown" speeches stable across 21 possible scenarios are good, convincing advocates for the simplicity of our models.

We are packaging the models and methods we used in this report in a Java archive, with the addition of an automated model selector for feature counts, and of an extensive study about possible parameterizations and functional forms for the priors of frequent bi-grams. We plan to extend the analysis of the asymptotic distribution of T_1 to include third order terms and assess the goodness of the relative p-values, and we will consider an extension of T_1 to be able to deal with the case of multiple authors. We are not able to assess the goodness of the 166 sparse Docu-Scope semantic features at this time, because of the unavailability of the source code. Eventually we plan to conclude our preliminary explorations about whether John McClaughry wrote successfully in Reagan style, both in general and compared to Peter Hannaford. [22] We hope to carry out a more extensive analysis of the literary style of Ronald Reagan and of that of his collaborators in the future, possibly taking into account more documents such as letters.

[22] We need to define the notion of "success" for these purposes. For example, we can consider "success" as the lack of perfect separability of the texts by two authors in the space of features, where the surfaces that possibly separate the texts are computed using support vectors corresponding to a polynomial kernel.

Appendix A: Predictions for the Unknown Speeches

In this appendix we present the predictions of our best classifiers, along with the Negative-Binomial log-odds for the speeches delivered by Reagan over the years 1975-79 whose author is uncertain. In the tables below a 1 indicates that the most likely author is Ronald Reagan, for all years, whereas a 0 indicates that the most likely author is not Reagan, for speeches delivered in 1975, and Peter Hannaford for speeches delivered over the years 1976-79. In order to better appreciate the strength of the log-odds we give the following conversion table as a reference.

Log Odds	Odds	Log Odds	Odds
0	1 to 1	4	55 to 1
.1	1.1 to 1	5	150 to 1
.5	1.6 to 1	10	$22,000$ to 1
1	$e \approx 2.7$ to 1	15	3.3×10^6 to 1
2	7 to 1	20	480×10^6 to 1
3	20 to 1	25	7.1×10^9 to 1

Table 22: Table of anti-logs.

Prog.	Speech ID	Logistic Regression	Naive Bayes Multinomial	Docu-Scope Neg.-Bin.		Full Bayes Poisson		Full Bayes Neg.-Bin.	
				Majority All sets β	Set $\beta^{(2)}$ log-odds	Majority All sets β	Set $\beta^{(2)}$ log-odds	Majority All sets β	Set $\beta^{(2)}$ log-odds
1	75-01-a1.txt	1	1	0	-0.16	1	6.6	1	1.1
2	75-01-a2.txt	0	0	1	0.16	0	-23.5	0	-6.5
3	75-01-a4.txt	0	0	0	-0.32	0	-1.5	0	0.1
4	75-01-a5.txt	1	1	1	0.31	1	22.6	1	9.2
5	75-01-b1.txt	1	1	1	0.33	1	25.8	1	13.2
6	75-01-b2.txt	1	1	0	0.01	1	24.5	1	12.7
7	75-01-b3.txt	1	1	0	-0.03	1	18.8	1	22.8
8	75-01-b4.txt	0	0	1	0.24	1	4.3	0	-0.1
9	75-01-b5.txt	1	1	0	-0.79	0	-12.6	0	-10.5
10	75-01-b6.txt	0	0	1	0.34	0	-2.2	0	-2.0
11	75-01-b7.txt	1	1	1	0.07	1	13.8	1	7.8
12	75-01-b8.txt	0	0	0	-0.51	0	2.6	0	-6.6
13	75-02-a1.txt	1	1	1	0.54	1	20.1	1	2.8
14	75-02-a2.txt	1	0	0	-0.08	1	0.7	0	-3.3
15	75-02-a3.txt	1	1	0	-0.40	0	-3.4	0	1.3
16	75-02-a4.txt	1	1	0	-0.38	0	-1.2	0	-6.6
17	75-02-a5.txt	1	1	1	0.24	1	14.3	1	3.6
18	75-02-b1.txt	1	1	0	-0.22	1	10.2	1	0.6
19	75-02-b2.txt	1	1	0	-0.10	1	10.2	1	7.9
20	75-02-b3.txt	1	1	0	-0.49	1	25.0	1	9.0
21	75-02-b4.txt	1	1	1	0.23	1	46.9	1	19.1
22	75-02-b5.txt	0	0	0	-0.06	0	-1.2	0	-1.4
23	75-02-b6.txt	1	1	0	-0.74	0	0.8	0	-6.0
24	75-03-a1.txt	1	0	0	-0.43	0	-12.1	0	-10.0
25	75-03-a2.txt	1	1	0	-0.83	1	3.5	0	-3.7
26	75-03-a3.txt	1	0	0	-0.17	0	-13.9	0	-12.2
27	75-03-a5.txt	1	1	1	0.01	1	38.6	0	0.8
28	75-03-b5.txt	1	1	0	-0.16	1	18.9	1	22.4
29	75-03-b6.txt	1	0	1	0.19	0	-10.7	0	-6.9
30	75-04-a3.txt	1	1	0	0.02	1	5.2	1	2.8
31	75-04-a4.txt	1	0	1	0.13	0	4.8	1	3.9
32	75-04-a5.txt	1	1	1	0.06	1	10.5	1	6.1
33	75-04-a6.txt	1	0	0	-0.47	0	-1.9	0	2.6
34	75-04-b1.txt	1	0	1	0.28	0	-1.0	1	1.2
35	75-04-b4.txt	1	1	0	-0.24	1	8.4	0	-0.8
36	75-04-b5.txt	1	0	1	0.19	0	-6.4	0	-2.8
37	75-05-a1.txt	1	1	0	-0.29	0	-1.8	0	-5.9
38	75-05-a2.txt	1	1	1	0.11	1	18.5	1	20.7
39	75-05-a3.txt	0	0	1	0.50	0	-8.3	0	-6.2
40	75-05-a4.txt	1	1	0	-0.46	0	4.1	1	4.5
41	75-05-b2.txt	1	0	1	0.19	0	-4.3	0	0.5
42	75-05-b4.txt	0	0	0	-0.03	0	-2.7	0	-0.7
43	75-05-b6.txt	0	1	0	-0.49	0	-4.8	0	-5.2
44	75-06-a1.txt	1	0	0	-0.38	0	-15.3	0	-10.9
45	75-06-a2.txt	1	1	0	-0.10	1	0.1	0	-5.5
46	75-06-a3.txt	1	1	0	-0.16	1	11.7	1	3.5
47	75-06-a4.txt	1	1	1	0.29	1	6.8	1	1.0
48	75-06-a5.txt	1	1	0	-0.12	0	-7.0	0	-6.5
49	75-07-a3.txt	1	0	0	-0.05	0	-8.3	0	-9.2
50	75-07-a4.txt	0	0	1	0.02	0	-5.6	0	-7.8
51	75-07-a5.txt	1	0	1	0.74	0	-5.7	0	-10.2
52	75-07-b1.txt	1	1	1	0.05	1	2.2	1	0.0
53	75-07-b2.txt	0	0	0	-0.02	0	-4.1	0	-4.5
54	75-08-a1.txt	1	1	1	0.45	1	43.5	1	22.1
55	75-08-a2.txt	1	1	1	0.36	1	15.4	1	5.7
56	75-08-a3.txt	1	1	0	-0.55	0	-3.3	0	-5.0
57	75-08-a4.txt	1	1	0	-0.03	1	1.2	1	0.8
58	75-08-a5.txt	1	1	0	-0.13	0	1.3	0	-1.6
59	75-08-b1.txt	1	1	0	-0.15	1	8.9	1	4.0
60	75-08-b3.txt	0	1	0	-0.05	0	-6.3	0	-10.2
61	75-08-b5.txt	1	1	1	0.04	1	33.0	1	10.2
62	75-08-b6.txt	1	1	1	0.05	1	9.0	1	4.3
63	75-09-a1.txt	0	1	0	-0.28	1	2.3	0	-8.3
64	75-09-a2.txt	1	0	0	-0.31	0	-21.1	0	-18.7
65	75-09-a3.txt	1	1	0	-0.01	1	24.0	0	-1.4
66	75-09-a5.txt	1	0	0	-0.40	0	-10.0	0	-7.7
67	75-09-b2.txt	1	1	1	0.32	1	12.6	1	6.5
68	75-09-b3.txt	1	0	0	-0.12	0	-10.5	0	-9.0
69	75-10-a1.txt	1	0	0	-0.31	0	-10.9	0	-9.2
70	75-10-a4.txt	1	1	1	0.00	1	4.2	1	17.0
71	75-10-a5.txt	0	1	1	0.29	0	-12.3	0	-13.2
72	75-10-a6.txt	0	1	0	-0.23	0	2.0	0	-3.7
73	75-10-b1.txt	1	1	0	-0.60	0	-0.9	0	-2.4
74	75-10-b2.txt	1	1	1	0.53	1	4.3	1	1.5
75	75-10-b3.txt	1	1	1	0.23	0	-3.4	0	-3.5

