Universitat Politècnica de Catalunya
Department of Statistics and Operations Research

Phd thesis

Bayesian Analysis of Textual Data

Author: Martí Font Valverde
Advisors: Josep Ginebra i Molins and Xavier Puig i Oriol

November 2015

Contents

Summary

In this thesis I develop statistical methodology for analyzing discrete data to be applied to stylometry problems, always with the Bayesian approach in mind. The statistical analysis of literary style has long been used to characterize the style of texts and authors, and to help settle authorship attribution problems. Early work in the literature used word length, sentence length, and proportion of nouns, articles, adjectives or adverbs to characterize literary style. I use count data that goes from the frequency of word frequency, to the simultaneous analysis of word length counts and more frequent function words counts. All of them are characteristic features of the style of author and at the same time rather independent of the context in which he writes.

Here we intrude a Bayesian Analysis of word frequency counts, that have a reverse J-shaped distribution with extraordinarily long upper tails. It is based on extending Sichel's non-Bayesian methodology for frequency count data using the inverse gaussian Poisson model. The model is checked by exploring the posterior distribution of the Pearson errors and by implementing posterior predictive consistency checks. The posterior distribution of the inverse gaussian mixing density also provides a useful interpretation, because it can be seen as an estimate of the vocabulary distribution of the author, from which measures of richness and of diversity of the author's writing can be obtained. An alternative analysis is proposed based on the inverse gaussian-zero truncated Poisson mixture model, which is obtained by switching the order of the mixing and the truncation stages.

An analysis of the heterogeneity of the style of a text is proposed that strikes a compromise between change-point, that analyze sudden changes in style, and cluster analysis, that does not take order into consideration. Here an analysis is proposed that strikes a compromise by incorporating the fact that parts of the text that are close together are more likely to belong to the same author than parts of the text far apart. The approach is illustrated by revisiting the authorship attribution of Tirant lo Blanc.

A statistical analysis of the heterogeneity of literary style in a set of texts that simultaneously uses different stylometric characteristics, like word length and the frequency of function words, is proposed. It clusters the rows of all contingency tables simultaneously into groups with homogeneous style based on a finite mixture of sets of multinomial models. That has some advantages over the usual heuristic cluster analysis approaches

as it naturally incorporates the text size, the discrete nature of the data, and the dependence between categories. All is illustrated with the analysis of the style in plays by Shakespeare, El Quijote, and Tirant lo Blanc.

Finally, authorship attribution and verification problems that are usually treated separately are treated jointly. That is done by assuming an open-set classification framework for attribution problems, contemplating the possibility that neither one of the candidate authors, with training texts known to have been written by them is the author of the disputed texts. Then the verification problem becomes a special case of attribution problems.A formal Bayesian multinomial model for this more general authorship attribution is given and a closed form solution for it is derived. The approach to the verification problem is illustrated by exploring whether a court ruling sentence could have been written by the judge that signs it or not, and the approach to the attribution problem is illustrated by revisiting the authority attribution of the Federalist papers.

Resumen

En esta tesis se desarrolla, siempre con el enfoque bayesiano en mente, una metodología estadística para el análisis de datos discretos en su aplicación en problemas estilometría. El análisis estadístico del estilo literario se ha utilizado para caracterizar el estilo de textos y autores, y para ayudar a resolver problemas de atribución de autoría. Para caracterizar el estilo literario trabajos anteriores usaron la longitud de las palabras, la longitud de las oraciones, y la proporción de los sustantivos, artículos, adjetivos o adverbios. Los datos que aqu se utilizan van, desde la frecuencia de frecuencias de palabras, hasta el análisis simultáneo de frecuencias de longitud de palabra y de las palabras funcionales más frecuentes. Todos estos datos son característicos del estilo de autor y al mismo tiempo independiente del contexto en el que escribe.

De esta forma, se introduce un análisis bayesiano de la frecuencia de frecuencias de palabra, que tiene una distribución en forma de J inversa con las colas superiores extraordinariamente largas. Se basa en la extensión de la metodología no bayesiana de Sichel para estos datos utilizando el modelo Poisson inversa gaussiana. Los modelos se comprueban mediante la exploración de la distribución a posteriori de los errores de Pearson y por la implementación de controles de consistencia de la distribución predictiva a posteriori. La distribución a posteriori de la inversa gausiana tiene una interpretación útil, al poder ser vista como una estimación de la distribución vocabulario del autor, de la cual se pueden obtener la riqueza y diversidad de la escritura del autor. Se propone también un análisis alternativo basado en la mixtura inversa gaussiana - poisson truncada en el cero, que se obtiene cambiando el orden de la mezcla y truncamiento.

También se propone un análisis de la heterogeneidad de estilo, que es un compromiso entre el modelo de punto de cambio, que busca un cambio repentino de estilo, y el análisi de conglomerados, que no tiene en cuenta el orden. Aquí se propone un análisis que incorpora el hecho de que partes prximas de un texto tienen más probabilidades de pertenecer al mismo autor que partes del texto ms separadas. El enfoque se ilustra volviendo a revisar la atribución de autoría del Tirant lo Blanc.

Para el análisis de la heterogeneidad del estilo literario, se propone también un análisis estadístico que utiliza simultáneamente diferentes características estilométricas, como la longitud palabra y la frecuencia de las palabras funcionales más frecuentes. Las filas de todas tablas de contingencia se agrupan simultáneamente basandose en una mezcla finita

de conjuntos de modelos multinomiales con un estilo homogéneo. Esto tiene algunas ventajas sobre las heurísticas utilizadas en el análisis de conglomerados, ya que incorpora naturalmente el tamao del texto, la naturaleza discreta de los datos y la dependencia entre categorías. Todo ello se ilustra a través del análisis del estilo en las obras de teatro de Shakespeare, el Quijote y el Tirant lo Blanc.

Finalmente, los problemas de atribución y verificación de autoría, que se tratan normalmente por separado, son tratados en forma conjunta. Esto se hace asumiendo un escenario abierto de clasificación para el problema de la atribución, contemplando la posibilidad de que ninguno de los autores candidatos, con textos conocidos para aprendijaje, es el autor de los textos en disputa. Entonces, el problema de verificación se convierte en un caso especial de problema de atribución. El modelo multinomial bayesiano propuesto permite obtener una solución exacta y cerrada para este problema de atribución de autoría más general. El enfoque al problema de verificación se ilustra mediante la exploración de si un fallo judicial condenatorio podría haber sido escrito por el juez que firma o no, y el enfoque del problema de la atribución se ilustra revisando el problema de la autoría de los Federalist Papers.

Chapter 1

Introduction

This thesis deals with methods for the analysis of discrete data in the context of the statistical analysis of literary style. The statistical analysis of literary style has long been used to characterize the style of texts and authors, and to help settle authorship attribution problems. Early work used word length and sentence length to characterize literary style. Other characteristics widely used for this purpose have been the proportion of nouns, articles, adjectives or adverbs, the frequency of use of function words, which are independent of the context. In Chapter 2 the frequencies of word frequency count is the one used in the analysis while in Chapters 3, 4 and 5 deal with with the analysis of data like word length counts and the frequency of function words.

Moreover, one can also characterize literary style by analyzing word frequency counts. Given that most words appear very few times and very few words are repeated many times, word frequency count data have reverse J-shaped distributions with extraordinarily long upper tails. In Chapter 2 word frequency counts are use as data in the analysis.

In Chapter 2, it is shown that the zero truncated inverse gaussian-Poisson model, obtained by first mixing the Poisson model assuming its expected value has an inverse gaussian distribution and then truncating the model at zero, is very useful when modeling frequency count data. A Bayesian analysis based on this statistical model is implemented on the word frequency counts of various texts, and its validity is checked by exploring the posterior distribution of the Pearson errors and by implementing posterior predictive consistency checks. The analysis based on this model is useful because it allows one to use the posterior distribution of the model mixing density as an approximation of the posterior distribution of the density of the word frequencies of the vocabulary of the

author, which is useful to characterize the style of that author. The posterior distribution of the expectation and of measures of the variability of that mixing distribution can be used to assess the size and diversity of his vocabulary. An alternative analysis is proposed based on the inverse gaussian-zero truncated Poisson mixture model, which is obtained by switching the order of the mixing and the truncation stages. Even though this second model fits some of the word frequency data sets more accurately than the first model, in practice the analysis based on it is not as useful because it does not allow one to estimate the word frequency distribution of the vocabulary.

In Chapter 3, one proposes a classification analysis of literary style that takes order into consideration. The statistical analysis of the heterogeneity of the style of a text often leads to the analysis of contingency tables of ordered rows. When multiple authorship is suspected, one can explore that heterogeneity through either a change-point analysis of these rows, consistent with sudden changes of author, or a cluster analysis of them, consistent with authors contributing exchangeably, without taking order into consideration. Here an analysis is proposed that strikes a compromise between change-point and cluster analysis by incorporating the fact that parts close together are more likely to belong to the same author than parts far apart. The approach is illustrated by revisiting the authorship attribution of *Tirant lo Blanc*.

In Chapter 4, one proposes a statistical analysis of the heterogeneity of literary style in a set of texts that simultaneously uses different stylometric characteristics, like word length and the frequency of function words. Data consist of several tables with the same number of rows, with the i-th row of all tables corresponding to the i-th text. The analysis proposed clusters the rows of all these tables simultaneously That has the advantage over the usual heuristic cluster analysis approaches that it naturally incorporates in the analysis the text size, the discrete nature of the data, and the dependence between categories. All this is illustrated through an analysis of the heterogeneity in the plays by Shakespeare and in *El Quijote*, and by revisiting again as in Chapter 3 the authorship-attribution of *Tirant lo Blanc*.

Finally, in Chapter 5, a unified approach to authorship attribution and verification problems is proposed. In authorship attribution problems one needs to assign a text or a set of texts from an unknown author to either one of two or more candidate authors on the basis of the comparison of the disputed texts with texts known to have been written by the candidate authors. In authorship verification problems one needs to decide whether a text or a set of texts could have been written by a given single author or not. These two problems are usually treated separately. By assuming an open-set classification framework for the attribution problem, contemplating the possibility that neither one of the candidate authors is the unknown author, the verification problem becomes a special

case of attribution problem. Here both problems are posed as a formal Bayesian multinomial model selection problem and are given a closed form solution. The approach to the verification problem is illustrated by exploring whether a court ruling sentence could have been written by the judge that signs it or not, and the approach to the attribution problem is illustrated by revisiting the authorship attribution of the Federalist papers.

Note that, Chapters 3, 4 and 5 deal with classification analysis techniques. In Chapters 3 and 4 the techniques are for unsupervised classification and in Chapter 5 they are for supervised classification.

Chapter 2

Bayesian Analysis of Frequency Count Data

The zero truncated inverse gaussian-Poisson model, obtained by first mixing the Poisson model assuming its expected value has an inverse gaussian distribution and then truncating the model at zero, is very useful when modelling frequency count data. A Bayesian analysis based on this statistical model is implemented on the word frequency counts of various texts, and its validity is checked by exploring the posterior distribution of the Pearson errors and by implementing posterior predictive consistency checks. The analysis based on this model is useful because it allows one to use the posterior distribution of the model mixing density as an approximation of the posterior distribution of the density of the word frequencies of the vocabulary of the author, which is useful to characterize the style of that author. The posterior distribution of the expectation and of measures of the variability of that mixing distribution can be used to assess the size and diversity of his vocabulary. An alternative analysis is proposed based on the inverse gaussian-zero truncated Poisson mixture model, which is obtained by switching the order of the mixing and the truncation stages. Even though this second model fits some of the word frequency data sets more accurately than the first model, in practice the analysis based on it is not as useful because it does not allow one to estimate the word frequency distribution of the vocabulary.

2.1 Introduction

To characterize literary style one often relies on the analysis of word frequency counts. Texts written by an author are treated as samples from his vocabulary and word frequency counts are used to help distinguish his style from the style of others (see, e.g., Holmes, 1985). Given that most words appear very few times and very few words are repeated many times, word frequency count data have reverse J-shaped distributions with extraordinarily long upper tails.

Typically, the process generating frequency count data can be modelled through a two stage process, with each count being Poisson distributed but with an expected value randomly changing from count to count with a distribution that relates to the class frequency distribution in the population. That naturally leads one to the use of Poisson mixture models for this kind of data.

The inverse Gaussian-Poisson mixture model was introduced by Holla (1966) to model highly skewed non-negative integer data, and it has been widely used ever since in many different fields of application involving frequency count data. In particular, this model has been widely used in the analysis of the frequency of word or species frequency data ever since Sichel (1975), where given that one can not count unobserved words or species it is necessary to truncate this model at zero. Even though this model is typically recommended because it provides good fits, what makes it useful is that it allows one to interpret the inverse gaussian mixing distribution as the distribution of the word frequencies of the vocabulary from which the text is coming from.