Prog.	Speech ID	Logistic Regression	Naive Bayes Multinomial	Docu-Scope Neg.-Bin.		Full Bayes Poisson		Full Bayes Neg.-Bin.	
				Majority All sets β	Set $\beta^{(2)}$ log-odds	Majority All sets β	Set $\beta^{(2)}$ log-odds	Majority All sets β	Set $\beta^{(2)}$ log-odds
76	75-10-b4.txt	1	1	0	-0.26	0	-6.5	0	-4.4
77	75-11-b1.txt	1	1	1	0.09	1	10.0	1	0.9
78	75-11-b2.txt	0	1	0	-0.48	0	1.4	0	-1.6
79	75-11-b3.txt	1	1	1	0.75	1	25.2	1	18.1
80	75-11-b4.txt	1	0	0	-0.24	0	1.9	1	0.2
81	75-12-a1.txt	0	1	1	0.09	1	10.6	1	7.8
82	75-12-a3.txt	0	0	0	-0.12	0	-18.4	0	-15.9
83	75-12-a4.txt	1	1	0	-0.11	0	-14.9	0	-13.0
84	75-12-a5.txt	0	0	0	-0.41	0	-13.0	0	-13.0
85	75-12-a6.txt	1	0	0	-0.19	0	-7.1	0	-9.6
86	75-12-b1.txt	1	0	0	-0.75	0	-8.4	0	-6.3
87	75-12-b2.txt	1	0	1	0.14	0	-8.6	0	-7.4
88	75-12-b3.txt	0	1	1	0.01	1	11.1	1	4.5
89	75-12-b4.txt	1	1	1	0.20	1	7.1	0	-1.0
90	75-12-b5.txt	1	1	1	0.01	1	3.0	1	-0.6
91	75-12-b6.txt	1	0	1	0.60	0	-4.1	0	-4.1
92	75-13-b1.txt	1	1	0	-0.13	0	0.3	0	-2.1
93	75-13-b2.txt	1	1	1	0.36	1	10.6	1	2.3
94	75-13-b3.txt	0	0	1	0.08	1	7.1	1	4.2
95	75-13-b4.txt	1	1	1	0.33	1	1.5	1	-2.0
96	75-13-b5.txt	1	0	0	-0.47	0	-10.4	0	-4.7
97	75-13-b6.txt	0	0	1	0.02	0	-11.5	0	-10.9
98	75-14-a2.txt	1	1	1	0.13	1	14.9	0	-2.9
99	75-14-a3.txt	0	1	0	-0.44	0	-6.1	0	-1.2
100	75-14-a4.txt	0	0	0	-0.08	0	-14.7	0	-12.7
101	75-14-a5.txt	1	1	1	0.65	1	32.9	1	27.6
102	75-15-a1.txt	1	0	0	-0.11	0	-6.6	0	-10.0
103	75-15-a2.txt	1	1	0	-0.12	1	1.3	1	-2.4
104	75-15-a3.txt	1	1	1	0.20	0	-4.7	0	-4.2
105	75-16-a2.txt	1	0	1	0.18	1	2.6	1	0.3
106	75-16-a3.txt	1	1	0	-0.18	1	4.7	1	2.6
107	75-17-a1.txt	1	0	1	0.20	0	-12.2	0	-9.8
108	75-17-a10.tx	1	1	1	0.12	0	-0.2	0	-3.9
109	75-17-a2.txt	1	1	0	-0.14	1	19.2	1	21.1
110	75-17-a3.txt	1	0	1	0.14	1	2.3	1	0.0
111	75-17-a4.txt	1	1	1	0.03	1	4.9	1	2.4
112	75-17-a5.txt	1	1	0	-0.13	0	-6.0	0	-5.2
113	75-17-a7.txt	1	1	0	-0.64	0	-2.7	1	3.8
114	75-17-a8.txt	1	1	1	0.10	1	3.1	1	0.6
115	75-17-a9.txt	1	0	0	-0.47	0	-7.6	0	-7.1
116	75-18-a1.txt	1	1	1	0.20	1	19.3	1	13.3
117	75-18-a2.txt	1	0	0	-0.07	0	1.7	0	-0.5
118	75-18-a3.txt	1	1	0	-0.16	1	14.8	1	14.5
119	75-18-a6.txt	1	1	1	0.04	1	21.7	1	6.3
120	75-18-a8.txt	1	1	0	-0.24	0	0.8	0	4.1
121	75-18-a9.txt	0	0	0	-0.06	0	-3.2	0	-1.8
122	75-19-a1.txt	0	0	0	-0.04	0	-5.3	0	-6.5
123	75-19-a2.txt	0	0	1	0.29	0	-6.8	0	-8.0
124	75-19-a3.txt	1	1	0	-0.15	1	0.6	1	-0.4
125	75-19-a4.txt	0	0	0	-0.22	0	-12.1	0	-11.6
126	75-19-a5.txt	0	0	1	0.31	0	-12.8	0	-11.3
127	75-19-a6.txt	1	1	0	-0.08	0	1.0	1	2.4
128	75-19-b1.txt	1	0	0	-0.25	0	-6.8	0	-0.5
129	75-19-b2.txt	0	0	0	-0.33	0	-33.6	0	-24.3
130	75-19-b4.txt	1	1	0	-0.18	1	9.8	1	6.3
131	75-19-b6.txt	1	1	1	0.64	1	15.2	1	7.5
132	75-20-a1.txt	1	1	1	0.16	1	19.4	1	11.1
133	75-20-a4.txt	1	1	0	-0.32	1	16.3	1	7.8
134	75-20-a5.txt	0	0	1	0.10	0	-7.0	0	3.8
135	75-20-a6.txt	0	1	0	-0.08	1	15.8	1	5.4
136	75-20-b1.txt	0	0	0	-0.47	0	-0.1	0	-0.8
137	75-20-b2.txt	1	0	0	-0.28	0	-10.7	0	-11.9
138	75-20-b4.txt	1	1	0	-0.15	1	1.7	1	1.0
139	75-21-a1.txt	1	1	1	0.08	0	-10.8	0	-11.5
140	75-21-a2.txt	1	1	1	0.10	1	17.1	1	8.1
141	75-21-a3.txt	1	1	0	-0.09	1	2.4	1	1.8
142	75-21-a4.txt	1	1	0	-0.56	0	-4.6	0	-8.0
143	75-21-a5.txt	1	0	1	0.10	1	15.0	1	5.7
144	75-21-a6.txt	1	1	0	-0.10	0	-0.6	0	-1.9
145	75-21-a7.txt	1	1	0	-0.21	1	9.4	1	3.6
146	75-21-a8.txt	1	1	0	-0.17	1	15.0	1	2.1
147	75-21-a9.txt	1	1	0	-0.21	1	4.5	1	18.7
148	75-22-a1.txt	1	1	1	0.31	1	18.0	1	8.0

Prog.	Speech ID	Logistic Regression	Naive Bayes Multinomial	Docu-Scope Neg.-Bin.		Full Bayes Poisson		Full Bayes Neg.-Bin.	
				Majority All sets β	Set $\beta^{(2)}$ log-odds	Majority All sets β	Set $\beta^{(2)}$ log-odds	Majority All sets β	Set $\beta^{(2)}$ log-odds
149	76-01-a2.txt	1	1	0	-0.70	1	15.0	1	12.9
150	76-01-a5.txt	1	1	0	-0.65	1	9.9	1	4.6
151	76-01-b2.txt	1	1	1	0.13	1	5.1	1	5.6
152	76-01-b4.txt	1	1	0	-0.28	1	7.7	1	3.9
153	76-02-a1.txt	0	0	1	0.31	0	-40.5	0	-17.0
154	76-02-a2.txt	1	0	0	-0.01	1	1.2	0	-0.1
155	76-02-a3.txt	1	1	1	0.37	0	-12.7	0	-10.5
156	76-02-a7.txt	0	0	0	-0.39	0	-0.4	1	1.3
157	76-02-b2.txt	1	0	0	-0.26	0	-6.9	0	-3.4
158	76-02-b5.txt	1	1	0	-0.12	1	3.0	1	2.6
159	76-02-b7.txt	1	1	0	-0.46	1	5.4	1	2.8
160	76-03-a1.txt	0	0	1	0.14	0	-22.2	0	-9.1
161	76-03-a2.txt	1	0	0	-1.25	0	-6.2	0	-7.1
162	76-03-a4.txt	1	1	1	0.26	0	-4.9	0	-4.8
163	76-03-a5.txt	0	0	0	-0.26	0	-4.9	0	-2.8
164	76-03-a6.txt	1	1	1	0.27	0	-0.5	0	-1.1
165	76-03-b1.txt	1	1	1	0.33	1	3.8	1	4.2
166	76-03-b2.txt	1	1	0	-0.15	1	8.5	1	6.4
167	76-03-b5.txt	1	1	1	0.20	1	10.4	1	9.2
168	76-03-b6.txt	1	1	0	-0.49	1	12.2	1	8.8
169	76-04-a1.txt	1	1	0	-0.01	0	-2.4	0	-2.7
170	76-04-a4.txt	0	1	0	-0.03	1	0.0	0	-0.2
171	76-04-a7.txt	0	0	0	-0.28	1	4.1	1	3.8
172	76-04-b2.txt	0	0	0	-0.20	0	-12.5	0	-3.4
173	76-04-b4.txt	1	0	1	0.04	0	-5.6	0	-3.0
174	76-04-b5.txt	1	1	1	0.42	1	0.3	0	-0.3
175	76-04-b6.txt	0	0	0	-0.62	0	-12.6	0	-6.0
176	76-04-b8.txt	0	0	0	-0.08	0	-20.1	0	-13.3
177	76-05-a1.txt	1	1	0	-0.13	1	3.4	1	2.6
178	76-05-a3.txt	1	1	1	0.15	1	8.2	1	5.3
179	76-05-a4.txt	1	1	1	0.07	1	4.3	1	3.2
180	76-05-a5.txt	0	0	0	-0.03	0	-3.2	0	-0.5
181	76-05-b2.txt	0	0	0	-0.63	0	-16.8	0	-11.5
182	76-05-b7.txt	1	1	0	-0.25	1	1.3	1	4.4
183	76-06-a1.txt	1	1	1	0.29	1	9.2	1	6.0
184	76-06-a2.txt	0	1	0	-0.24	0	-5.2	0	-4.6
185	76-06-a3.txt	0	1	0	-0.46	1	4.0	1	3.8
186	76-06-a4.txt	0	0	0	-0.09	0	-2.5	0	-0.3
187	76-06-a5.txt	1	1	0	-0.76	1	4.7	1	3.7
188	76-06-b8.txt	1	1	0	-0.15	1	2.5	1	2.3
189	76-07-a1.txt	1	1	0	-0.22	1	4.9	1	1.8
190	76-07-a2.txt	0	1	0	-0.27	0	-3.8	0	-1.5
191	76-07-a3.txt	0	1	0	-0.37	1	2.0	1	3.4
192	76-07-b3.txt	0	1	0	-0.73	0	-11.7	0	-6.0
193	76-07-b4.txt	1	1	1	0.30	1	3.5	1	3.0
194	76-07-b6.txt	1	0	0	-0.14	1	3.1	1	2.8
195	76-07-b7.txt	0	0	1	0.16	0	-16.6	0	-7.8
196	76-07-b8.txt	0	1	1	0.06	1	3.2	1	2.8
197	76-08-a1.txt	1	1	0	-0.55	1	11.4	1	0.6
198	76-08-a3.txt	1	0	1	0.06	0	-8.6	0	-6.7
199	76-09-b1.txt	1	1	1	0.09	1	9.0	1	7.5
200	76-09-b3.txt	0	0	0	-0.28	1	2.1	1	1.4
201	76-09-b4.txt	1	1	0	-0.23	1	5.4	1	3.1
202	76-10-b7.txt	0	1	0	-0.11	0	-0.8	0	-0.1
203	76-10-b8.txt	1	1	0	-0.23	1	7.9	1	6.8
204	76-11-a2.txt	1	0	1	0.12	0	-9.5	0	-5.5
205	76-11-a5.txt	1	1	1	0.18	1	9.4	1	7.4
206	76-12-a7.txt	0	0	1	0.06	0	-1.3	0	-1.8
207	76-12-b3.txt	0	0	1	0.03	1	5.2	1	4.4
208	76-13-a1.txt	1	1	1	0.09	1	17.6	1	15.6
209	76-13-b4.txt	1	1	1	0.09	0	-0.2	0	-0.1
210	76-13-b8.txt	0	1	1	0.89	0	-2.3	0	-0.5
211	76-14-a5.txt	1	0	0	-0.31	0	-5.7	0	-5.3
212	76-15-a1.txt	1	0	0	-0.30	0	-18.6	0	-6.6
213	76-16-b2.txt	1	1	0	-0.05	1	9.3	1	2.9
214	76-17-a2.txt	1	1	1	0.24	1	7.0	1	5.6
215	76-17-a3.txt	1	1	0	-0.33	1	8.0	1	4.1
216	76-17-a4.txt	1	1	1	0.40	1	20.3	1	8.1
217	76-17-a5.txt	1	1	1	0.38	1	6.9	1	6.5
218	76-17-a6.txt	0	1	1	0.40	0	-5.5	0	-2.5
219	76-17-a7.txt	1	1	1	0.21	1	11.8	1	7.5
220	76-17-a8.txt	1	1	0	-0.09	1	21.8	1	12.9
221	76-17-b1.txt	1	1	1	0.12	1	27.7	1	17.1
222	76-17-b2.txt	1	1	0	-0.21	1	19.5	1	11.4
223	76-17-b3.txt	1	1	1	0.16	1	9.4	1	3.8
224	76-17-b4.txt	1	1	1	0.27	1	7.8	1	6.0
225	76-17-b5.txt	1	1	1	0.56	1	6.0	1	4.9
226	76-17-b6.txt	1	1	0	0.00	1	10.7	1	7.6
227	76-17-b7.txt	1	1	0	-0.16	0	-1.8	0	-0.7
228	76-17-b8.txt	1	0	0	-0.32	0	-6.4	0	-5.3
229	77-20-a3.txt	0	0	0	-0.05	0	-18.3	0	-5.6
230	77-20-a4.txt	1	0	0	-0.19	0	-39.8	0	-13.8