The first goal of the paper is to propose a Bayesian analysis based on this statistical model, and to illustrate how it allows one to use the posterior distribution of the inverse of the mean and of measures of the variability of the model mixing distribution to estimate the size and lack of diversity of vocabulary. The second goal is to explore the usefulness of an alternative Bayesian analysis based on the statistical model that results from switching the mixing and the truncation stages and leading to the inverse Gaussian-Truncated Poisson mixture model.

The paper is organized as follows. Section 2.2 describes word frequency count data and it motivates the use of the truncated inverse gaussian-Poisson mixture model in the analysis of that type of data. Section 2.3 proposes a Bayesian analysis based on this later model and it uses it on the word frequency counts of texts by Macaulay, Carroll, Wells and Doyle. The validity of this Bayesian model is checked by exploring the posterior distribution of the Pearson errors and by implementing various posterior

predictive consistency checks. The texts considered were purposely chosen to be long to test the limitations of the model and to illustrate the type of departures found through the model checking diagnostic tools proposed as part of the Bayesian analysis. Section 2.3 also investigates the role that the posterior distribution of the model mixing density plays as an approximation of the posterior distribution of the density of vocabulary, and its use as a fingerprint of the literary style of the author in his texts.

Section 2.4 considers an alternative analysis based on the inverse gaussian-truncated Poisson mixture model, first considered in Puig, Ginebra and Font (2010). In Section 2.5 the two analysis are compared based on the posterior distribution of the sum of the squares of the Pearson errors and on the value taken by overall goodness of fit test statistics; even though the analysis in Section 2.4 based on the model that first truncates and then mixes is not as meaningful as the one in Section 2.3 based on the model that first mixes and then truncates, because it does not allow one to link the data with the distribution of the word frequencies of the vocabulary of the author, this alternative model fits some of the word frequency count data sets a bit more accurately than the usual inverse gaussian-Poisson model. Finally, Section 2.6 ponders some of the practical implications of what is exposed in the paper.

2.2 Word frequency count data and statistical model

2.2.1 Description of the data

To characterize the style of an author through its vocabulary the basic assumption made is that the author has available a list of all the words that he knows, and that the i-th word in that list is characterized through the proportion of times that that word would be found in a text of infinite length by that author, which is denoted by π_i. The set of probabilities π_j when j ranges over all the v words known by an author, $(\pi_1, \ldots, \pi_v)$, with $\sum_{i=1}^{v} \pi_i = 1$, constitute the distribution of the vocabulary of that author.

For mathematical convenience, one treats the π_j's as a continuous variable with a density function $\psi(\pi)$. This frequencies density function characterizes the vocabulary of the author and it should be of interest to anyone characterizing the style of an author. In particular, the larger the number of words in the vocabulary of an author, v, the smaller the π_j's, which links a small expected value for $\psi(\pi)$ with a rich vocabulary. Furthermore, given v, the closer the distribution $(\pi_1, \ldots, \pi_v)$ is to the uniform distribution, the more peaked $\psi(\pi)$ is around $1/v$, which links variability of $\psi(\pi)$ with lack of diversity of

vocabulary, as recently discussed in detail in Ginebra and Puig (2010).

As an approximation, texts written by an author will be treated as if they were random samples drawn from his vocabulary. If one denotes the total number of words (tokens) in a given text by n, the number of occurrences of the i-th word by n_i, and the proportion of occurrences of that word in that text by $\hat{\pi}_i = n_i/n$, the expected value of $\hat{\pi}_i$ is π_i.

Let v_n denote the number of different words (types) in a text of size n, and let $v_{r:n}$ denote the number of different words appearing exactly r times in it. The proportion of different words appearing exactly r times in a text of size n will be denoted by $\hat{p}_{r:n} = v_{r:n}/v_n$ and its expectation, which depends on n, will be denoted by $p_{r:n}$.

By counting the number of words used once, $v_{1:n}$, the number of words used twice, $v_{2:n}$, and so on, one obtains the vector $(v_{1:n}, v_{2:n}, \ldots, v_{n:n})$ of word frequency counts. Table 2.1 presents the word frequency count for the nouns in the Macaulay's essay on Bacon, considered in Sichel (1975), and of all the words in a Turkish archeology text, in *Alice in Wonderland* and in *Through the Looking Glass* by Carroll, in *The Hound of the Baskervilles* by Doyle, and in *The War of the Worlds* by Wells, which are all considered in Baayen (2001). Other than for the essays on Bacon, in these data sets all parts of speech are counted including articles, prepositions, conjunctions, nouns, adjectives, verbs and adverbs.

For example, the third row in Table 2.1 indicates that Alice in Wonderland has a total of $n = 26505$ words out of which $v_n = 2651$ are different words; in it 1176 words appear once, 402 words appear twice, 233 words appear three times and so on, with the most frequent word appearing 1631 times. Given that most of the words appear only a few times and few words are repeated many times, the distribution of $(v_{1:n}, v_{2:n}, \ldots, v_{n:n})$ is reverse J-shaped with a very long upper tail.

2.2.2 The zero truncated IG-Poisson mixture model

If a specific word, i, has a probability π_i of being used each time that an author writes a word, the number of times that this word appears in one of its texts with a total of n words would be distributed as a binomial(n, π_i). Hence, if its distribution of vocabulary was $\psi(\pi)$, the probability that a word from that vocabulary appears exactly r times in a text of size n, $p_{r:n}$, can be modelled through a $\psi(\pi)$-binomial mixture model. Usually n will be large and all the π_i will be small, and one can approximate $p_{r:n}$ through a $\psi(\pi)$-Poisson mixture model.

	$v_{1:n}$	$v_{2:n}$	$v_{3:n}$	$v_{4:n}$	$v_{5:n}$	$v_{6:n}$	$v_{7:n}$	$v_{8:n}$	$v_{9:n}$	$v_{10:n}$	$v_{11:n}$	...	n	v_n
Turkish A.	2326	477	178	107	53	33	22	26	7	7	12	...	6939	3302
E. Bacon	990	367	173	112	72	47	41	31	34	17	24	...	8049	2048
Alice in W.	1176	402	233	154	99	57	65	52	32	36	23	...	26505	2651
Through L.	1491	460	259	148	113	78	61	47	28	26	26	...	28767	3085
Hound B.	2836	889	449	280	208	137	116	92	86	52	48	...	59241	5741
War of W.	3613	1138	567	340	250	177	135	93	72	67	44	...	59938	7112

Table 2.1: Part of the word frequency count data sets of the nouns in the Macaulay's essay on Bacon, and of all the words in a Turkish archeology text, in *Alice in Wonderland*, in *Through the Looking Glass*, in *The Hound of the Baskervilles* and in *The War of the Worlds*.

Given that one can not count the words that an author knows but are not observed in the text, one needs to consider the zero truncated version of it,

$$p_{r:n}^{tpm} = \frac{1}{1 - \int_{R^+} e^{-n\pi}\psi(\pi)d\pi} \int_{R^+} \frac{(n\pi)^r e^{-n\pi}}{r!}\psi(\pi)d\pi, \qquad \text{for } r = 1, 2, \ldots. \tag{2.1}$$

This argument entitles one to interpret the model mixing density $\psi(\pi)$ as the density of the word frequencies of the vocabulary. Following a recommendation in Good (1953) , Sichel (1975, 1986a) models the mixing distribution through an inverse gaussian distribution, denoted by IG(b, c), which is defined on R^+ and has a density function

$$\psi(\pi|b, c) = \frac{b}{2}\sqrt{\frac{c}{pi}} e^b \pi^{-3/2} e^{-\frac{\pi}{c} - \frac{b^2 c}{4\pi}}, \tag{2.2}$$

where b is in $(0, \infty)$, c is in $(0, \infty)$, and where pi is the known irrational number. Even though the support of (2.2) is $(0, \infty)$, under the values of (b, c) that one considers in practice (2.2) is negligible for $\pi > .1$. For details on this distribution see for example Seshadri (1998).

By replacing (2.2) in (2.1) and solving the integral one obtains that the probability function of the zero truncated IG-Poisson mixture model is

$$p_{r:n}^{tigp}(b, c) = \frac{1}{(1 + cn)^{-1/4} K_{-1/2}(b) - K_{-1/2}(b\sqrt{1 + cn})} \frac{(\frac{1}{2}\frac{bcn}{\sqrt{1+cn}})^r}{r!} K_{r-1/2}(b\sqrt{1 + cn}), \tag{2.3}$$

for $r = 1, 2, \ldots$, where $K_a(\cdot)$ is the modified Bessel function of the third kind of order a. The support of (2.3) is unbounded but in practice $p_{r:n}^{tigp}(b, c)$ dies out very fast with increasing r. This two parameter model is actually a special case of the three parameter generalized inverse gaussian-Poisson mixture model considered in Sichel (1975).

Sichel (1975, 1986a), Pollatschek and Radday (1981), Holmes (Holmes1992), Holmes and Forsyth (1995), Baayen (2001) and Riba and Ginebra (2006) fit this model to word frequency count data, and find that it provides very good fits when the texts are in English and have less than $n = 10000$ words. The texts considered in this paper were purposely chosen to have n larger than that in order to illustrate the type of departures from the model found through the model checking diagnostic tools considered next.

2.3 Bayesian analysis based on the zero truncated IG-Poisson

2.3.1 Posterior distributions

If one assumes that word frequencies are independent and identically distributed as a zero truncated IG-Poisson distribution, the likelihood function is such that:

$$L^{tigp}_{(v_{1:n},\ldots,v_{n:n})}(b,c) \propto \Pi_r (p^{tigp}_{r:n}(b,c))^{v_{r:n}}, \tag{2.4}$$

and the posterior distribution of (b,c), is

$$\pi(b,c|Data) \propto \pi(b,c) L^{tigp}_{(v_{1:n},\ldots,v_{n:n})}(b,c), \tag{2.5}$$

where $\pi(b,c)$ is the prior distribution. We report the results based on a reference prior assuming that b and c are independently distributed Gamma$(.001, .001)$.

The posterior distribution (2.5) is too complex to be computed analytically. Instead, we simulated samples from the posterior distribution of (b,c) through the Markov Chain Monte Carlo method implemented through WinBUGS (Spiegelhalter et al., 2003). Unfortunately, not all the distributions needed to simulate from our models are available in WinBugs. To solve this problem one can use the WinBUGS Development Interface (Lunn, 2003; Wetzels et al., 2009) to program functions and distributions that are unavailable in WinBUGS; in particular for this model we used this WBDev to simulate from the zero truncated IG-Poisson.

We have monitored the convergence of every chain by visual inspection of graphical histories and by computing the $\hat{R}$ statistic proposed by Gelman and Rubin (1992) based on four initially overdispersed sampling chains. The burning period of 4000 iterations has been determined from this preliminary analysis, by checking that it is what is required for the $\hat{R}$ statistic to be less than 1.05 for all parameters. The MCMC based estimation

has been performed with the subsequent 2500 values of each series. Our descriptions of posterior distributions are thus based on sample size of 10000 values.

Figure 2.1 presents samples from the posterior distributions of (b, c) for the word frequency count data sets in Table 2.1, under our reference prior. That figure also presents a non-parametric kernel posterior density estimate based on these samples. When we tried different priors, we obtained very similar results, which is a combined consequence of having word frequency count data sets from very large texts and hence being very informative, coupled with the lack of information about (b, c) in the prior distribution used. That also explains that, except for the Turkish archeology text, all the posterior distributions have a very normal like behavior.

For the Turkish archeology text the maximum likelihood estimate of (b, c) is $(0., 0.0013)$, on the boundary of the parameter space, which explains that its posterior distribution is concentrated near that boundary. For these situations, Puig et al. (2009) proposes an extension of the parameter space of the IG-Poisson model that allows for better model fits but which does not allow one to interpret the extended part of the model as a Poisson mixture model, and hence it does not allow one to make inferences about the model mixing distribution.

2.3.2 Model checking

To check the validity of the model we explore the posterior distribution of the Pearson errors,

$$\epsilon^{p}_{r:n}(b, c) = \frac{v_{r:n} - v_n p_{r:n}(b, c)}{\sqrt{v_n p_{r:n}(b, c)}} \tag{2.6}$$

for each category r. To compute these errors the categories were aggregated the least so that the posterior expected count in each category was at least 5.

The samples from the posterior distributions of $\epsilon^{p}_{r:n}(b, c)$, in Figure 2.2, indicate that this model fits the word frequency count data of the *Essays on Bacon* very well, and it fits the word frequency count data of *Alice in Wonderland* and of *Through the Looking Glass* fairly well.