Prog.	Speech ID	Logistic Regression	Naive Bayes Multinomial	Docu-Scope Neg.-Bin.		Full Bayes Poisson		Full Bayes Neg.-Bin.	
				Majority All sets β	Set $\beta^{(2)}$ log-odds	Majority All sets β	Set $\beta^{(2)}$ log-odds	Majority All sets β	Set $\beta^{(2)}$ log-odds
231	77-20-b8.txt	1	1	0	-0.10	1	10.2	1	7.0
232	77-21-a2.txt	0	0	1	0.05	0	-23.0	0	-14.2
233	78-01-a1.txt	0	0	0	-0.12	0	-19.6	0	-13.8
234	78-02-b6.txt	1	0	0	-0.17	0	-1.5	0	-1.5
235	78-02-b7.txt	1	1	0	-0.27	1	4.7	1	3.9
236	78-03-a1.txt	0	0	0	-0.74	0	-15.8	0	-12.7
237	78-03-a4.txt	1	0	0	-0.05	0	-0.9	0	-0.4
238	78-03-a6.txt	1	1	0	-0.59	0	-11.7	0	-7.2
239	78-03-b6.txt	1	1	0	-0.54	1	7.6	1	3.4
240	78-06-b3.txt	0	1	0	-0.36	1	5.2	1	2.3
241	78-06-b7.txt	1	1	1	0.21	1	3.0	1	2.3
242	78-06-b8.txt	0	0	0	0.01	0	-19.0	0	-11.8
243	78-08-b7.txt	0	0	1	0.15	0	-11.8	0	-6.5
244	78-09-a1.txt	1	0	0	-0.74	0	-13.2	0	-13.6
245	78-09-a2.txt	1	1	0	-0.56	0	-0.9	0	-5.2
246	78-10-a7.txt	1	1	0	-0.29	0	-3.3	0	-1.7
247	78-10-b5.txt	1	1	1	0.36	1	2.1	1	2.9
248	78-13-a1.txt	0	0	1	0.24	0	-16.0	0	-9.9
249	78-13-a4.txt	1	1	1	0.13	1	1.7	1	2.3
250	78-13-b1.txt	0	1	0	-0.26	1	8.3	0	-3.3
251	78-14-b5.txt	0	1	0	-0.36	0	-1.1	0	0.2
252	78-14-b6.txt	0	0	1	0.31	0	-16.0	0	-10.5
253	78-14-b8.txt	0	0	0	-0.57	1	1.9	1	4.0
254	78-15-a1.txt	1	1	0	-0.06	1	3.8	1	2.6
255	78-15-a2.txt	0	1	1	0.08	1	3.3	1	2.6
256	78-15-b7.txt	1	1	0	-0.20	0	-0.3	0	-1.0
257	78-17-a1.txt	1	1	1	0.20	1	14.1	1	6.2
258	79-01-a1.txt	1	0	0	-0.23	0	-30.4	0	-14.9
259	79-02-a5.txt	1	0	0	0.00	0	-4.2	0	-2.2
260	79-02-a6.txt	0	0	1	0.15	0	-6.9	0	-2.1
261	79-03-a1.txt	1	1	1	0.40	0	-2.4	0	0.1
262	79-03-a2.txt	1	1	1	0.39	1	3.3	1	2.8
263	79-03-a3.txt	1	0	1	0.17	0	-12.0	0	-9.0
264	79-04-a2.txt	0	0	0	-0.04	0	-12.6	0	-9.3
265	79-04-b8.txt	1	0	0	-0.11	1	1.7	1	3.4
266	79-05-a4.txt	1	1	1	0.21	1	8.8	1	6.6
267	79-06-a2.txt	0	0	0	-0.15	0	-2.6	0	-2.4
268	79-06-a3.txt	1	1	0	-0.22	1	11.0	1	4.0
269	79-06-a4.txt	1	0	1	0.08	0	-9.3	0	-5.9
270	79-07-a4.txt	1	0	1	0.02	1	3.6	1	1.4
271	79-07-a5.txt	1	1	0	-0.66	0	-5.0	0	-3.1
272	79-07-b7.txt	0	1	1	0.48	1	4.8	1	5.3
273	79-07-b8.txt	1	1	1	0.23	1	6.0	1	4.1
274	79-08-a6.txt	1	1	0	-0.05	1	7.9	1	3.4
275	79-08-a7.txt	0	0	0	-0.14	0	-1.2	0	-1.4
276	79-08-b2.txt	1	0	0	-0.09	0	-6.6	0	-4.6
277	79-08-b5.txt	1	1	0	-0.06	1	35.4	1	15.2
278	79-08-b8.txt	1	0	0	-0.13	1	1.4	1	2.1
279	79-10-a1.txt	1	0	0	-0.45	0	-10.0	0	-7.1
280	79-10-a2.txt	0	1	1	0.12	1	2.2	0	-0.6
281	79-10-b3.txt	1	1	0	-0.35	1	3.6	1	2.4
282	79-10-b5.txt	1	1	1	0.19	1	10.1	1	8.0
283	79-10-b6.txt	1	1	1	0.43	1	6.5	1	6.3
284	79-10-b8.txt	1	1	0	-0.18	1	0.1	1	1.9
285	79-11-a1.txt	1	1	1	0.01	0	-3.5	0	-3.3
286	79-11-a2.txt	1	1	0	-0.72	1	14.2	1	10.3
287	79-11-a4.txt	1	0	0	-0.17	0	-7.2	0	-5.2
288	79-11-a6.txt	1	1	1	-0.01	1	5.7	0	-1.8
289	79-11-a7.txt	1	1	1	0.26	1	3.6	1	3.3
290	79-11-b2.txt	1	0	0	-0.60	1	0.1	1	2.1
291	79-11-b3.txt	1	1	1	0.18	0	-3.0	0	-1.5
292	79-12-a1.txt	1	0	0	-0.41	0	-5.3	0	-4.1
293	79-12-a2.txt	0	0	0	-0.52	0	-5.5	0	-4.3
294	79-12-a3.txt	0	1	0	-0.30	0	-6.7	0	-5.0
295	79-12-a4.txt	1	0	0	-0.26	0	-0.4	0	-0.7
296	79-12-a5.txt	1	0	0	-0.31	0	-4.6	0	0.5
297	79-12-a6.txt	1	1	1	0.06	1	4.3	1	2.2
298	79-12-a7.txt	1	0	1	0.45	1	7.1	1	6.3
299	79-12-b1.txt	1	0	0	-0.78	0	-4.1	0	-1.5
300	79-12-b3.txt	1	0	1	0.20	0	-5.4	0	-4.1
301	79-12-b8.txt	1	0	0	-0.49	1	3.3	1	3.6
302	79-13-b6.txt	0	0	0	-0.04	0	-1.5	0	-0.5
303	79-13-b7.txt	0	0	1	0.02	0	-8.0	0	-5.1
304	79-13-b8.txt	1	1	0	-0.33	1	7.7	1	3.9
305	79-14-a2.txt	1	1	1	0.05	1	6.4	1	3.3
306	79-14-a7.txt	1	1	0	-0.28	1	9.6	1	7.8
307	79-14-b7.txt	1	0	0	-0.47	0	-5.3	0	-1.2
308	79-15-a1.txt	1	1	0	-0.74	1	6.3	1	5.4
309	79-15-a2.txt	1	1	0	-0.10	0	-5.5	0	-3.1
310	79-15-a3.txt	1	1	0	-0.52	1	7.7	1	5.6
311	79-15-a4.txt	1	1	0	-0.86	1	14.8	1	6.0
312	79-15-a5.txt	1	1	0	-0.22	1	28.2	1	8.2

Appendix B: The Secrets of Ronald Reagan's Writing Style

In section 4.5 we showed that using certain words to predict the author of a speech with fully Bayesian methods yielded cross-validated accuracies of 90% or more, very stable across different sets of underlying constants. Here we look for the best selection of words, among those we obtained both using p-values for T_1[23] with FDR correction, and using the Information Gain (IG) computed according to Multinomial and multivariate Bernoulli models.