It is also clear from Figure 2.2 that for the Turkish archeology text this model *systematically* leads to positive errors, (and therefore larger observed $v_{r:n}$ counts than the expected $v_n p_{r:n}(b, c)$ counts), for all the categories except for $r = 1$ and for the categories representing the tail of the distribution for which negative errors with anomalously large

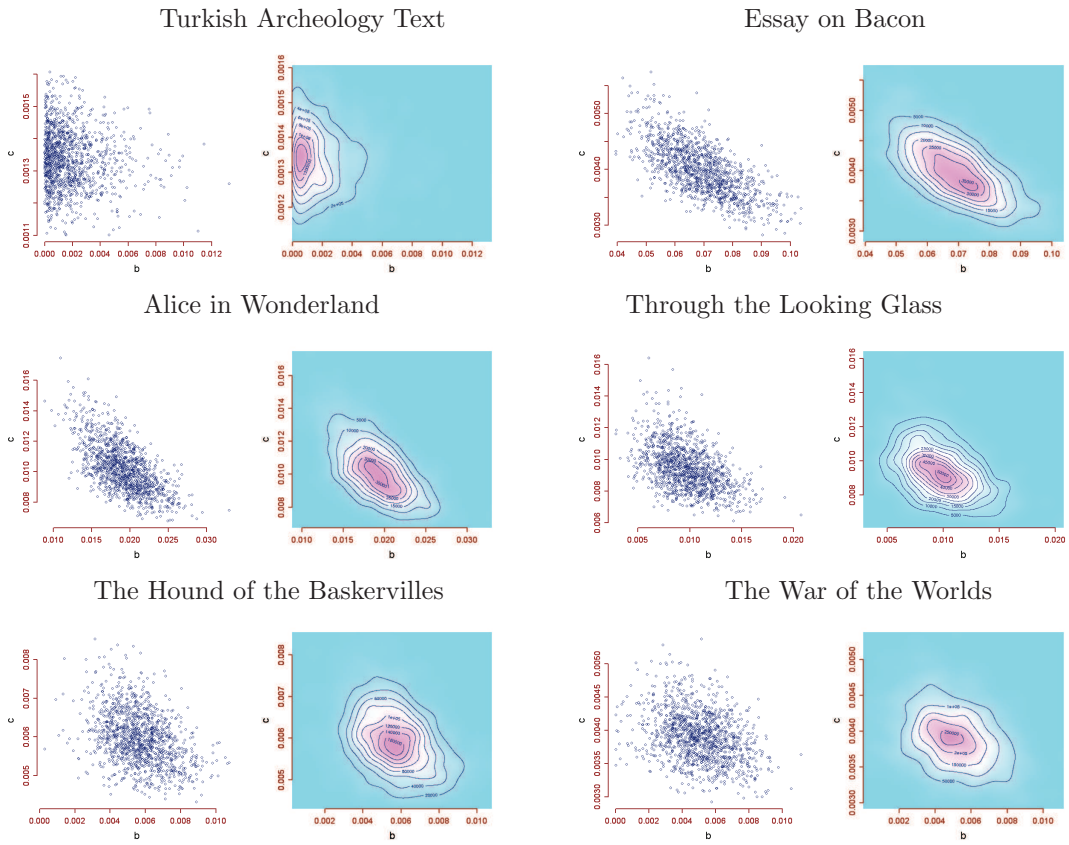


Figure 2.1: Sample of 10000 observations from the posterior distribution of (b, c) under the truncated IG-Poisson model, in (2.3), with independent Gamma(.001, .001) priors for b and c, together with a non-parametric posterior density estimate based on those samples.

absolute values occur.

To a smaller degree, these features are repeated in the posterior of the $\epsilon^{p}_{r:n}(b,c)$'s for *The Hound of the Baskervilles* and T*he War of the Worlds*, with the only difference that for them *systematic* negative errors with anomalously large absolute values happen for a few small r categories but not for $r = 1$. This partial failure follows from the fact that these two texts have a total of almost 60.000 words each, which puts them outside the range of applicability of this simple two parameter model because it fails to capture the large over-dispersion present in word frequency count data for texts of this length.

To further understand where does this model fail when it does, posterior predictive consistency checks were implemented along the lines advocated for in chapter 6 of Gelman et al. (2004). The idea is that if the model is accurate, replicates of the data obtained by simulation from the Bayesian model should look similar to the observed data. To simulate replicates of the data using the Bayesian model, we simulated a sample of 10000 (b,c)'s from its posterior distribution and for each simulated (b,c) we used the corresponding IG(b,c) distribution as if it was the vocabulary distribution and simulated a word frequency count set from it forcing all the simulated count sets to have the same total number of words, n, as the observed sample. To quantify the discrepancy between simulated and observed data we compared the number of words appearing once, $v_{1:n}$, the number of words appearing twice, $v_{2:n}$, and the total number of different words, v_n, in the various samples of the simulated data and in the observed data.

Figure 2.3 presents the results from these checks. Observe that this model only fails to explain the number of words observed once, $v_{1:n}$, for the *Turkish archeology text* for which almost all the simulated word frequency count data set samples have less than the 3302 words observed once in it. This Bayesian model adequately explains the number of different words, v_n, and the number of words observed twice, $v_{2:n}$, even though for the two longest texts the simulated values for $v_{2:n}$ tend to be smaller than the observed values.

2.3.3 Density, richness and diversity of vocabulary

The main advantage in using the zero truncated Poisson mixture models is that they allow one to interpret the mixing density as the density of the vocabulary of the author. When the Bayesian analysis based on the truncated IG-Poisson model reproduces adequately the features of interest in the data, one can use the posterior distribution of the density of IG(b,c) as an approximation to the posterior distribution of the density

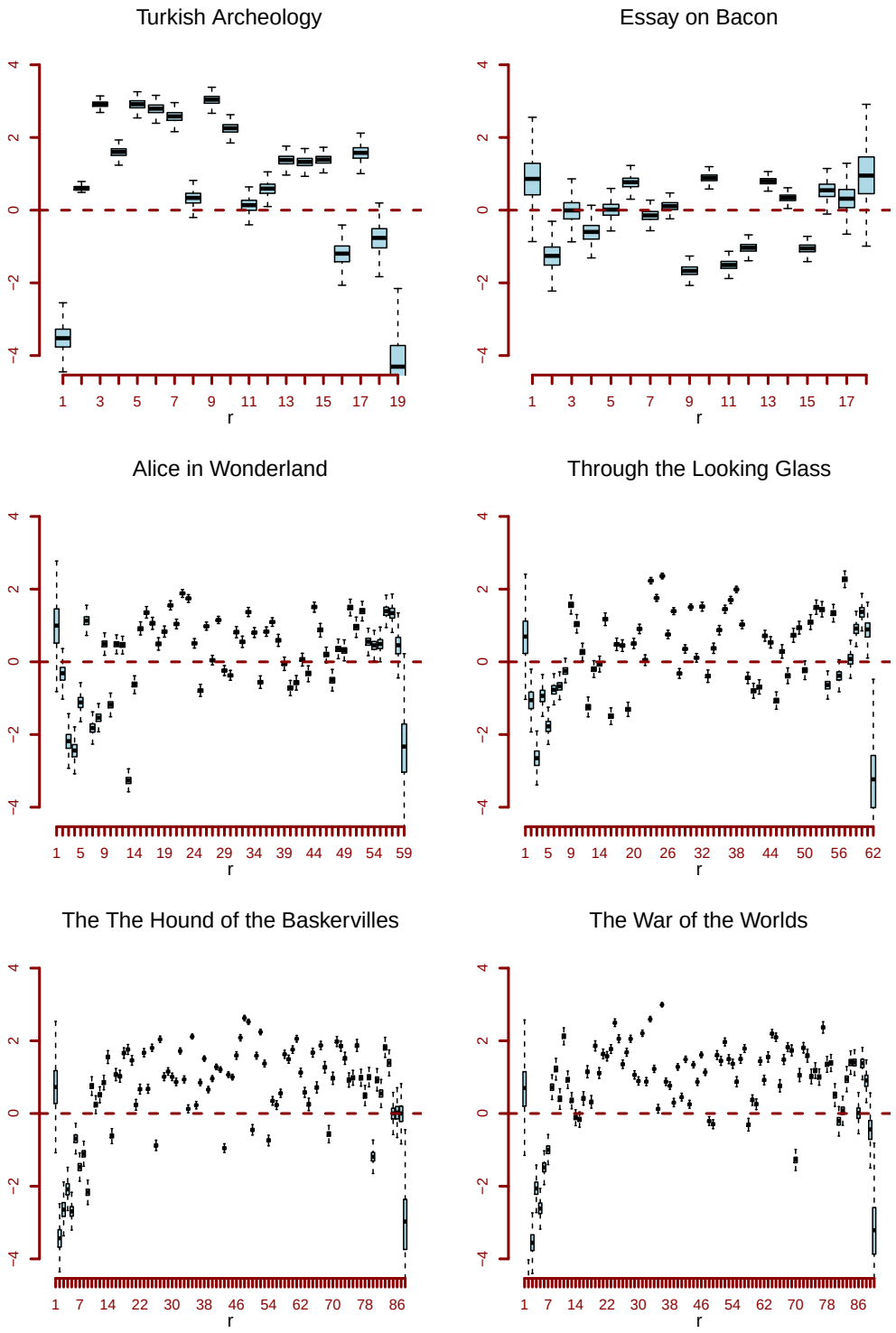


Figure 2.2: Box-plots of samples of 10000 observations from the posterior distribution of the Pearson errors, $\epsilon^p_{r:n}(b, c)$, under the zero truncated IG-Poisson model, in (2.3), with independent Gamma(.001, .001) priors for b and c.

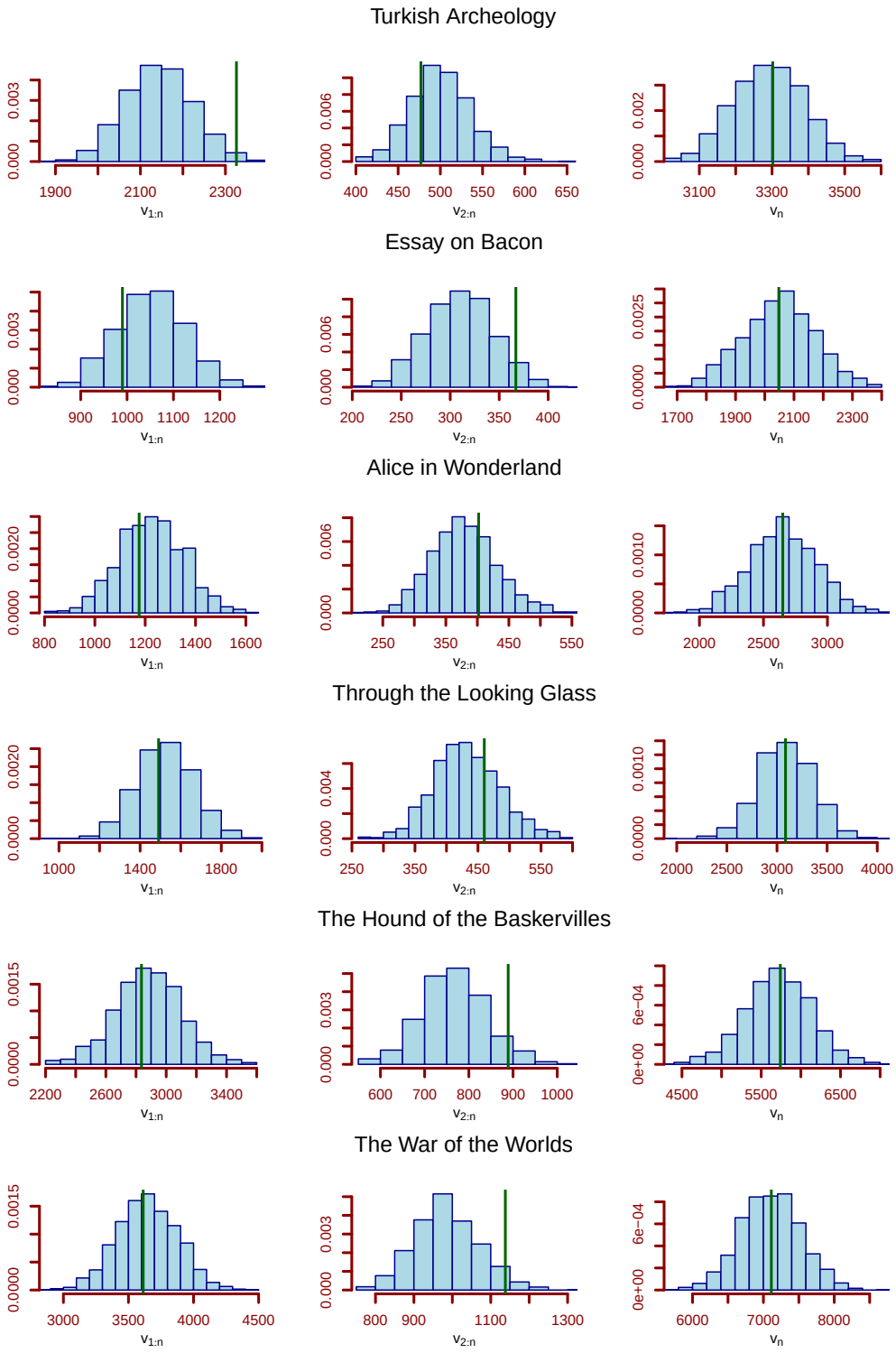


Figure 2.3: Observed value and sample of 10000 observations from the posterior predictive distribution of $v_{1:n}$, of $v_{2:n}$ and of v_n under the zero truncated IG-Poisson model, in (2.3), with independent Gamma(.001, .001) priors for b and c.

of vocabulary of the author.