As we noticed in section 4.2 the words we obtain with T_1 capture elements of writing style of an author that affect the frequency of use of words. Further the final set of words we obtained with T_1 included the words we obtained with high Information Gain on both Multinomial and multivariate Bernoulli models, with some minor exceptions. The problem of finding the best combination of words in terms of cross-validated accuracy is NP-hard. We adopted several strategies to obtain reasonably good combinations, and we acknowledge that some better ones may exist. The main idea was to look for some combinations with apparent accuracy above 95% on the texts of both authors, which is faster but still NP-hard, and then to cross-validate those. The threshold of 95% for the apparent accuracy was chosen because that was about the apparent accuracy of the best classifiers in section 4.3, whose cross-validated accuracy is above 90%. We implemented mainly 4 strategies to find good combinations of words, and played with them:

(AC) All Combinations: this is the exhaustive search. It became non-practical as the number of words in the pools we considered grew above 20.

(RS) Random Sampling: sampling 1000 random combinations of words of different sizes, from 1 to 117, returned combinations with apparent accuracy above 95% on all authors. All of these combinations would yield cross-validated accuracies below 90% on at least one author, and as high as 100% on the other.

(LS) Local Search: this iterative procedure starts from a given combination, and at every step includes or excludes the word that yields the maximum increase in apparent accuracy. There are plenty of local maxima, though, and the average number of iterations before stopping was about 3 on many experiments. Eventually we combined RS and LS to perform random searches locally optimal. This strategy returned good combinations, with apparent accuracy as high as 98% on Reagan texts, and 96% on Hannaford texts.

(RW) Random Walk: finally this strategy starts from a given combination and includes or excludes a words from a pool according to a set of probabilities, at each step. First we need to choose the word whose status in the combination (presence/absence) we want to modify, say, with equal probabilities or according to P_1, and then we modify its status according to the probabilities in $P_2(word)$. For example, in a no-information situation the set $P_2, (word)$ may be composed of probabilities $P(\text{word is included}|\text{chosen}) = P(\text{word is excluded}|\text{chosen}) = 0.5$. In our case, given that the sets of about 30 words obtained with Information Gain gave a cross-validated accuracy of about 90% on all authors for almost all the sets of underlying constants, we assumed that they were more likely to be part of good combinations, and we increased their probabilities of inclusion accordingly, say, to be directly proportional to their IG score or to their t_1 value.

[23] For each word p-values were computed on 3 simulated distributions for T_1. Specifically 1'000'000 t_1 values were obtained simulating word counts according to Negative-Binomial models for Reagan, the alternative author, and both. For the FDR correction we picked the largest p-values, among the three, in order to be conservative, and we used $\alpha = 0.05$.

Notice that equi-probable RW would find the best combination eventually, but the time get there may not be finite! In fact the probability of moving towards the best combination would decrease exponentially, the closer to it we would get. The biased RW had reasonable chances to find good combinations.

Below we present tables with a final pool of words, starting point for future analyses. The tables contain words obtained with IG models and T_1, for discriminating texts of different authors, hence a word may appear in different tables. The differences from the tables in appendix 23 is that here we used a further batch of 4 news columns, not available before.

Reagan versus Hannaford — all 1-Grams						
DDD	after	are	assumption	basic	big	carter
cents	chances	cheap	context	current	depriving	despite
devastating	endless	enough	entire	especially	free	future
get	group	groups	hear	heavy	her	huge
if	in	indeed	ironically	issue	it	its
joke	large	last	may	measure	measures	message
money	nearly	new	not	notion	of	on
ones	other	our	over	percent	plenty	popular
predictably	problems	rates	recent	scarcely	seems	sharp
she	similar	soon	special	story	strong	that
the	then	they	this	though	thus	till
to	total	under	us	was	we	week
were	white	will	worth	would	your	

Reagan versus Hannaford — all N-Grams						
DDD-percent	DDD-to	about-DDD	after-the	all-of	and-he	and-so
as-a-result	as-if	a-big	carter-administration	carter-s	for-DDD	he-would
human-rights	if-the	if-we	in-all	in-america	in-our	in-the
it-out	it-would	i-m	i-ve	last-week	may-be	more-than-DDD
mr-carter	no-wonder	of-all	of-our	our-own	per-cent	president-carter
put-it	seems-to	than-DDD	that-it	that-the	they-were	the-carter
the-issue	the-joke	the-other	through-the	to-its	to-say	we-d
we-had	we-ve	when-it-comes				

Reagan versus Others — all 1-Grams						
DDD	after	alternative	assumption	aware	basic	benefit
carter	center	continuous	current	despite	easy	efficiency
endless	equal	especially	expressed	fair	future	got
greatest	he	her	huge	ignored	important	improve
incidentally	initiative	intelligence	ironically	issue	it	its
joke	jokes	large	lot	may	me	measure
measures	men	message	model	most	nearly	negative
new	no	not	notion	now	of	on
only	ordinary	otherwise	our	over	people	percent
plenty	popular	positive	pretty	problems	rates	regard
rhetoric	scarcely	scheme	self	sharp	sharply	she
similar	since	small	something	soon	special	spirit
story	strength	strong	supposed	sure	surprisingly	the
their	then	there	though	thus	to	total
under	us	various	was	we	which	why
wife	will	would	yes			

Reagan versus Others — all N-Grams						
DDD-percent	DDD-to	DDD-years	about-DDD	about-to	after-the	all-of
america-s	and-all	and-of	and-of-course	and-other	and-our	as-a-result
a-big	a-few-of	a-lot	carter-administration	carter-s	don-t-know	d-like
for-DDD	how-many	if-we	in-all	in-america	in-our	in-the-case-of
in-the-future	it-s-not	it-would	i-m	i-m-sure	i-ve	last-week
let-me	let-s-hope	like-to	may-be	mr-carter	no-wonder	of-how
of-our	of-the	on-the-other	our-government	our-own	plenty-of	president-carter
put-it	quite-a-few	regard-to	right-to	seems-to	supposed-to	than-DDD
that-it	that-the	there-s	they-don	the-carter	the-issue	the-joke
the-law	the-notion	the-right	this-was	told-of	to-its	to-say
to-the	was-DDD	we-can	we-could	we-d	we-had	we-re
we-ve	we-were	when-it-comes	with-regard			

Appendix C: Exploratory Data Analysis

In this section we briefly discuss how we polished the original dataset, and how we parsed the text into the words and N-grams we used in the rest of the analysis.

Typos

We used the Unix script `ispell` to filter all the texts of the speeches, and to eliminate typos. There were 559 files that contain typos out of 1032 total, each file contained more than one typo on average. We corrected only obvious typos, following the general heuristic of not being intrusive.

A number of things popped to our attention. Unexpectedly there seemed to be no rules about when a word is listed in the dictionary as a dashed word or not; we only separated (with a dash) words that did not exist in the Merriam-Webster on-line dictionary like `manhours -> man-hours`, `poohpooh -> pooh-pooh` or `oldfashioned -> old-fashioned`. Many words existed in two forms; we left all the words that were listed in two variants like `labeled` and `labelled` the way we found them[24], without any concern about the consistency of the vocabularies, since the OCR process correctly captured whatever was on the paper copies of the speeches, a fact that was actually quite true as we show below. Reagan probably made up some words; we did not change words that supposedly Reagan made up like `depoliticalization`[25] or `sneeringly`.

The most curious situation arose when we considered British versus American spelling of words ending in `ter` as opposed to `tre`. The big surprise is that the President of the United States used the British variation, like in `theatre`, more than once! The natural question to ask at this point is whether something went wrong with the OCR application used to scan the original paper documents. Even more surprisingly it was not the OCR, in fact the paper copies of the speeches themselves clearly show a `theatre` in `75-09-a7.txt`, but a `theater` in `75-18-a7.txt`, for example. We turned to the experts at this point, quite puzzled.

Annelise Anderson explained that: "Reagan spelled it both ways. The spelling could have been changed from an original manuscript during retyping or editing. Before a radio address was given, no matter who did the original draft, it was always read and edited by Reagan; most of them were edited by Hannaford (except for a few when Hannaford might have been on vacation). Reagan had the opportunity to make last-minute changes just before he recorded a batch of addresses. Reagan was very inconsistent about spelling and punctuation. Sometimes Hannaford's editing would break something up into two sentences.

The original drafts were typed, both in the offices of Deaver & Hannaford and in the offices of Harry O'Connor (where they were recorded) by many different people over the years, who had different views of spelling, capitalization, and so forth. So these differences would not be indicators of authorship".

Dictionary Rules

The set of parsing rules in our PERL script were the same rules used by automated systems for text classification like Bag of Words (BOW) developed at Carnegie-Mellon Computer Science Department [17], and Gerard Salton's SMART developed at Cornell CSD — in this report we refer

[24] It is true that we lost frequency in doing this, but luckily enough such words were all too contextual to be considered markers in our analysis anyway.

[25] Instead of depoliticization.

to this set of rules as CS rules. We implemented the possibility to retrieve N-tuples of adjacent words, which are commonly referred to as N-grams (or N-tokens[26]).

The main feature of the CS set of rules is that we collapsed digits into three categories `DDD`, `DD:DD` and `$DDD`. Further a preliminary look at the N-tokens showed that both expressions like `$20` and `20 dollars` appeared in the dictionary, so we implemented a further correction to collapse such expressions into the category `$DDD`. This was useful to the extent of not underestimating the frequencies of occurrence of numeric expressions, which will turn out to be an indicator of literary style. An example of the importance in retrieving a correct dictionary is to consider the `CS` set of rules that collapsed digits into `DDD`, `$DDD` and `DD:DD`. Looking at the speeches written by `JMCC` we noticed that the 2-token `in + DDD` appeared among the most frequent words, whereas using a previous version of our parser its frequency would get divided among all the 2-tokens `in + 'number in digits'`. It is evident that the importance of a dictionary that is somehow sensible to the problem does not have to be underestimated. A bad dictionary could flaw the analysis from the very beginning. The previous example brought up an issue that we also mentioned when discussing the correction of typos; when we counted words we would underestimate the frequency of occurrence for some words. One source of error was given by words that could be written in two variations, as we mentioned above, since we preferred to stick to the original version of the texts rather than aiming at consistency, i.e. `labeling` and `labelling`. Other sources of potential inaccuracies were plurals, and words with different endings, like verbs. The nature of these errors gave us further reasons to focus on function words, that is high-frequency words that do not have any particularly strong connection to the content of the speech.

Just a note to say that our code was a simple implementation, but one that would very much fit our purposes. In general there are three phases that texts may go through before qualifying for input of a classifier: (1) `parsing` or tokenizing is the first step where we remove punctuation, or if we want to include it we separate it from other tokens by spaces, by means of a `text parser` such as the one we implemented; (2) `stemming`, usually with Porter method [26], is the second step where we collapse words with the same stem into a concept class — i.e. `go goes` into `go` or `policy police` into `polic` — with or without paying attention to an actual common meaning, by means of a `text stemmer`, which we decided to disregard as not relevant to our problem where function words are our core asset; (3) `text tagging` is the last step where we classify each `token` or `concept` into a categories such as name, adjective and so on, by means of a `text tagger` which has to be trained itself on a large corpus of words, such as the British National Corpus, and we will attempt this method as well, later on in this section.

Finally it is very interesting to notice that the words we call `function words` in this report are referred to as `stop words` in the Computer Science literature, and the default behavior of many text parsers would be to **skip** them because not likely to be helpful to classify documents **by subject**. Yes, the goal of most of the automatic text classifiers is different as well. Computer Scientists are more concerned with the automatic classification by topic of the texts[27], a different problem where capturing the best words to summarize the content of a discussion group is the key. This is the reason why we do not need a text stemmer, but we do keep function words.

[26] The process of parsing the texts to produce space-separated strings of one or more words is sometimes called tokenization.

[27] For example, Internet newsgroup postings.

The Parser

The algorithm below simply aims to split a line into words, where each word is ideally delimited by spaces or tabs. Particular attention must be devoted to `'` the single apostrophe, and to characters that might be meaningful to keep as part of a word in the broad sense, e.g. `10:30` `70's` `$3,000` `0.453` `agents' f-14` or `hit-and-run`. Briefly:

```
Parse_Speech
   skip first line
   loop lines that don't contain QQQ, 'thank for listening' or 'by ronald reagan'
      remove end of line \n
      add spaces to beginning and end of line
      set all lowercase
      substitute punctuation " ( ) ; ! ? / .. with space

      remove dashes, but keep:      back-to-back or 1-in-14 or ...
      remove dots, but keep:        u.s.a. as u.s.a or 0.453 or ...
      remove commas, but keep:      $1,600 or ...
      remove colon, but keep:       10:30 or ...
      remove apostrophe, but keep:  agents' but not that' or 'if
                                    compact pattern like '70's into '70s

      split the line into words using space/tab as delimiters
```

`Parse_Speech` compacted the expressions `'60s` `'60's` `60s` into `'60s` , threw away the character `/` unless within a date, and the character `&` unless within two words and so on, in order to return a consistent vocabulary. e.g. `3/5/1976` and `a&m` would be counted as separate words, but the characters `/ &` in the expressions `and/or` or `The Full Employment & Balanced Growth Act` would be discarded.

Appendix D: Pools of Words

In this appendix we report some details of the process of word selection, along with the actual lists of marker words we found.

Hi-Frequency Words

During the exploratory data analysis (Reagan vs. other authors in general) we freely selected 267 hi-frequency function words among the 3000 most frequent words in the two dictionaries of Reagan and other authors, then we tested for differences in the usage of each of those words with a Kolmogorov-Smirnov two-sample non-parametric test, and came up with a list of 14 markers. We further refined the list by correcting the type I error for the multiple comparisons we performed, to end up with a list of 5 markers: `I may which our we`. Plus we found by manual inspection a compelling pair of low-frequency substitutes `over and over` vs. `continuous`, `continuously`.

We reproduced the same analysis by selecting about 270 hi-frequency words, among the fist 3000 highest frequency words in the dictionaries of Reagan and Hannaford. We used both a two-sample Kolmogorov-Smirnov test (KS test) to pick the words whose whole distribution of frequencies of occurrence $Y_{n\,i,j}$ would differ in the two sets of speeches, and the Welch approximate t-test (WT test) to check for differences in the average frequencies of occurrence. In both cases we corrected for multiple testing using the False Discovery Rate (FDR) correction proposed by Benjamini and Hochberg (1995) [2]. Below we present the discriminating words that we ended up with.

```
Kolmogorov-Smirnov 2-sample test + FDR correction
-------------------------------------------------------------------------
 [1] "may"  "our"  "we"  "now"                    alpha=0.05
 [1] "no"                                         alpha=0.06
 [1] "if"                                         alpha=0.0675

Welch approximate t-test + FDR correction
-------------------------------------------------------------------------
 [1] "we|ve"     "indeed"    "we|d"      "easy"      "i|ve"      "worth"
 [7] "they|ve"   "shouldn|t" "her"       "you|d"     "actual"    "suggest"
[13] "totally"   "exactly"   "were"      "we"        "south"     "fear"
[19] "that|s"    "may"(*)    "she"       "down"      "somewhere" "our"
[25] "someone"   "worldwide" "deep"      "then"      "myself"    "he|ll"
[31] "why"       "they"      "evidently" "was"       "us"        "now"
[37] "wasn|t"    "must"      "openly"    "no"        "wherever"  "awhile"
[43] "foolish"   "sure"      "bother"    "we|re"     "those"     "than"
[49] "only"      "something" "half"      "i|m"       "there"     "each"

Note: (*) are PH markers, RR the rest
```

We can see that `may our we` are good discriminators also in comparing Reagan versus Hannaford. In the EDA comparison we had 39 texts for other authors only 12 of which had been authors by Hannaford. In the new analysis we included 26 more news column by Hannaford, published with Reagan signature; the final analysis will contain 30.

SMART Stop Words

A second set of words contained 523 so called `stop words`, used in the computer science text and information retrieval communities, which expands the list of 363 words in Miller, Newman and Friedman (1958) [20] used by Mosteller and Wallace (1964, 1984) [24] [25]. The words that appear in this list would be excluded from the dictionary in a typical automated classification of e-texts into newsgroups as *too common* and not able to capture information about the context. We freely removed all the single letters present in the list, but `a d i s t`. Some of them may be an indication of the differential usage of abbreviated forms (for example `s t d`) but we did not believe they had real discriminating power in analysis of the authorship, as they suffered of the same problem different spellings `theatre/theater` did, not constituting a case for discrimination (see 23 above). Again performing the KS test and the WT test we ended up with the following words.

```
Kolmogorov-Smirnov 2-sample test + FDR correction
--------------------------------------------------------------------------
 RR: 4
 [1] "may"  "our"  "we"  "now"                    alpha=0.05

Welch approximate t-test + FDR correction
--------------------------------------------------------------------------
 [1] "indeed"        "tried"         "therefore"     "her"
 [5] "happens"       "following"     "unfortunately" "exactly"
 [9] "were"          "we"            "oh"            "may"(*)
[13] "she"           "down"          "d"             "somewhere"
[17] "becoming"      "five"          "latter"        "our"
[21] "entirely"      "provides"      "placed"        "someone"
[25] "indicated"     "went"          "then"          "uses"
[29] "myself"        "appropriate"   "why"           "they"
[33] "was"           "indicates"     "fifth"         "etc"
[37] "us"            "now"           "thank"         "somewhat"
[41] "indicate"      "know"          "must"          "no"
[45] "besides"       "anyway"        "wherever"      "doing"
[49] "follows"       "allows"        "thorough"      "elsewhere"
[53] "sure"          "sometime"      "those"         "than"
[57] "only"          "something"     "thereby"       "six"
[61] "s"(*)          "took"

Note: (*) are PH markers, RR the rest
```

A Semantic Approach using DOCU-SCOPE

A possibly independent set of features came from `DocuScope` software, an application courtesy of Pantelis Vlachos (Statistics Department) and David S. Kaufer (English Department) which tagged the texts according to several semantic dimensions. At the coarser level of abstraction we could distinguish parts of the speeches as `Inner Thought`, `Relations` or `Description`; at a finer level of details we could find `First Person`, `Inner Thinking`, `Think Positive`, `Think Negative`, `Think Ahead`, `Think Back` as a breakdown for the first, `Reasoning`, `Share Society Ties`, `Direct Activity`, `Interacting`, `Notifying`, `Linear Guidance` as a breakdown for the second, and `Word Picture`, `Space Interval`, `Motion`, `Past Events`, `Time Interval`, `Shifting Events` as a breakdown for the last. There was one more level of details which we did not consider since the 140+ third level features were created by David S. Kaufer somewhat subjectively, whereas the 18 second level features were built with robustness and normality in mind as simple sums of third level features. KS and WT tests yielded the following results:

```
Kolmogorov-Smirnov 2-sample test + FDR correction
------------------------------------------------------------------------------
[1] "ThinkAhead"     "ThinkPositive"  "ThinkBack"     alpha = 0.05
[1] "PastEvents"                                      alpha = 0.0575
[1] "LinearGuidance"                                  alpha = 0.08

Welch approximate t-test + FDR correction
------------------------------------------------------------------------------
None
```

The features values would represent the number of times each semantic indicator is present in the text, normalized to represent the counts in text of 100 words. These features were not exactly normal; further as we looked at the 18 second-level features we observed many zero counts that detracted to the normality assumption, and pushed towards a mixture. Despite these departure from normality in the sample, these features were built to be robust to departure from normality on a wide variety of texts, and a generalization of the results based upon them may be more believable. Thus we wanted to explore the information they carried a little more using the jackknife, as a the first step of a technique proposed by Mosteller and Tukey (1968) [23]. The focus in this section is on feature selection, though, more about the full procedure and its implications in terms of predictive power will be examined in section 4.3.