Figure 2.4 presents samples from the posterior distribution of the model mixing $IG(b, c)$ density function for the data in Table 2.1, together with the densities of the $IG(\hat{b}_{pm}, \hat{c}_{pm})$ distribution summarizing those samples, where $(\hat{b}_{pm}, \hat{c}_{pm})$ is the posterior mode for (b, c) obtained by maximizing the kernel joint density estimate in Figure 2.1. Given that the analysis based on the truncated IG-Poisson model does not capture the main features of the word frequency counts in the *Turkish archeology text* and in the two longest texts well, one should interpret the samples of the posterior distribution of the mixing densities for these texts with caution.

One could compare these density samples with the help of functional data analysis tools (Ramsey and Silverman, 2005), but it is better to summarize them through real valued quantities that help characterize literary style. In particular, Note that the smaller the values of the π_j's, the larger the total number of words in it, v, the smaller the expected value of the π_j's under $\psi(\pi)$, and the richer the vocabulary. Sichel (1986a, 1986b) proposes estimating the size v through the closest integer to

$$v(\psi) = \frac{1}{E_\psi[\pi]} = \frac{2}{bc}, \tag{2.7}$$

where the last equality holds only when $\psi(\pi)$ is the IG(b, c).

To measure the diversity of $(\pi_1, \ldots, \pi_v)$, note that given v, the higher and narrower the peak of $\psi(\pi)$, the closer the vocabulary distribution $(\pi_1, \ldots, \pi_v)$ is to the uniform distribution, the smaller the variability of the π_j's under $\psi(\pi)$ and the more even and diverse the distribution of vocabulary. Simple measures of the diversity of the vocabulary of the author would be the negative or the inverse of $Var_\psi[\pi]$ or of any other measure of the variability of $\psi(\pi)$, like

$$e(\psi) = -\log Var_\psi[\pi] = -\log \frac{bc^2}{4}, \tag{2.8}$$

where the last equality holds only when $\psi(\pi)$ is the IG(b, c) distribution. Another useful measure of the diversity in $(\pi_1, \ldots, \pi_v)$ is the Gini-Simpson index, $D_1(\pi_1, \ldots, \pi_v) = 1 - \sum_{i=1}^{v} \pi_i^2$, which is the probability that two words picked at random from a text of infinite length would be different. If one assumes that the π_j's are identically distributed as $\psi(\pi)$. The expected value of this index is:

$$D_1(\psi) = 1 - \sum_{i=1}^{v} E_\psi[\pi^2] = 1 - \frac{c}{2}(1 + b), \tag{2.9}$$

where the last equality holds only when $\psi(\pi)$ is the IG(b, c) distribution. For more details on the relationship between measuring the variability of $\psi(\pi)$ and measuring the lack of

diversity of the corresponding vocabulary or population, see Ginebra (2007) and Ginebra and Puig (2010).

Figure 2.5 presents samples from the posterior distribution of $\log_{10} v(\psi)$, of $e(\psi)$ and of $D_1(\psi)$. We also sampled from the posterior distribution of the expectation of the entropy of $(\pi_1, \ldots, \pi_v)$, which is another measure of the diversity in the vocabulary of the author, but it had a huge dispersion and it was not as useful as the Gini-Simpson index based measure.

According to Figure 2.5 the richest vocabulary is the one from which the Turkish archeology text was produced. That figure also indicates that the word frequency count set coming from the least rich vocabulary seems to be one for the essays on Bacon, which makes a lot of sense because that is the only case in which word frequency counts refer only to the names in the text and not to all types of words. According to that figure the texts by Carroll, Alice in Wonderland and Through the Looking Glass are the ones that come from the least diverse vocabulary of all the texts under consideration.

A non-Bayesian way of assessing richness and diversity of vocabulary would estimate (2.7), (2.8) and (2.9) by replacing (b, c) by its maximum likelihood estimator, which would be close to the posterior modes for v, e and D_1. The advantage of the Bayesian way of assessing richness and diversity of vocabulary through Figure 2.5 is that it also provides a convenient estimate of the uncertainty in those richness and diversity measure estimates, which is something that is a lot more difficult to obtain in the non-Bayesian setting.

2.4 Bayesian analysis based on the IG-Truncated Poisson

As an alternative to (2.1) the order of the mixing and the truncation stages can be switched, leading to a mixture of the truncated Poisson model. That is, let the probability of a word being repeated exactly r times in a text of size n be modelled through

$$p_{r:n}^{mtp} = \int_{R^+} \frac{(n\pi')^r e^{-n\pi'}}{(1 - e^{-n\pi'})r!} \psi'(\pi') d\pi', \qquad \text{for} \quad r = 1, \ldots, n. \tag{2.10}$$

As discussed in Puig, Ginebra and Font (2010), the model mixing density $\psi'(\pi')$ in (2.10) represents the v_n words that have appeared at least once in the given text of size n, and not all the v words in the vocabulary of the author. Hence here $\psi'(\pi')$ heavily depends

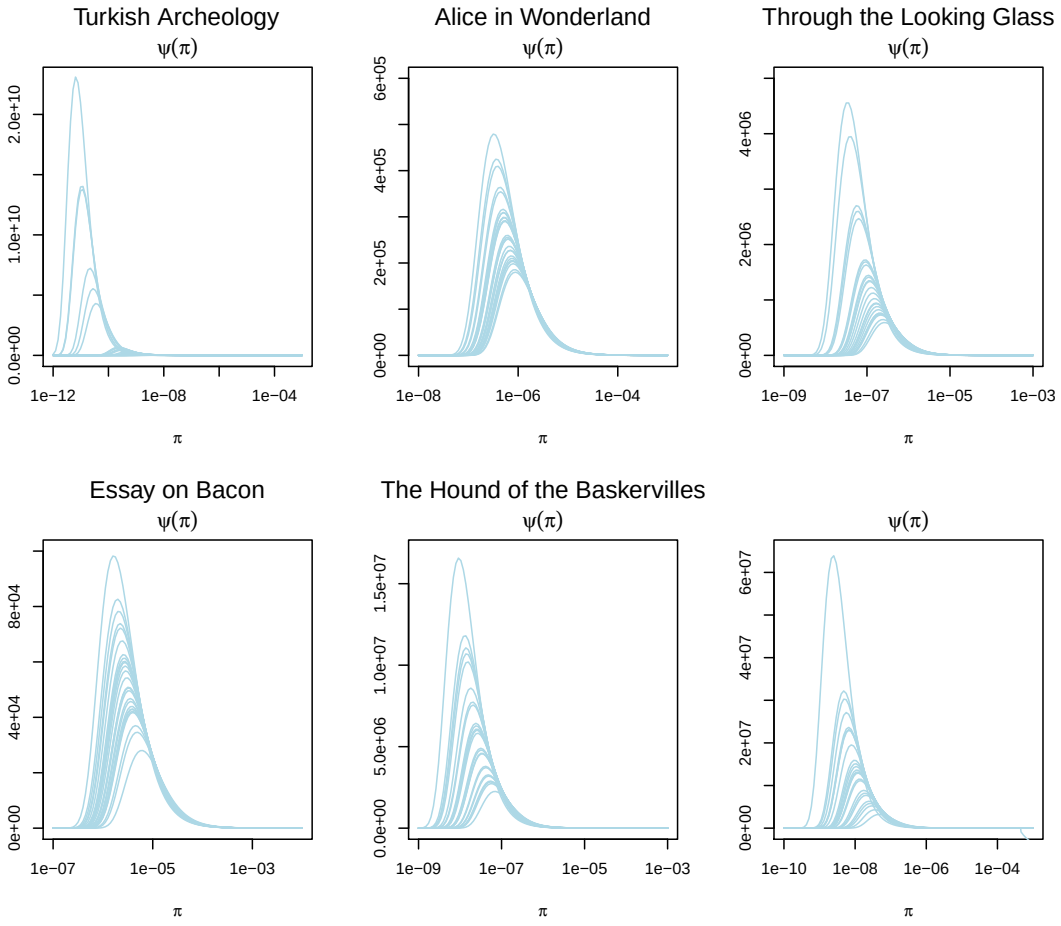


Figure 2.4: Samples of 25 densities of the posterior distribution of the mixing density, IG(b, c), under the zero truncated IG-Poisson(b, c) model with independent Gamma$(.001, .001)$ priors for b and c. The density in red is the one of $IG(\hat{b}_{pm}, \hat{c}_{pm})$. These samples serve as an approximation to the posterior distributions of the density of vocabulary of the authors.

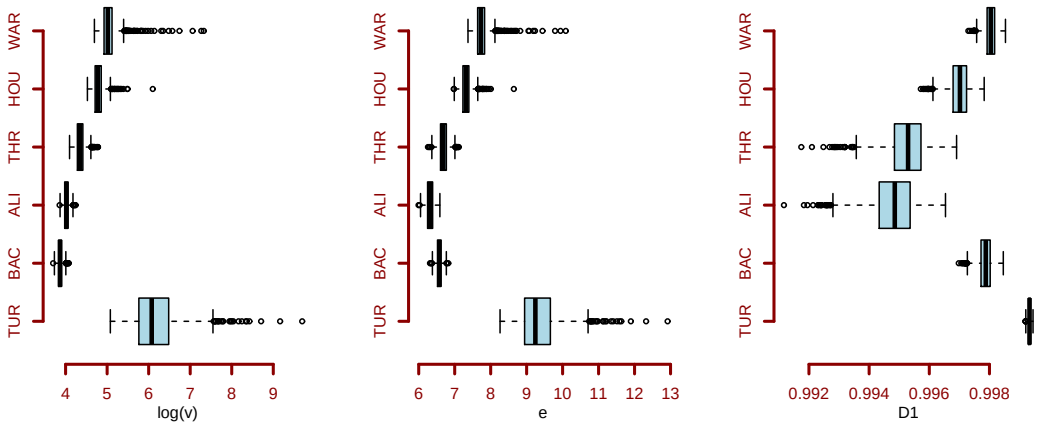


Figure 2.5: Box-plots of samples of 10000 observations from the posterior distribution of $\log_{10} v(\psi)$, which measures the richness, and of $e(\psi) = -\log_{10} Var_{\psi}[\pi]$ and $D_1(\psi)$, which measure the diversity of the vocabulary of the author. The model is the zero truncated IG-Poisson with independent Gamma(.001, .001) priors for b and c.

on the text size n and it can not be interpreted as the density of vocabulary of the author in the way the mixing density $\psi(\pi)$ associated with (2.1) was interpreted in subsection 2.3.3. That puts the IG-Truncated Poisson model in a disadvantage when it is compared with the truncated IG-Poisson model.

The model obtained from (2.10) when $\psi'(\pi')$ is an inverse gaussian distribution, IG(b, c), is the IG-TruncatedPoisson mixture model and the corresponding $p_{r:n}$ is denoted as $p^{igtp}_{r:n}(b, c)$.

The right panel of Figure 2.6 presents samples from the posterior distribution of (b, c) for the word frequency count data in Table 2.1, assuming that the likelihood function is proportional to:

$$L^{igtp}_{(v_{1:n},\ldots,v_{n:n})}(b, c) \propto \Pi_r (p^{igtp}_{r:n}(b, c))^{v_{r:n}}, \tag{2.11}$$

and that the prior is such that b and c are independent Gamma(.001, .001). The posterior distribution here is again too complex to be computed analytically, and we again simulated samples from the posterior distribution of (b, c) through the MCMC method implemented through WinBUGS (Spiegelhalter et al., (2003). Here we used the WBDev Interface (Lunn, 2003; Wetzels et al., 2009) to simulate from the inverse gaussian distribution and from the zero-truncated Poisson distribution because they were not originally available in WinBugs.