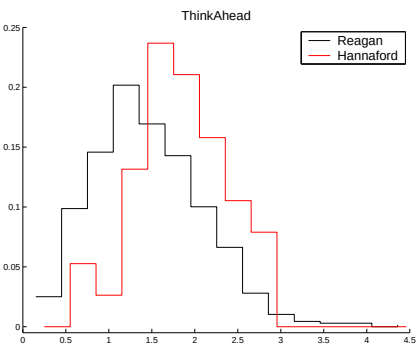

Figure 4: Think Ahead.

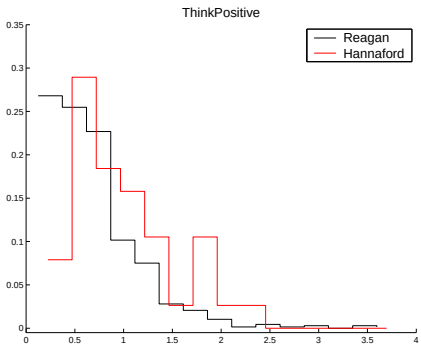

Figure 5: Think Positive.

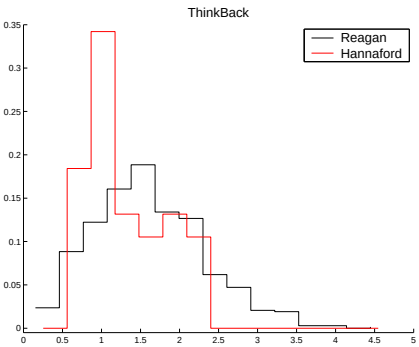

Figure 6: Think Back.

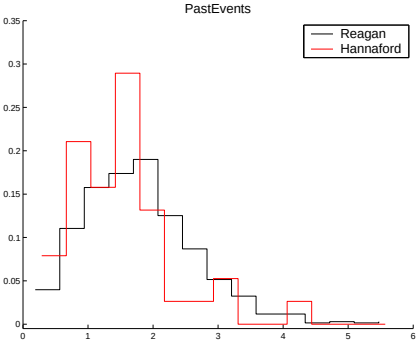

Figure 7: Past Events.

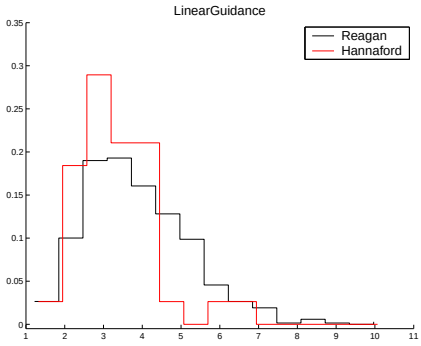

Figure 8: Linear Guidance.

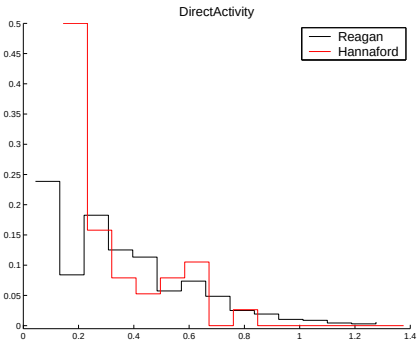

Figure 9: Direct Activity.

Mosteller and Tukey used a linear discriminant function as classifier, and their idea was to use a nested procedure to estimate the variability of the coefficients and the predictive power of the classifier on separate data. We were interested in the variability of the coefficients, so we did not need a nested procedure. We divided the data in 38 batches, so that each batch would contain

one speech authored by Hannaford and 18 by Reagan. The jackknife consisted in first estimating the coefficients β_{all} using all the data, and the coefficients $\beta_{(i)}$ removing the i^{th} batch of texts for $i = 1, ..., 38$ from the data, then we adjusted the coefficients of each batch as in $\beta_{*i} = 38\,\beta_{all} - 37\,\beta_{(i)}$, and finally we computed the jackknifed coefficients β_* by averaging the adjusted coefficients β_*.

Preliminary runs showed that none of the averaged coefficients was significantly different from zero. We tried both Fished LDA and variation where we adjusted the variance-covariance matrices involved in the procedure for the fact that the batches were not exactly random, as we fixed the proportions of Hannaford to Reagan texts in each batch to be constant, with no difference.

Finally we looked at the distributions of the jackknifed coefficients to look for some indications of style not carried by means and variances. In each batch there is only one document authored by Hannaford and 18 by Reagan, and out of 18 features three of them, `Think Positive`, `Think Back` and `Direct Activity` had more than 75% of the jackknifed coefficients with the same sign.

	Summary statistics for β_*				
Features	Min	1^{st} Qu.	Median	3^{rd} Qu.	Max
ThinkPositive	-2.350	-1.122	-0.565	-0.259	1.307
ThinkBack	-1.387	0.153	0.484	0.883	1.966
Direct Activity	-1.078	0.001	0.289	0.828	1.651

Table 23: Summary statistics for `Think Positive`, `Think Back` and `Direct Activity`.

Two-stage Selection on all 4-Grams

Using the new set of CS rules we started from scratch and looked for markers to distinguish Reagan's style from Hannaford's one. We retrieved all 4-grams[28] to look for expressions as well as words, following the intuition that expressions are much more personal, and enough to discriminate between authors. For the manual screening we made use of the statistic T_1 in equation 3.2 along with the thresholds derived in section 3.2.2. We used a cut-off value for T_1 of 3.85, derived along the lines suggested by Mosteller and Wallace (1964, 1984) [24] [25]. We corrected for the difference in the number of speeches, adjusting the ratios as if we had 38 speeches for Reagan as well[29] as for Hannaford as explained in section 3.2.3.

The first stage consisted of filtering all the words in the dictionaries using two groups of documents. We split the speeches into two groups, the first contained speeches broadcasted over the years 1975-77 and the second contained the speeches broadcasted over the years 1978-79. For each word we made use of T_1 with the counts obtained from the speeches in the first group, and if the words qualified we then re-filtered it using the counts in the speeches of the second group. We did not corrected for multiple testing at this stage, this is just a first-stage filtering were we considered the overall frequencies of occurrence, and we wanted to cut down the number of words and 2-3-4-grams from the hundreds of thousands to a manageable number, while getting a feel for which kind of words were likely to be used at different rates. The purpose of the two groups was to mitigate the selection effect.

[28] Note that in a text of ℓ words there are $\ell - k + 1$ k-grams, that is $\ell - k + 1$ overlapping sequences of k adjacent words. i.e. the phrase `I go to the doctor` contains the $(5 - 3 + 1 = 2)$ 3-grams `I-go-to`, `go-to-the`, and `to-the-doctor`.

[29] This is a delicate issue. The most sensible thing to do was to act as if we had 38 Reagan texts. Doing so meant to summarize the speeches authored by Reagan, which sounded better than augmenting the speeches by Peter Hannaford, thus making implicit inferences.

The second stage consisted in using the usual KS and WT tests and the FDR correction for multiple testing on all the words that passed through the first stage.

— Stage One —

We did a visual exploration of whether the words that passed the first stage did have some discriminating power or not, since we did not correct for multiple tests. We considered the distribution of Reagan markers vs. Hannaford markers on all the speeches in the years 1978-79 (the second group contained 372 texts for Reagan and 18 for Hannaford). Below we present the plots of the density estimates for the quantities $\left(\frac{\sum_j X_{n\, i_1\, j}}{372} - \frac{\sum_j X_{n\, i_2\, j}}{18}\right)$. These quantities should be about zero if the markers had no discriminative power. Notice that the selection effect should not be too bad, since we filtered the words in two waves, and we were computing the quantity above only for the speeches in the second wave. We used bootstrap to estimate the variability of the distributions of the markers, and plotted the median density estimate.

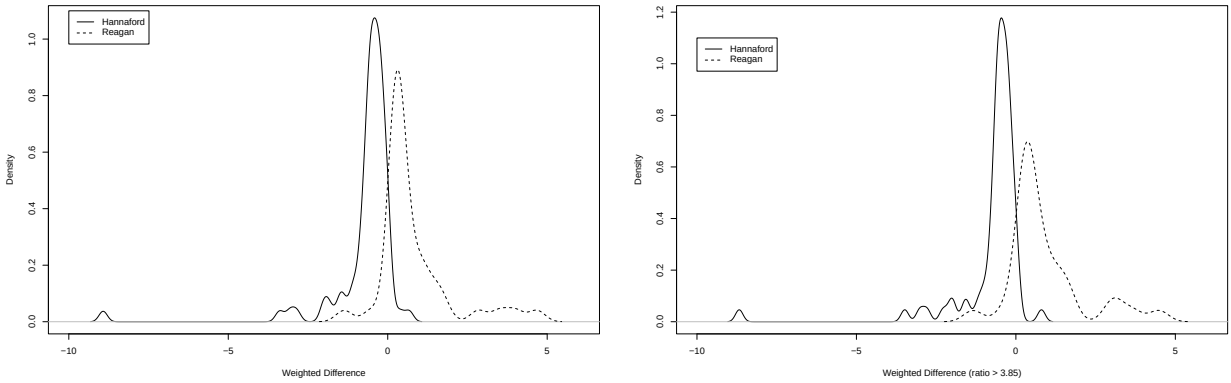

Figure 10: Median density estimates (out of 1000 bootstrap samples) to assess the discriminative power of the markers on the speeches 1978-79. Left panel 121 markers with $ratio > 3.6$, right panel 103 markers with $ratio > 3.85$.

In this first stage, 103 words made the list. We can divide them into 31 (+8) Reagan markers and 72 (+10) Hannaford markers.

```
### Hannaford markers (T1 > 3.85)
 [1] "the"                  "carter"                 "to"
 [4] "its"                  "may"                    "would"
 [7] "it"                   "will"                   "are"
[10] "mr|carter"            "the|carter"             "on"
[13] "strong"               "nearly"                 "after"
[16] "measure"              "not"                    "new"
[19] "that|the"             "president|carter"       "issue"
[22] "jokes"                "special"                "problems"
[25] "current"              "in|the"                 "joke"
[28] "carter|s"             "west"                   "that"
[31] "california"           "carter|administration"  "if"
[34] "it|would"             "sharp"                  "the|issue"
[37] "after|the"            "go"                     "a|big"
[40] "to|its"               "last|week"              "measures"
[43] "if|the"               "large"                  "recent"
[46] "may|be"               "basic"                  "seems"
[49] "other"                "he|would"               "the|other"
```

```
[52] "till"                 "ones"                  "future"
[55] "is"                   "put|it"                "your"
[58] "through|the"          "as|if"                 "no|wonder"
[61] "under"                "seems|to"              "over"
[64] "thus"                 "that|it"               "most"
[67] "the|joke"             "plenty|of"             "role"
[70] "huge"                 "big"                   "with"

### Hannaford markers (3.65 < T1 < 3.85)
 [1] "you|go|to" "as"        "plenty"    "though"    "on|the"    "you|go"
 [7] "this"      "chances"   "it|out"    "up|in"

### Reagan markers  (T1 > 3.85)
 [1] "we"          "our"          "her"          "they|were"   "in|our"
 [6] "she"         "were"         "we|ve"        "ve"          "of"
[11] "i|ve"        "of|our"       "in"           "was"         "free"
[16] "than|DDD"    "DDD"          "total"        "$DDD"        "of|all"
[21] "percent"     "for|DDD"      "about|DDD"    "$DDD|a"      "entire"
[26] "worth"       "DDD|percent" "and|so"      "and|he"      "and|the"
[31] "indeed"

### Reagan markers  (3.65 < T1 < 3.85)
 [1] "if|we"         "power"         "we|d"          "DDD|to"
 [5] "in|all"        "we|had"        "our|own"       "more|than|DDD"
```

Looking at the markers, we can attempt a free interpretation. Reagan liked to enrich his arguments with numbers; numbers that the typical worker and his wife could understand. As we shall also see below, he often liked to tell the impact of policies or new and old laws on simple citizens' incomes to make the issues he talked about more *touching*, or just to make facts more *concrete* by adding some practical effect that his listeners could understand. He was a great communicator, and he would not keep a distance. So all the markers like `we`, `our`, `of+our`, `in+our`, `our+own`, and the other like `DDD`, `$DDD`, `for+DDD`, `about+DDD`, `$DDD+a`, `more+than+DDD`, `percent`, `DDD+percent`, `worth`, fit into the picture. It is interesting that he also used more often than Hannaford adjectives and hi-frequency words like `free`, `entire`, `indeed`, `of`, `in`, `and+so`, `of+all`, `in+all`, `if+we`, and addressed subtle critics to Carter's doings by explaining more often what he had done `i+ve` than telling what Carter did not do.

As far as Hannaford's markers are concerned, two things popped up; first he attacked Carter more directly as in `carter`, `carter+s`, `the+carter`, `mr+carter`, `president+carter`, and second he kept a distance as in `your`, `you+go`, `you+go+to`. Hannaford used more frequently adjectives like `sharp`, `big`, `a+big`, `plenty`, `plenty+of`, `large`, `huge`, he used more often both `may` and `will`, he referred to Reagan experience in California explicitly as in `west` or `california`, he talked about `problems`, `issue`, `the+issue` and about `measure`, `measures`, `chances` and he referred to events in time as in `current`, `recent`, `last+week`, `future`. Hannaford also made a different use of higher-frequency words like `the`, `on`, `thus`, `with`, `to`, `its`, `till`[30], `may+be`, `that+the`, `in+the`, `put+it`, `after+the`, `other`, `as+if`, and some other words like `new`, `basic`, `nearly`, `special`, `most`, `over`, `under`, `no+wonder`, and so on.

Some of these expressions are indeed colloquial, but if we were willing to ignore the FDR correction for multiple testing in this first-stage screening, as maybe too severe, then what we would

[30] Even as `until` was used more often by Reagan it did not qualify as a marker since, over all the papers, we could observe a corrected ratio of $\frac{(38/679*102-7)^2}{(38/679*102+7)} \approx 0.13$. `till` qualified as a marker with a ratio $\frac{(38/679*22-8)^2}{(38/679*22+8)} \approx 4.96$.

find would suggest that Hannaford used his own favorite informal expressions, but not Reagan's favorite ones.

There are two more lists that we want to present. The first one is a list of words that popped up as we loosed the cut-off to 2.72 in the two-wave filtering procedure above, which would correspond to simulated p-values less than 13%.

```
### Hannaford markers  (2.72 < T1 < 3.65)
 [1] "i|have"         "to|get"         "just|how"       "when|it|comes"
 [5] "to|the"         "meanwhile"      "may|have"       "since|it"
 [9] "since"          "bad"            "seems|to|be"    "the|other|day"
[13] "sharply"        "despite|the"    "closely"        "a|few|of"
[17] "everywhere"     "through"        "with|the"       "recent|years"
[21] "about|to"       "after|all"      "scarcely"       "seem|to"
[25] "in|california" "all|but"        "up|to|DDD"      "soon"
[29] "this|year"

### Reagan markers  (2.72 < T1 < 3.65)
 [1] "us"           "more|than"    "i|m"          "this|was"     "all|of"
 [6] "then"         "great"        "matter|of"    "now"          "few|years"
[11] "supposed|to" "times|as"     "DDD|billion" "easy"         "therefore"
[16] "any|of"       "how|many"     "and"          "why"
```

The second list was obtained by some artificial data augmentation, specifically by comparing the counts we would obtain in a 100 *reference papers*. We present it only because many of the words in there may lead to very compelling interpretations. We did not use these words in the classification, though, since there was not strong enough evidence in the data we observed to include them.

```
### Hannaford markers  (T1 > 3.85, not in the lists before)
 [1] "the|notion"         "notion"             "notion|that"
 [4] "especially"         "i|have"             "on|the"
 [7] "to|the"             "closely"            "sharply"
[10] "a|few|of"           "a|strong"           "about|to"
[13] "time|as"            "scarcely"           "same|ones"
[16] "jokes|about"        "may|be|a"           "do|it"
[19] "up|to|DDD"          "but|in|my"          "to|prove|it"
[22] "argue|that"         "my"                 "soon"
[25] "note"               "will|not"           "this|year"
[28] "ways"               "no|doubt"           "it|may"
[31] "few|of"             "sense|of"           "for|them|to"
[34] "just|the|same"      "is|one|of"          "got"
[37] "as|the"             "most|of"            "fact"
[40] "between|the"        "while|the"          "says|that|the"
[43] "rather"             "in|fact"            "so|long|as"
[46] "on|the|average"     "end|up"             "suggests|that"
[49] "in|the|case|of"     "on|the|other|hand" "at"
[52] "it|is|hard|to"      "boost"              "as|a|result"
[55] "on|behalf|of"       "it|comes|to"        "pride"
[58] "more|not|less"      "a|lot|to|do"        "these|jokes"
[61] "more|and|more"      "back|and|forth"     "put|it|this|way"
[64] "a|way|of"           "how|do|you"

### Reagan markers  (T1 > 3.85, not in the lists before)
 [1] "remember"       "in|all"         "and|of"         "DDD|billion"
 [5] "picture"        "therefore"      "any|of"         "how|many"
 [9] "must|be"        "in|an"          "regard"         "named"
[13] "and"            "why"            "equal"          "is|it"
[17] "god"            "for|every"      "a|few|years"    "regard|to"
```

```
[21] "is|being"       "use|of"         "told|of"        "in|DDD|the"
[25] "DDD|to|DDD"     "allowed|to"     "trouble"        "and|of|course"
[29] "and|all"        "sure"           "tried|to"       "out|of|the"
[33] "made|it"        "no"             "the|entire"     "so|forth"
[37] "a|matter"       "a|matter|of"    "and|our"        "with|regard|to"
[41] "and|so|forth"   "let|me"         "the|best"       "and|if"
[45] "i|d|like|to"    "aware"          "too|long|ago"   "too|long"
[49] "unfortunately"  "planned"        "easy|to"        "supposed"
[53] "i|don|t|know"   "some|time"      "half|the"       "truth|is"
[57] "in|spite|of"    "pretty"         "must|have"      "but|then"
[61] "pay"            "exactly"        "someone"
```

As we lowered the cut-off the discriminative power of the words we found decreased slightly, as the figure below testifies. Note that the ranges are much narrower than before, passing from about $[-10, 8]$ to $[-2, 3]$ and $[-1, 1]$ respectively.

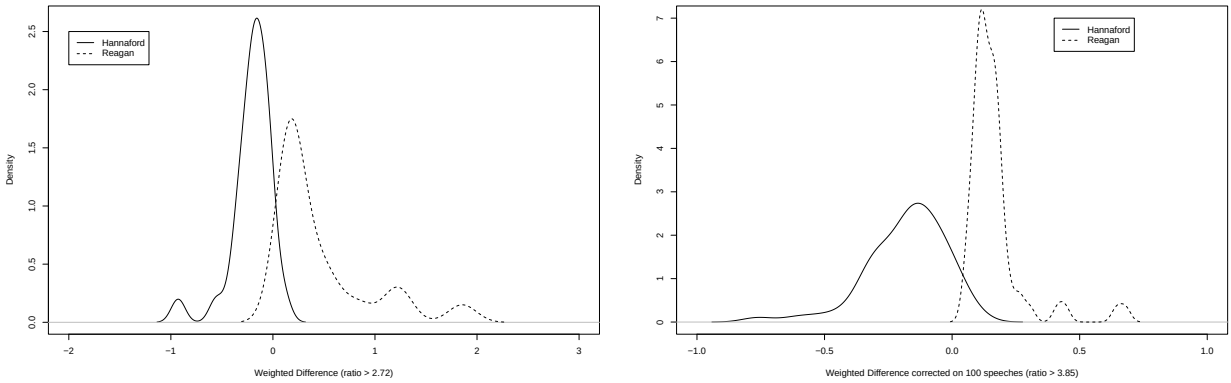

Figure 11: Median density estimates (out of 1000 bootstrap samples) to assess the discriminative power of the markers on the speeches 1978-79. Left panel 48 markers with $ratio > 2.72$, right panel 128 markers with $ratio > 3.85$ applied to *projected* counts on 100 speeches.