Bayesian analysis based on truncated mixture models are easier to interpret than the

ones based on the mixture of truncated models, and yet both approaches can be helpful when discriminating between the style of different authors as illustrated in Figure 2.6 through the simultaneous representation of samples from the posterior distributions of (b, c) under both models. Observe that the samples for *Alice in Wonderland* and for *Through the Looking Glass* are very close, in line with the fact that both texts share the same author.

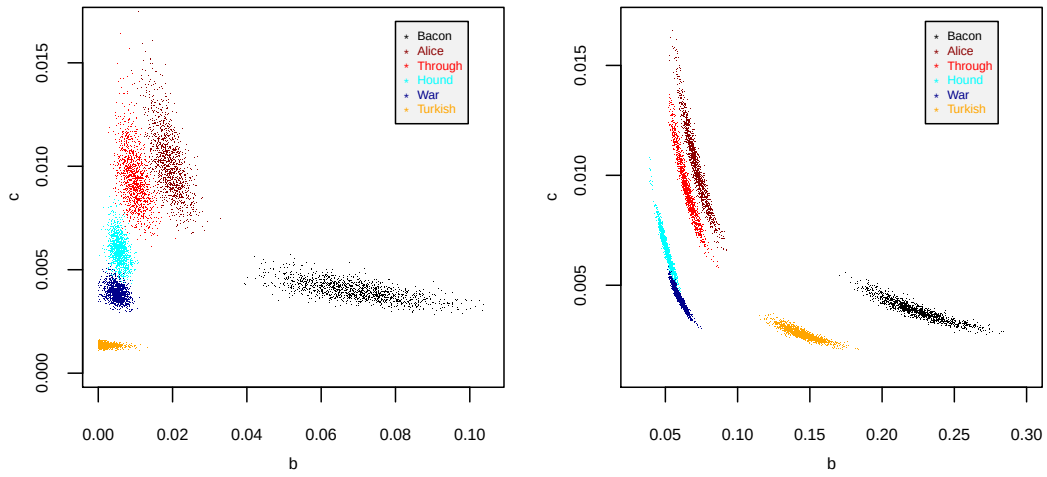


Figure 2.6: Samples of 10000 observations from the posterior distribution of (b, c) under the truncated IG-Poisson model, in the left hand side panel, and under the IG-TruncatedPoisson model, in the right hand side panel, both under independent Gamma$(.001, .001)$ priors for b and c and for the word frequency count sets in Table 2.1.

The posterior of (b, c) for the Turkish archeology text in the right panel of Figure 2.6, is not concentrated near the boundary the way it is for the posterior of (b, c) in the left panel of Figure 2.6, because the maximum likelihood estimate of b under the IG-Truncated Poisson model is not in that boundary. The strong inverse dependence in the posterior distributions of (b, c) in the right hand side panel of Figure 2.6, which is not present in the posterior distributions in the left hand side panel of that same Figure 2.6, follows from the fact that here the mixing ψ'=IG(b, c) distribution represents only the observed vocabulary with a size known to be v_n, which links b and c through $v_n = 2/bc$, as in (2.7).

Figure 2.7 explores the posterior distribution of the Pearson errors in (2.6), $\epsilon^p_{r:n}(b, c)$, for the same aggregated categories used in Figure 2.2. The samples of the posterior of $\epsilon^p_{r:n}(b, c)$ for the Turkish archeology text and for the essays on Bacon indicate that

this model fits their word frequency counts very well. That figure also indicates that the model fits the word frequency counts of Alice in Wonderland and of Through the Looking Glass fairly well, only mildly failing with the frequency of a few categories with a small r and with the frequency of the category of the most frequent words. These mild failures become more serious for the two longest texts, which require three parameter models.

Figure 2.8 presents the results of the posterior predictive consistency checks described in Section 2.3 for this alternative Bayesian IG-TruncatedPoisson model. Different from what happens in Figure 2.3, here the word frequency counts simulated under this model have values for $v_{1:n}$, $v_{2:n}$ and v_n that closely match the observed values for all the six texts considered.

2.5 Model comparison

The truncated mixture models in (2.1), like the one in Section 2.3, are more natural to formulate and to interpret than the mixture of truncated models in (2.10), like the one in Section 2.4, because they let one make inferences about the density of the vocabulary of the author. Nevertheless, the later models might be theoretically easier to treat and they might yield better fits.

One could formally chose between the Bayesian models in Sections 2.3 and 2.5 by computing the corresponding Bayes factor, but it is more meaningful to compare them through the posterior distribution of their Pearson errors in Figures 2.2 and 2.7, because that points towards the differing behavior of both models. In our case for example, Figure 2.7 indicates that the IG-Truncated Poisson model captures the overdispersion in the word frequency counts of the *Turkish archeology text* and of the two longest texts than the truncated IG-Poisson model, which is a fact that would be missed by just computing the Bayes factor.

To compare their overall goodness of fit one can also explore the posterior distribution of the sum of the squares of their Pearson errors,

$$\chi^2(b,c) = \sum_r \epsilon^p_{r:n}(b,c)^2 = \sum_r (\frac{v_{r:n} - v_n p_{r:n}(b,c)}{\sqrt{v_n p_{r:n}(b,c)}})^2. \tag{2.12}$$

The samples of the posterior distribution of $\chi^2(b,c)$ in Figure 2.9 indicate that the alternative IG-Truncated Poisson model provides a better overall fit than the truncated IG-Poisson model for the word frequency count data sets of the *Turkish archeology text*

and of the two longest texts in Table 2.1. The performance of these two models on the word frequency count data sets of the *essays on Bacon* and of the two texts by Carroll is very similar. Note that, differently from what happens in the non-Bayesian model comparison approach based on the values adopted by goodness of fit test statistics, the posterior distributions in Figure 2.9 capture the degree of the uncertainty behind the conclusion reached.

Table 2.2 presents the posterior modes for (b, c), $(\hat{b}_{pm}, \hat{c}_{pm})$, obtained by maximizing the smoothed estimate of the joint posterior densities for (b, c) in Figures 2.1 and 2.6, next to their maximum likelihood estimates, $(\hat{b}_{ml}, \hat{c}_{ml})$, under both models. Note that these two estimates are very similar. Table 2.2 also includes the maximum of the loglikelihood function and the values taken by the goodness of fit test statistic obtained as the sum of the squares of the Pearson residuals,

$$X^2(\hat{b}, \hat{c}) = \sum_r \left(\frac{v_{r:n} - v_n p_{r:n}(\hat{b}, \hat{c})}{\sqrt{v_n p_{r:n}(\hat{b}, \hat{c})}}\right)^2. \tag{2.13}$$

To evaluate it the categories are aggregated the least so that their expected count is at least 5. An alternative goodness of fit test statistic that we have tried is the one obtained by replacing $p_{r:n}(\hat{b}, \hat{c})$ in (2.13) by an estimate of the posterior expected value of $p_{r:n}(b, c)$ based on a sample from the posterior distribution of (b, c).

The values in Table 2.2 are in agreement with the conclusions reached elsewhere. The truncated IG-Poisson model in Section 2.3 fits the count data sets of the *essays on Bacon* and of the two texts by Carroll fairly well. On the other hand the IG-Truncated Poisson model in Section 2.4 fits fairly well the count data sets of all the texts except the ones of the two longest texts, which are still better fit by this model than by the model in Section 2.3.

2.6 Concluding remarks

The zero truncated IG-Poisson(b, c) model used in Sections 2.3 is known to provide good fits for word frequency count data sets from texts with less than 10000 words. Nevertheless, we purposely chose to illustrate our Bayesian analysis based on this two-parameter model with data from texts that are considerably longer than that in order to test the limits of this model and to check the model checking diagnostic tools. Even though we were surprised by the flexibility allowed by this simple two-parameter model, we indeed find this model fails to capture the large degree of overdispersion present in the count data from the longer texts.

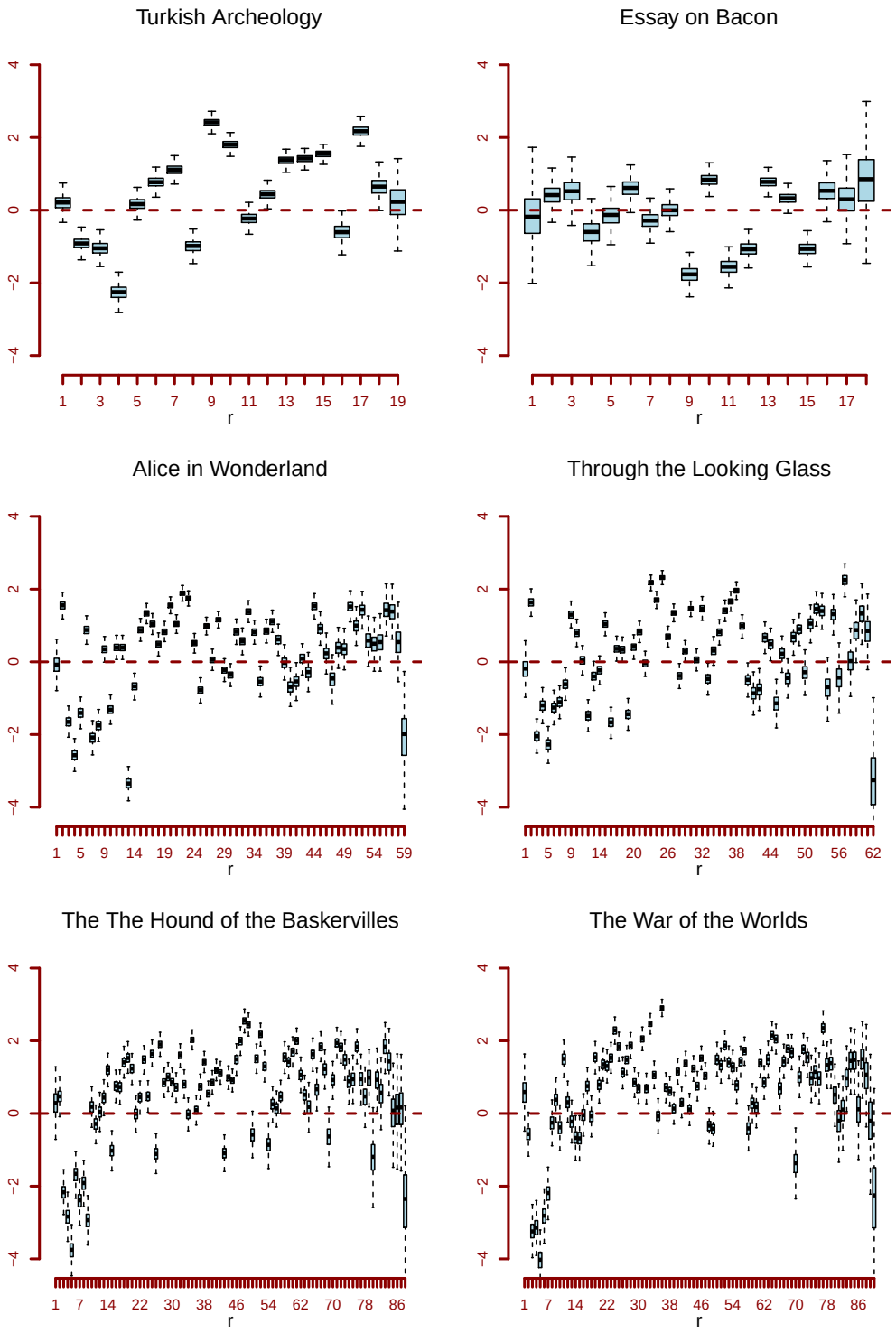


Figure 2.7: Box-plots of samples of 10000 observations from the posterior distribution of the Pearson errors, $\epsilon^p_{r:n}(b, c)$, under the IG-TruncatedPoisson model with independent Gamma$(.001, .001)$ priors for b and c.