— Stage Two —

In the second stage, we implemented formal testing for differences in the overall and average use of words, corrected for multiple testing via FDR. We present below the words that made the final list.

```
Kolmogorov-Smirnov 2-sample test + FDR correction
-------------------------------------------------------------------------
T1 > 3.85
RR: "we"  "our"  "were"                (at FDR alpha=0.1 add: "ve" "DDD")
PH: "carter"  "may"  "if"

T1 > 2.72
RR: "now"

Welch approximate t-test + FDR correction
-------------------------------------------------------------------------
T1 > 3.85
RR:
```

```
 [1] "we"          "our"         "her"         "they|were"   "in|our"
 [6] "she"         "were"        "we|ve"       "ve"          "i|ve"
[11] "of|our"      "was"         "free"        "than|DDD"    "DDD"
[16] "total"       "$DDD"        "of|all"      "percent"     "for|DDD"
[21] "about|DDD"   "$DDD|a"      "entire"      "worth"       "DDD|percent"
[26] "and|so"      "and|he"      "indeed"
 PH:
 [1] "carter"      "its"         "may"         "mr|carter"   "the|carter"
 [6] "nearly"      "president|carter"          "current"     "carter|s"
[10] "carter|administration"     "after|the"   "last|week"   "large"

 T1 > 3.65
 RR:
 [1] "if|we"         "power"         "we|d"          "DDD|to"
 [5] "in|all"        "we|had"        "our|own"       "more|than|DDD"

 T1 > 2.72
 RR:
 [1] "us"          "more|than"   "this|was"    "all|of"      "then"
 [6] "matter|of"   "now"         "few|years"   "supposed|to" "times|as"
[11] "DDD|billion" "easy"        "therefore"   "any|of"      "how|many"
[16] "why"
```

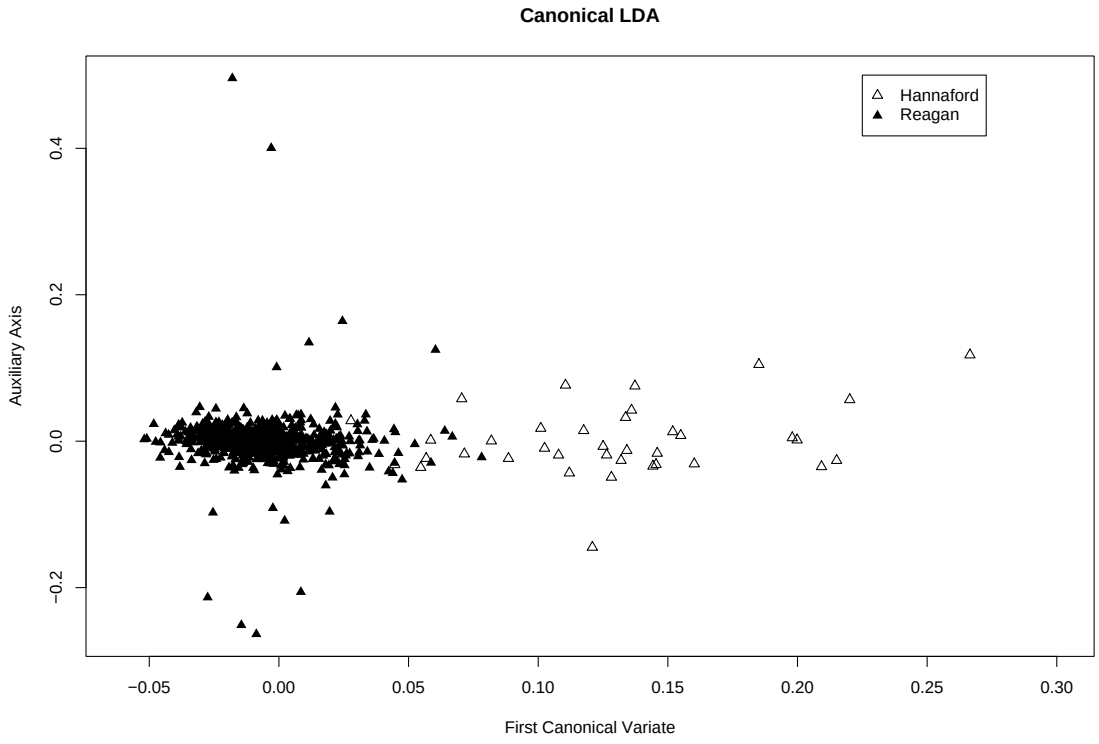

Figure 12: Separation obtained with Fisher LDA, the vertical axis is auxiliary since with two groups there is only one linear discriminant function. Train and test on all texts using all words.

Fisher LDA (by training and testing on all the speeches using all the 103 words whose $ratio >$

3.85) achieved an accuracy of 99.85% on Reagan texts and of 84.21% on Hannaford texts. This was an encouraging result since Fisher LDA does not make distributional assumptions on the joint p.d.f. of the words, but simply projects data on the *best separating* direction. If we assume multivariate normality and use the probabilistic version of LDA due to Rao (1948) [27] we could specify equal prior beliefs on the authorship of each text to achieve an accuracy of 99.41% on Reagan texts of 84.21% on Hannaford texts. Multivariate normality did not help. We did not attempt transformations to stabilize the variance, or to make the data more *Normal*, so we cannot exclude at this point that some improvements in the accuracy could be achieved.

Information Gain

A last attempt to collect good discriminating words made use of the information gain.

Hannaford		Hannaford, McClaughry	Other authors
Bernoulli	Multinomial	Multinomial	Multinomial
carter	may	we	carter
our	current	our	we
may	notion	may	our
her	nearly	her	her
its	message	small	may
jokes	strong	its	ve
joke	worries	ve	she
ethnic	ironically	she	its
was	scarcely	would	us
strong	despite	jokes	scarcely
week	our	joke	sharp
nearly	ve	negative	despite
were	sharp	was	message
ve	week	alternative	strong
measure	problems	you	notion
sharp	assumption	strong	sharply
she	plenty	week	model
current	measures	us	which
notion	efficiency	self	large
despite	other		spirit
measures	ones		continuous
special	large		self
will	huge		total
they	till		alternative
percent	future		nearly
would	sharply		thus
efficiency	contradictions		me
	context		
	devastating		
	predictably		
	cheap		

Table 24: Word selected via Information Gain.

Briefly, any word yields a partition into documents that contain it and documents that do not. The information gain of a word measures how much that word is able to separate the two authors, in terms distance between the distributions of the class labels before and after the split. In order to compute the information gain for a word we used two model that have been proposed in the recent literature: multivariate Bernoulli, and multinomial. The multivariate Bernoulli model takes

into account the presence or absence of a word in a speech, whereas the multinomial model takes into account how many times a word appeared in a speech.

References

[1] D. Beeferman, A. Berger, and J. Lafferty. A model of lexical attraction and repulsion. In P. R. Cohen and W. Wahlster, editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 373–380. Association for Computational Linguistics, 1997.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995.

[3] J. Binongo. Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, 16:18–25, 2003.

[4] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] J.F. Burrows. Not unless you ask it nicely: The interpretative nexus between analysis and information. *Literary and Lingustic Computing*, 7:91–109, 1992.

[6] J. Collins and D.F. Kaufer. Docu-Scope: A Java application for statistical literary style modeling. Technical report, Carnegie Mellon University, 2001.

[7] A.C. Davison and D.V. Hinkley. *Bootstrap Methods and Their Application.* Cambridge University Press, 1999.

[8] R. Dawes and B. Corrigan. A unit weighted model for discriminating the Federalist Papers. Technical Report 5, Oregon Research Institute, 1976.

[9] B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78:316–331, 1983.

[10] R.A. Fisher. The use of multiple treatments in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[11] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer-Verlag, 2001.

[12] D. Holmes. Stylometry and the Civil war: the case of the Pickett letters. *Chance*, 16:9–17, 2003.

[13] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In C. Nédellec and C. Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 137–142. Springer-Verlag, 1998.

[14] N.L. Johnson, S. Kotz, and A.W. Kemp. *Univariate Discrete Distributions.* John Wiley, 1992.

[15] I.T. Jolliffe. *Principal Component Analysis.* Springer-Verlag, 1986.

[16] J. Kleinberg. Two algorithms for nearest-neighbor search in high dimensions. Unpublished manuscript. http://simon.cs.cornell.edu/home/kleinber/kleinber.html, 1997.

[17] A. McCallum. Bag Of Words: An Ansi C library for statistical language modeling, text retrieval, classification and clustering. http://www-2.cs.cmu.edu/∼mccallum/bow/, 1996.

[18] T.C. Mendenhall. The characteristic curves of composition. *Science*, 11:237–249, 1887.

[19] G.A. Miller. Communication. In C.P. Stone, editor, *Annual Review of Psychology*, pages 137–142. Banta Publishing Company, 1954.

[20] G.A. Miller, E.B. Newman, and E.A. Friedman. Length-frequency statistics for written English. *Information and Control*, 1:370–389, 1958.

[21] T. Mitchell. *Machine Learning.* McGrow-Hill, 1997.

[22] T. Mitchell (Director). Text Learning Group. http://www-2.cs.cmu.edu/∼TextLearning/. Carnegie Mellon University.

[23] F. Mosteller and J.W. Tukey. Data analysis including statistics. In G. Lindsey, editor, *Handbook of Social Psychology*, volume 1, pages 80–203. Knopf, 1968.

[24] F. Mosteller and D.L. Wallace. *Inference and Disputed Authorship: The Federalist.* Addison-Wesley, 1964.

[25] F. Mosteller and D.L. Wallace. *Applied Bayesian and Classical Inference: The Case of The Federalist Papers.* Springer-Verlag, 1984.

[26] M.F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[27] C.R. Rao. The utilization of multiple measurements in problems of biological classification (with discussion). *Journal of the Royal Statistical Society, Series B*, 10:159–203, 1948.

[28] B.D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, 1996.

[29] K. Skinner, A. Anderson, and M. Anderson, eds. *Reagan, in His Own Hand: The Writings of Ronald Reagan that Reveal his Revolutionary Vision for America.* Free Press, 2001.

[30] K. Skinner, A. Anderson, and M. Anderson, eds. *Stories in His Own Hand: The Everyday Wisdom of Ronald Reagan.* Free Press, 2001.

[31] B.L. Welch. The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29:350–362, 1938.

[32] U. Yule. *The Statistical Study of Literary Vocabulary.* Cambridge University Press, 1944.

[33] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Research and Development in Information Retrieval*, pages 334–342, 2001.

[34] G.K. Zipf. *Selected Studies of the Principle of Relative Frequency in Language.* Harvard University Press, 1932.