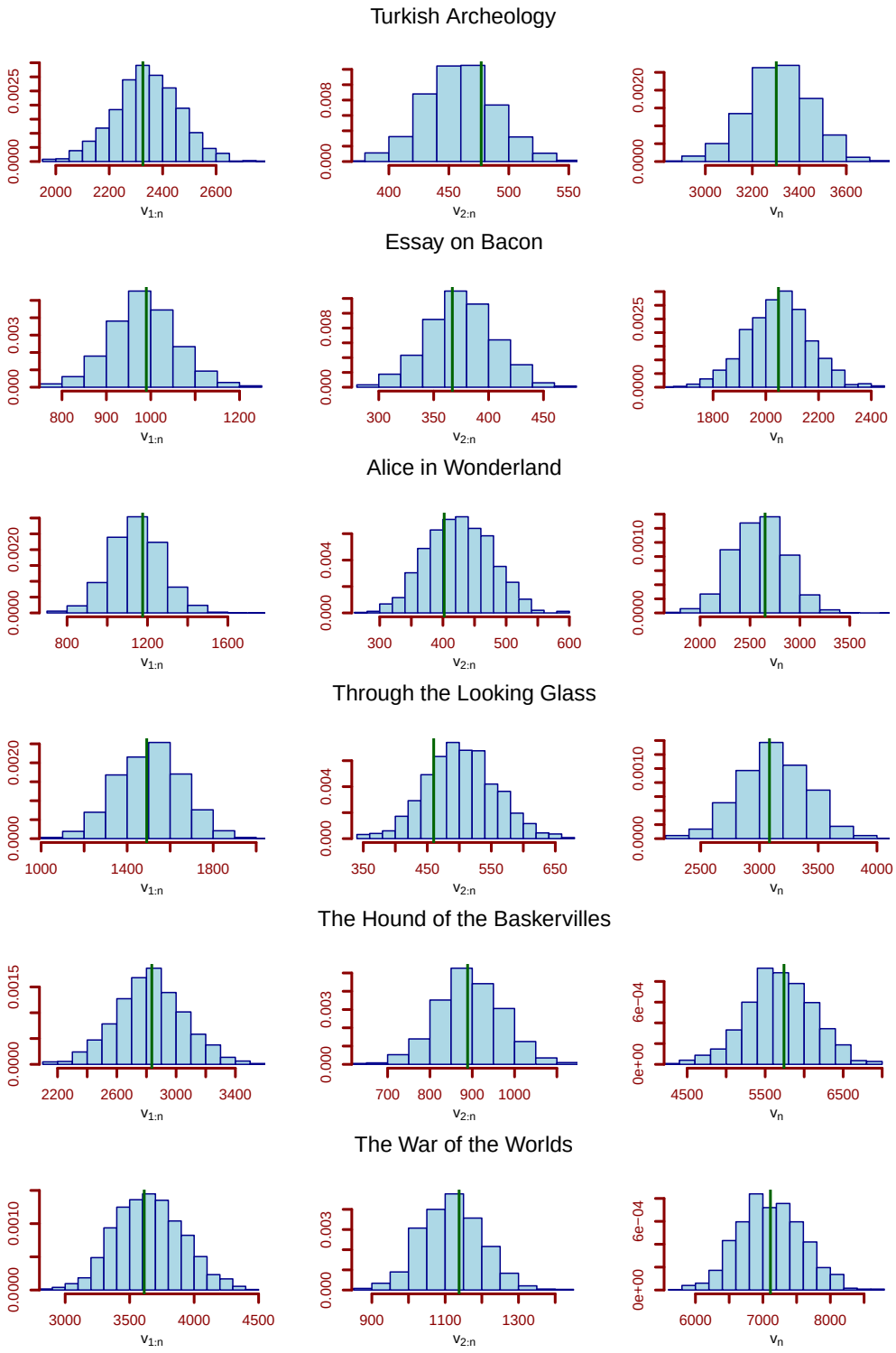


Figure 2.8: Observed value and sample of 10000 observations from the posterior predictive distribution of $v_{1:n}$, of $v_{2:n}$ and of v_n under the IG-Truncated Poisson model with independent Gamma(.001, .001) priors for b and c.

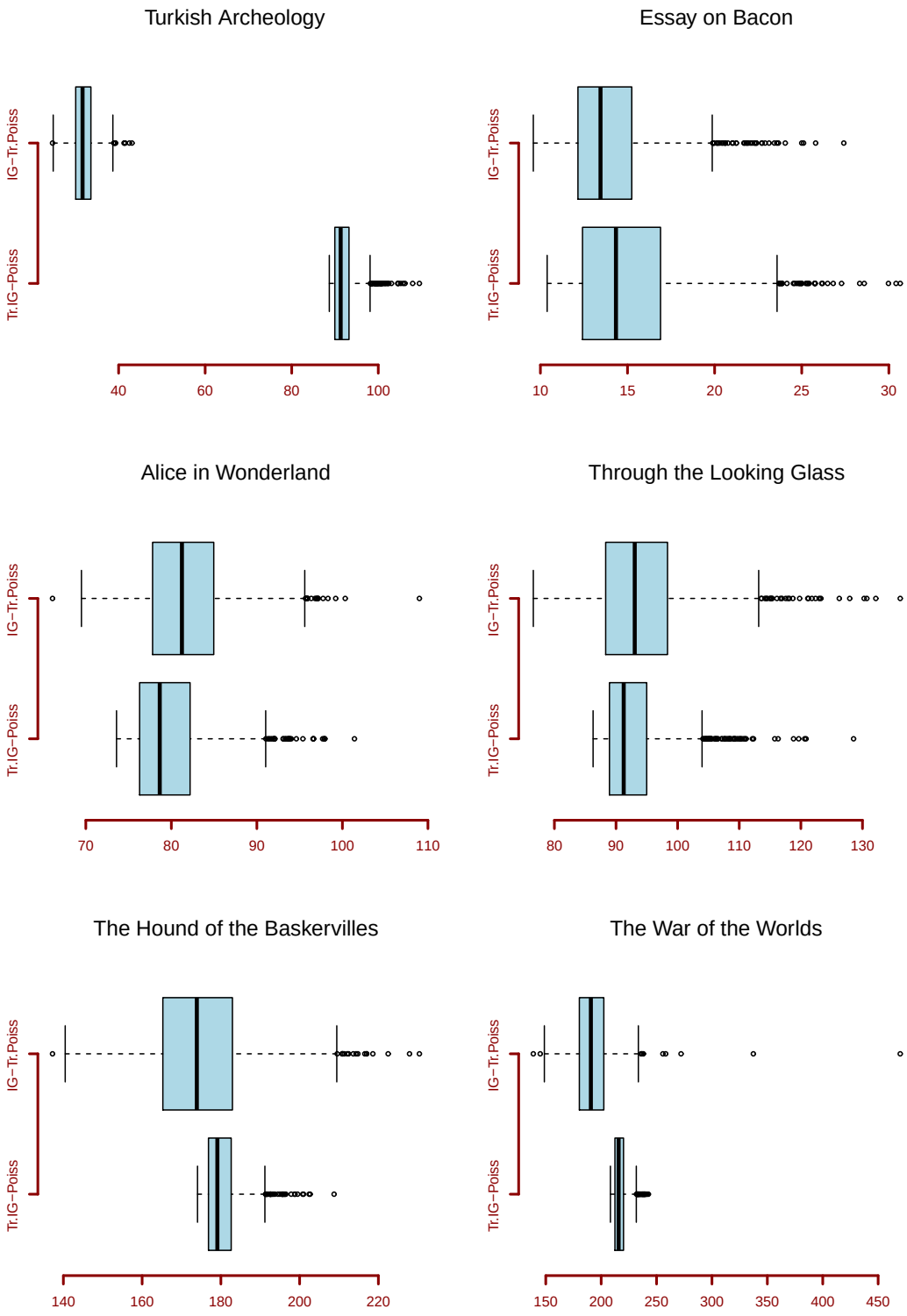


Figure 2.9: Box-plots of samples of 10000 observations from the posterior distribution of $\chi^2(b,c) = \sum_r \epsilon^p_{r:n}(b,c)^2$ under the truncated IG-Poisson and the IG-TruncatedPoisson models with independent Gamma(.001, .001) priors for b and c.

Text	Model	$\hat{b}_{ml}$	$\hat{c}_{ml}$	max lglik	$X^2(\hat{b}_{ml}, \hat{c}_{ml})$	$\hat{b}_{pm}$	$\hat{c}_{pm}$	$X^2(\hat{b}_{pm}, \hat{c}_{pm})$
Turkish	Tr.IG-Poiss	0.	.0013	-3882.65	88.68 (19)	.00056	.0013	89.31 (19)
	IG-TrPoiss	.1458	.0027	-3831.61	35.10 (22)	.1495	.0026	33.92 (22)
E. Bacon	Tr.IG-Poiss	.0836	.0037	-4008.89	17.69 (30)	.0739	.0037	17.75 (29)
	IG-TrPoiss	.2228	.0038	-4008.86	18.91 (30)	.2169	.0040	19.53 (30)
Alice	Tr.IG-Poiss	.0229	.0095	-6281.12	85.85 (59)	.0195	.0097	86.83 (59)
	IG-TrPoiss	.0734	.0098	-6283.07	90.56 (59)	.0702	.0106	85.19 (60)
Through	Tr.IG-Poiss	.0119	.0089	-6887.62	82.56 (61)	.0100	.0091	83.32 (61)
	IG-TrPoiss	.0635	.0097	-6887.45	88.76 (61)	.0645	.0094	85.12 (61)
Hound	Tr.IG-Poiss	.0068	.0057	-12445.73	181.26 (89)	.0057	.0058	186.03 (89)
	IG-TrPoiss	.0515	.0064	-12437.07	175.66 (88)	.0524	.0063	176.43 (88)
War	Tr.IG-Poiss	.0061	.0038	-14654.54	216.11 (90)	.0048	.0039	216.07 (90)
	IG-TrPoiss	.0598	.0044	-14631.83	188.88 (90)	.0592	.0045	188.62 (90)

Table 2.2: Maximum likelihood estimate, $(\hat{b}_{ml}, \hat{c}_{ml})$, and posterior mode, $(\hat{b}_{pm}, \hat{c}_{pm})$, maximum of the log-likelihood function, and $X^2(\hat{b}, \hat{c})$ goodness of fit test statistics for the posterior mode and maximum likelihood fits, under the truncated IG-Poisson and the IG-Truncated Poisson models with independent Gamma$(.001, .001)$ priors for b and c. Between brackets, the number of categories that intervene in the computation of $X^2(\hat{b}, \hat{c})$.

The large amount of information in word frequency count sample data from texts with more than 10000 words would over-ride the information that one might want to incorporate into the analysis through informative priors making use of substantive information about literary style. That is why instead of requiring more informative priors, a more precise Bayesian analysis of word frequency counts in longer texts requires that it be based on more flexible three parameter Poisson mixture models with mixing distributions that better adapt to the typical word frequency distributions of the vocabulary of most authors.

The first candidate that comes to mind for that is the three parameter zero truncated generalized inverse Gaussian-Poisson model considered in Sichel (1975, 1986a). A Bayesian analysis based on this model would be very convenient computationally speaking, because the generalized inverse Gaussian distribution is a conjugate prior for the Poisson model. A different Bayesian analysis for word frequency count data from texts with more than 10.000 words could be based on the zero truncated version of the three-parameter Tweedie-Poisson mixture model first considered by Gerber (1991) and Hougaard et al. (1997).

One nice feature of both the extended approach based on the generalized inverse gaussian mixing model as well as of the one using the Tweedie mixing model is that both mixing

models include the gamma and the inverse gaussian models as special cases. Hence by resorting to either one of these three-parameter Poisson mixture models one can always test wether the simpler negative binomial or inverse gaussian-Poisson models provide a good enough fit for any particular word frequency count data set analyzed.

Under either one of these extended approaches we recommend the use of the three-parameter models obtained by first mixing the Poisson model and then truncating it, which generalize the approach in Section 2.3, instead of the models obtained by first truncating the Poisson model at zero and then mixing it as in Section 2.4. Thanks to the flexibility gained through the additional parameter it is expected that one will obtain good fits for counts from long texts in either case, but using models that mix first and truncates later allows one to estimate the frequency distribution in the population, which is not the case if one uses models truncating first and mixing later.

Even though the usefulness of the bayesian analysis of frequency count data using Poisson mixture models, with a focus on the use of the mixing distribution, has been illustrated in the context of the analysis of word frequency count data in stylometry, everything can be trivially extended to the analysis of frequency of frequency data in many other fields. In particular, this type of analysis should be very useful when modelling species frequency count data in ecology, with the goal of learning about the species distribution in the population of an ecosystem, and in particular about the size, evenness and diversity of that population.

Chapter 3

Classification of Literary Style that Takes Order into Consideration

The statistical analysis of the heterogeneity of the style of a text often leads to the analysis of contingency tables of ordered rows. When multiple authorship is suspected, one can explore that heterogeneity through either a change-point analysis of these rows, consistent with sudden changes of author, or a cluster analysis of them, consistent with authors contributing exchangeably, without taking order into consideration. Here an analysis is proposed that strikes a compromise between change-point and cluster analysis by incorporating the fact that parts close together are more likely to belong to the same author than parts far apart. The approach is illustrated by revisiting the authorship attribution of *Tirant lo Blanc*.

3.1 Introduction

The statistical analysis of literary style has often been used to settle authorship attribution problems both in the academic as well as in the legal context. Early work used word length and sentence length to characterize literary style. Other characteristics widely used for this purpose have been the proportion of nouns, articles or adjectives, the frequency of use of function words, which are independent of the context, and the diversity of the vocabulary used by the author. As a consequence, data in this context is almost always categorical.

In the particular case where one suspects that there might be more than one author, one typically carries out an heterogeneity analysis of the style of the text or corpus of texts

after splitting it down into smaller pieces. Under most of the stylometric characteristics listed above, that leads to the analysis of a contingency table that will often have ordered rows, with each row corresponding to a different piece of the text or corpus, and each column corresponding to the counts of a given category, like of a function word or words or sentences of a given length.

One approach to that problem is through single change-point analysis, assuming that the ordered rows share style and hence the same distribution all the way up to a given point of the row sequence, where the author changes and hence the style and that distribution changes and stays the same for the remaining sequence of rows in the table. The goal in that type of analysis is estimating both the change-point, as well as the before and after the change-point distributions that help characterize the differences in style between authors. This naturally generalizes to multiple change-point analysis, and it is useful in settings where one can assume that the change of author has been sudden.

An alternative approach is through cluster analysis, also recognized as unsupervised classification, which consists on partitioning the rows of the table into groups that are more homogeneous than the whole and could be sharing the same style, without imposing any order restriction when forming the groups. That approach can be implemented based on finite mixture models and it is useful when authors can be assumed to be intervening exchangeably.

Between change-point analysis that force all consecutive observations except the ones at change-points to belong to the same group, and cluster analysis, that assign observations to groups without taking order into consideration, there is a whole range of analysis that incentive but do not force consecutive observations to belong to the same group. That fits well the authorship attribution settings where one is willing to assume that consecutive parts are more likely to belong to the same author than parts that are far apart.

Here one such analysis is proposed based on an extension of the finite mixture models that incorporate the fact that the role of authors could be changing along the text. By letting neighboring observations be related, the model will also capture the correlation that one expects to find as a consequence of the way the writing process works.

Most of the alternative classification methods that are used in the literature of authorship attribution and of the analysis of the heterogeneity of literary style assume data to be continuous, when in practice most of the time data is categorical. We avoid that continuity assumption. Furthermore, the usual classification methods employed by the authorship attribution literature use ad hoc heuristic partitioning algorithms that tend to be easy to apply and work well, but do not allow one to assess cluster uncertainties

and do not provide rigorous inference based methods to allocate individual observations to clusters, (see, e.g., Kaufman and Rousseeuw, 1990, Gnanadesikan, 1997, or Gordon, 1999).

Instead, in this manuscript Bayesian model based clustering approaches are adopted, under which observations are assumed to come from one of two sub-populations, each with a distinctive distribution. These approaches provide a complete probabilistic framework assuming a finite mixture model under which observations (texts) belonging to the same cluster (author) have the same distribution, and then estimating the mixed distributions and assigning observations to these distributions. Each one of the two distributions involved in the mixture characterize each one of the two styles. Model based approaches simultaneously group objects and estimate the distribution of each group, and that avoids the biases appearing whenever these two stages are tackled separately.

Model based Bayesian methods also have the advantage over the usual heuristic classification methods of providing a measure of the uncertainty in the allocation of individual observations into clusters, and of casting the decision of the number of clusters (authors) as a statistical testing problem. Good introductions to Bayesian and non Bayesian model based classification methods can be found in Bock (1996), McLachlan and Peel (2000) and Fraley and Raftery (2002).

To illustrate our novel approach, the authorship attribution problem of *Tirant lo Blanc* will be revisited by analyzing the *word lengths* and the use of *function words* in its chapters, and the results will be compared with the ones of the change-point and cluster analysis of this data carried out in Giron, Ginebra and Riba (2005).

The paper is organized as follows. Section 3.2 presents the authorship attribution problem that will be used to illustrate the method and motivate its need. In Section 3.3 the model proposed is presented and compared with the multinomial change-point and cluster models. In Section 3.4 the results of the analysis for *Tirant lo Blanc* is presented, and in Section 3.5 possible extensions are discussed.

3.2 Description of the authorship problem

Tirant lo Blanc is a chivalry book written in catalan, hailed to be "the best book of its kind in the world" by Cervantes in El Quixote, and considered by many to be the first modern novel in Europe, (see, e.g., Vargas Llosa, 1991, 93). The main body of the book was written between 1460 and 1464, but it was not printed until 1490, and there has been a long lasting debate around its authorship, originating from conflicting information in

its first edition.

Where in the dedicatory letter at the beginning of the book it is stated that "*So that no one else can be blamed if any faults are found in this work, I, Joanot Martorell, take sole responsibility for it, as I have carried out the task singlehandedly*," in the colophon at the end of the book it is stated that "*Because of his death, Sir Joanot Martorell could only finish writing three parts of it. The fourth part, which is the end of the book, was written by the illustrious knight Sir Martí Joan de Galba. If faults are found in that part, let them be attributed to his ignorance.*" Over the years, experts have split between the ones defending the existence of a single author for all its 487 chapters, in line with the dedicatory letter, and the ones backing a change of author somewhere between chapters 350 and 400, in line with the colophon. For a detailed overview of this debate, see Riquer (1990).

It is well accepted by all medievalists that the main (and maybe single) author, Joanot Martorell, died in 1465, and did not start work on the book before 1460, and that if there were any additions, they would be close to the end of the book and made by the second author much later, when the book was printed in 1490. Neither Martorell nor the candidate to be the book finisher left any other texts comparable with this one.

An analysis of the diversity of the vocabulary carried out in Riba and Ginebra (2006) finds that it becomes significantly less diverse after chapter 383. Giron et al (2005) carried out a multinomial change-point analysis and a multinomial two-cluster analysis based on word lengths and on the frequency of words that do not depend on context, called function words; under both characteristics a stylistic boundary is detected between chapters 371 and 382, apparently with a few chapters misclassified by that boundary. Section 3.1 describes and motivates these two types of analysis. As in these previous studies, here the edition of *Tirant lo Blanc* by Riquer is used; after excluding from consideration the titles of chapters, the quotations in latin and the chapters with less than 200 words, that leads to the analysis of a total of 398242 words split down into 425 chapters.

The literature on the statistical analysis of style characterized through word length and through the use of function words is far too large to be covered in detail here. Early uses of word length can be found for example in Mendenhall (1887), Mosteller and Wallace (1984), Brinegaar (1963), Bruno (1974), Williams (1975), Morton (1978), Smith (1983) and Hilton and Holmes (1993). Early uses of function words can be found in some of these references as well as in Burrows (1987, 92), Holmes (1985, 92), Binongo (1994) or Oakes (1998). Function words are proven to be more sensitive than word length when trying to tell authors apart.

Word length counts												
Chapter	1	2	3	4	5	6	7	8	9	10+	N_i	$\overline{wl}_i$
1	21	59	44	19	33	20	16	17	9	17	285	4.47
2	53	113	80	49	52	33	28	36	16	16	476	4.14
...	...	...	...	...	...	...	...	...	...	...	...	...
487	48	49	62	53	41	36	21	9	16	13	348	4.20
Function word counts												
Chapter	e	de	la	que	no	l	com	molt	és	jo	si	dix
1	12	15	9	8	1	7	2	1	6	0	3	0
2	26	28	19	9	3	2	3	8	3	1	3	1
...	...	...	...	...	...	...	...	...	...	...	...	...
487	29	13	8	10	2	10	3	9	0	0	0	0

Table 3.1: Part of the 425×10 table of word length counts in chapters of more than 200 words of *Tirant lo Blanc*, and of the 425×12 table of counts of twelve function words in them. N_i is the total number of words and $\overline{wl}_i$ is the average word length. Authors will provide the full data set to anyone requesting it.

In the example of *Tirant lo Blanc* the analysis of word length leads to the analysis of the 425×10 table of word length counts partially presented in Table 3.1, and the analysis of the twelve function words used in Giron et al. (2005) leads to the analysis of the 425×12 table of function words partially presented in that table. These twelve function words were chosen in that paper by first doing a change-point and a cluster analysis of the chapters of the book based on the 25 most frequent words, and then selecting the subset of these words that best discriminated between the estimated two groups of chapters.

If the book had been written by a single author, one might expect the proportion of words of each length and the frequency of use of each function words to be similar in all chapters. As a consequence, one would expect that once taken into account the fact that chapters have different lengths, all the rows in each one of the two sub-tables of Table 3.1 would have similar distributions. If instead, the distribution of these rows either changed suddenly or kept switching back and forth between two different distributions, it could indicate the existence of a second author that either took over at some point and completed the book, or contributed chapters all over the book.

Figure 3.1 presents the sequence of the proportions of words of each length in each chapter, the sequence of the average word length per chapter and the sequence of the ratio between the number of long words, (with six or more letters), and of short words, (with less than six letters). Note that, for example, the average word length and the proportion of single lettered words and of ten or more lettered words seems to be larger at the end of the book. Figure 3.2 presents the sequence of frequencies of the twelve

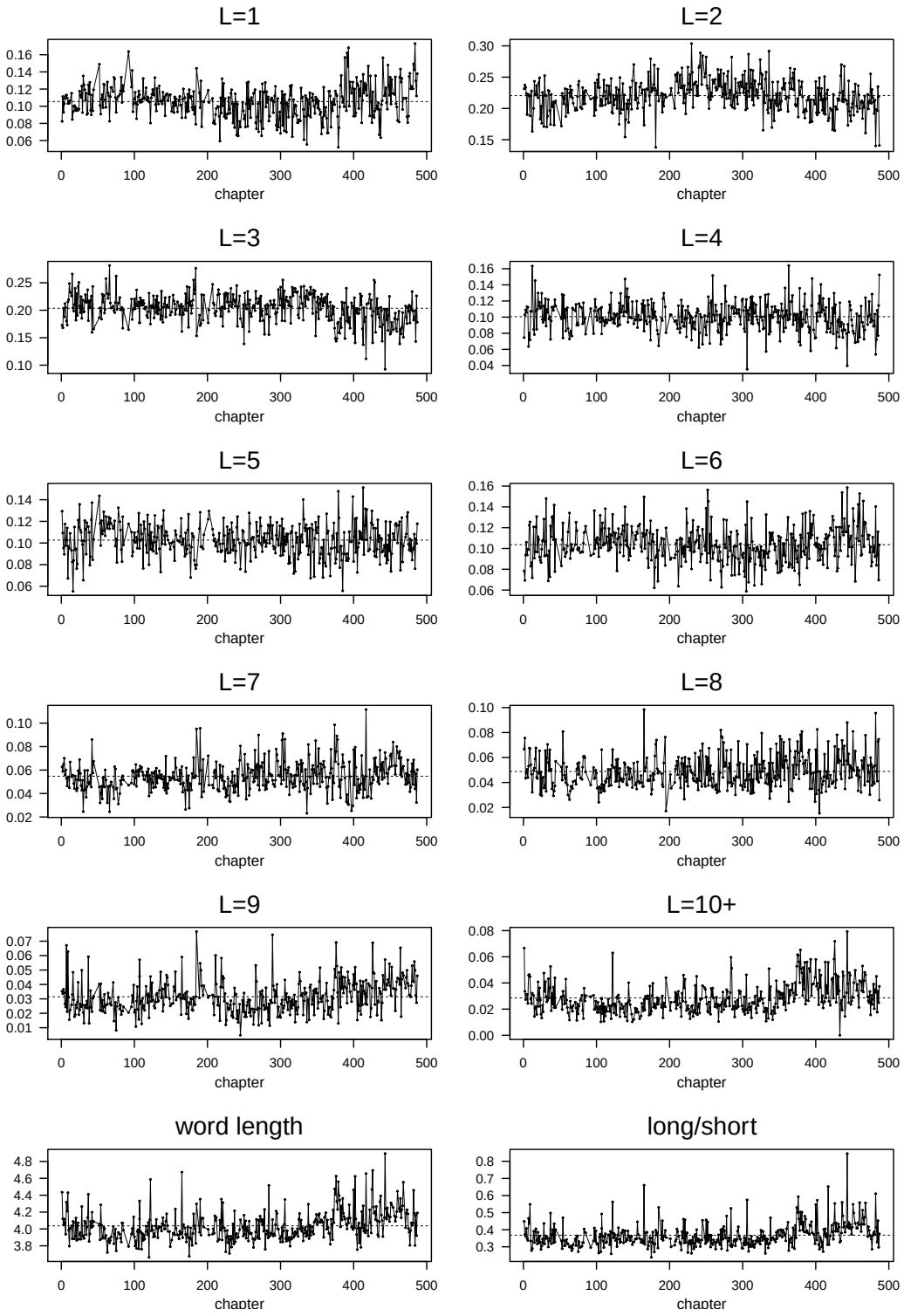


Figure 3.1: Sequence of proportion of words of each length in each chapter of *Tirant lo Blanc*, with $L = l$ meaning words of l characters, sequence of average word length, and sequence of the ratio between the number of long words and of short words in them.

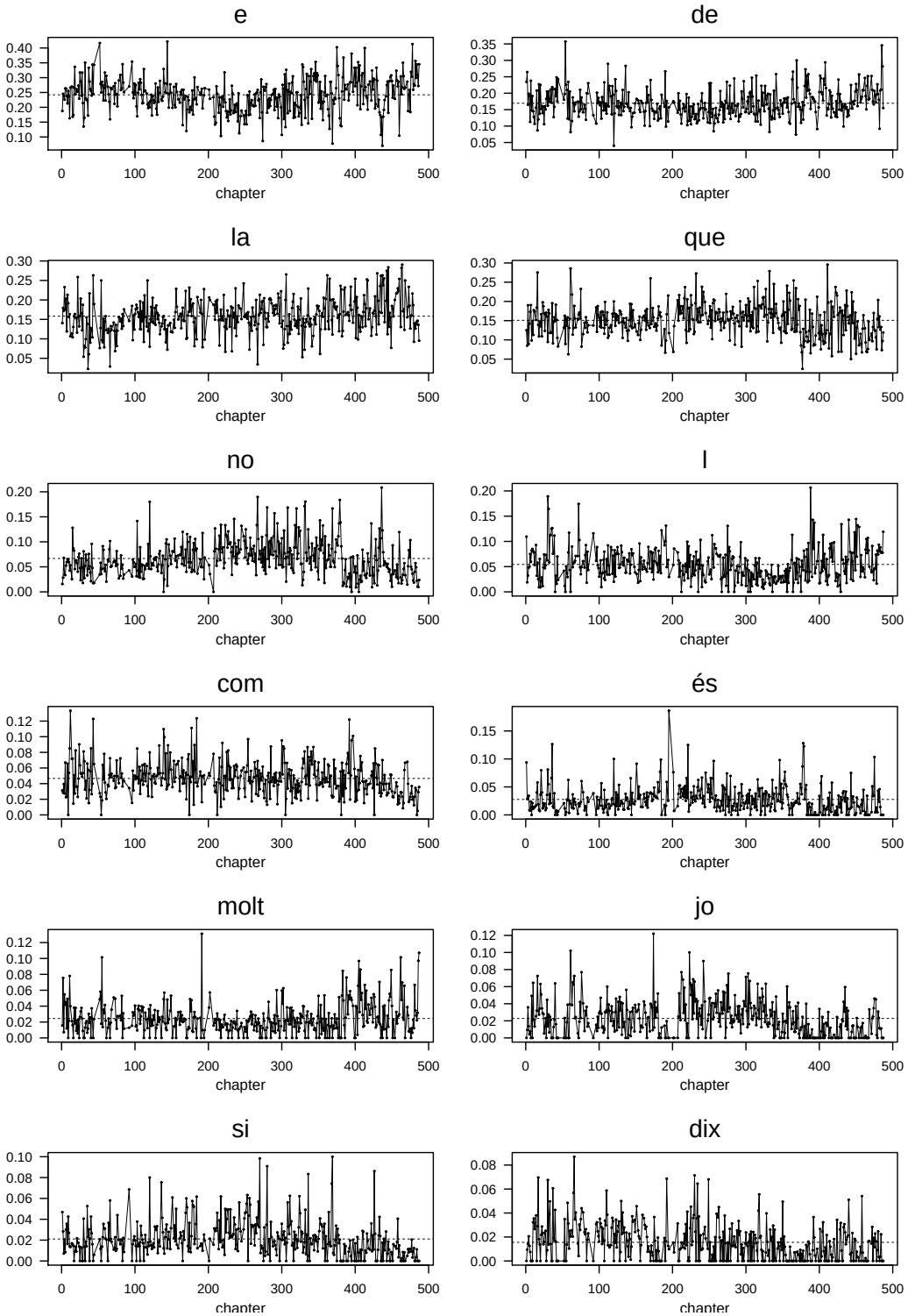


Figure 3.2: Frequency of appearance in the chapters of *Tirant lo Blanc* of the twelve function words used in the analysis.

function words selected. Note that there is also a clear shift in the level of use of words like *e, que, no, l, molt, jo* or *dix* towards the end of the book. What is found in both figures might be consistent both with the existence of two authors and a single change-point, as well as with the existence of a second author filling in material mostly at the end of the book.

In some instances, one might explain changes in style through differences in chronology or topic, specially when one is dealing with works that were written during a long span of time. In our case though it is known that the main author (single author according to some) of the book worked on the book during a short span of time, shortly before his death, and therefore in our example differences in style should not be attributed to breaks in writing. That the estimated changes in style do not coincide with shifts in topic needs to be checked after the chapters are classified according to style.

The three models considered next assess whether the observations in these sequences can be adequately classified into two different populations, each corresponding to a different style. The first model assumes that the change happens once suddenly, the second model assumes that the two styles alternate exchangeably all over the text, and the third model strikes a compromise somewhere in between.

3.3 Description of the models

For each chapter in the book (or part in the corpus of texts), i with $i = 1, \ldots, n$, one has a vector valued categorical observation, $y_i = (y_{i1}, \ldots, y_{ik})$, where k denotes the number of categories of the stylistic characteristic. In our example, y_i will be the ten dimensional vector of word length counts in the i-th chapter, presented as the i-th row in the first sub-table of Table 3.1, and the twelve dimensional vector of frequency counts of the function words selected in that chapter, presented as the i-th row in the second sub-table. The set of all the rows in each sub-table will be denoted by $y = (y_1, \ldots, y_n)$.

Under all the three models considered next, the i-th row of the table, y_i, will always be assumed to be multinomially distributed, $\text{Mult}(N_i, \theta_i)$, where $N_i = \sum_j^k y_{ij}$ denotes the i-th row total and hence the total number of words considered in that row, and where $\theta_i = (\theta_{i1}, \ldots, \theta_{ik})$ is such that $\sum_{j=1}^k \theta_{ij} = 1$, with θ_{ij} being the probability of the j-th category for the i-th row. In our example k will be ten for the first table of word lengths and twelve for the second table of function words. Thus, the rows of these two tables will be considered to form sequences of conditionally independent observations

with probability density function (pdf):

$$\text{Mult}(y_i|N_i,\theta_i) = \frac{N_i!}{\Pi_{j=1}^k y_{ij}!}\Pi_{j=1}^k \theta_{ij}^{y_{ij}}. \tag{3.1}$$

The vector of probabilities, $\theta_i = (\theta_{i1},\ldots,\theta_{ik})$, can be seen as a fingerprint of the style of the author in his texts, because one expects that on average he will use different categories of words with the same relative frequencies. That will lead to the texts by the same author sharing the same set of average probabilities, θ_i. Under that assumption, θ_i characterizes the style of the author while N_i naturally takes into account the text size and therefore the weight to be allocated in the analysis to each row of each table.

If all the chapters belong to the same author and were written at about the same time, it is reasonable to expect that they will share the same style and therefore one would expect the vector of probabilities, θ_i, for all the rows in the two sub-tables considered to stay approximately constant along the whole sequence of 425 chapters. In that case, the rows of these sub-tables could be modeled as a random sample of $\text{Mult}(N_i,\theta)$ distributions.

On the other hand, if one detects a sudden shift in the vector of probabilities, θ_i, through a change-point analysis, that might indicate a sudden change in style and therefore a sudden change of author, of topic, or of writing time. If, instead, one identifies the rows of the tables as belonging to two distinct populations through a cluster analysis, with each population of rows sharing a different vector of probabilities, that might indicate the existence of two different styles and therefore of two different authors intervening more or less exchangeably all along the book. Next, these two settings are modeled probabilistically.

3.3.1 Multinomial change-point and cluster models

In a multinomial single change-point analysis one assumes that $y = (y_1,\ldots,y_n)$ is a sequence of conditionally independent multinomial random variables such that $\theta_i = \theta_b$ for $i \leq r$ and $\theta_i = \theta_a$ for $i > r$, and thus with a probability density function (pdf):

$$p(y|r,\theta_b,\theta_a) = \prod_{i=1}^{r} \text{Mult}(N_i,\theta_b) \prod_{i=r+1}^{n} \text{Mult}(N_i,\theta_a). \tag{3.2}$$

This model assumes that the first r chapters (rows) before the change-point have been written by the first author with a style characterized by the first set of probabilities θ_b, while the remaining set of $n-r$ chapters (rows) after that change-point have been written by the second author with a style characterized by the second set of probabilities θ_a. The goal in change-point analysis is to learn about the change-point, r, as well as about the

before and after the change-point multinomial probabilities, θ_b, θ_a, characterizing the two styles.

As an alternative, in multinomial two-cluster analysis, the n rows of the table, $y = (y_1, \ldots, y_n)$, are considered to be conditionally independent and identically distributed according to a finite mixture of two multinomial distributions, with pdf:

$$p(y|\omega, \theta_1, \theta_2) = \prod_{i=1}^{n} (\omega * \text{Mult}(N_i, \theta_1) + (1 - \omega) * \text{Mult}(N_i, \theta_2)), \tag{3.3}$$

where $\theta_s = (\theta_{s1}, \ldots, \theta_{sk})$ for $s = 1, 2$ determine the distribution of the rows in the s-th cluster, and hence characterize the style in that cluster, and where ω is a weight determining the proportion of rows belonging to the first cluster and hence the probability that any given row will be allocated to that cluster. This model assumes that the chapters (rows) allocated to the cluster 1 were written by an author with a style characterized by the set of probabilities θ_1, while the remaining chapters (rows) allocated to the cluster 2 were written by a different author with a style characterized by θ_2.

To allocate rows into clusters, which is an essential feature in cluster analysis, one has to introduce a vector of unobserved (latent) categorical variables $\zeta = (\zeta_1, \ldots, \zeta_n)$, where ζ_i takes values in $\{0, 1\}$ and is such that $\zeta_i = 1$ when the i-th row belongs to the first cluster and $\zeta_i = 0$ when it belongs to the second cluster. A variable is considered to be latent whenever one can not observe it but is willing to estimate it, very much like one does for a parameter. Here the ζ_i are assumed to be conditionally independent and identically distributed, with $\pi(\zeta_i = 1|\omega) = \omega$ and $\pi(\zeta_i = 0|\omega) = 1 - \omega$. As a consequence the joint pdf for $y = (y_1, \ldots, y_n)$ and $\zeta = (\zeta_1, \ldots, \zeta_n)$ becomes:

$$p(y, \zeta|\omega, \theta_1, \theta_2) = \prod_{i=1}^{n} (\omega * \text{Mult}(N_i, \theta_1))^{\zeta_i} ((1 - \omega) * \text{Mult}(N_i, \theta_2))^{1-\zeta_i}. \tag{3.4}$$

The allocation of rows into clusters can be inferred through point estimates of ζ.

Fitting these multinomial change-point and cluster models through the classical frequentist inference techniques is complicated, specially when it turns to assessing the uncertainty of the estimates of the multinomial probabilities and to estimating ζ. Instead, we adopt the Bayesian inference approach, that requires eliciting a prior distribution on the parameters of the models that summarize the knowledge one has about them, and then updating these distributions in the light of the data. For an introduction to the Bayesian approach to data analysis, see, e.g., Gelman et al. (2013) or Carlin and Louis (2008).

As a prior distribution, one typically assumes by default that the vectors of multinomial probabilities in the change-point analysis, (θ_b, θ_a), and in cluster analysis, (θ_1, θ_2), are

independent and Dirichlet$(a_{s1}, \ldots, a_{sk})$ distributed, with either $s = a, b$ or $s = 1, 2$, and hence with pdf:

$$\pi(\theta_s) = \pi(\theta_{s1}, \ldots, \theta_{sk}) = \frac{\Gamma(\sum_{j=1}^{k} a_{sj})}{\prod_{j=1}^{k} \Gamma(a_{sj})} \theta_{s1}^{a_{s1}-1} \ldots \theta_{sk}^{a_{sk}-1}, \tag{3.5}$$

where $\Gamma(\cdot)$ stands for the Gamma function. Depending on the values chosen for $(a_{s1}, \ldots, a_{sk})$, the prior can go from being very subjective to reflecting vague information about the multinomial vector of probabilities, (θ_b, θ_a) and (θ_1, θ_2). In particular, note that the prior expected value for $\theta_s = (\theta_{s1}, \ldots, \theta_{sk})$ will be $(a_{s1}, \ldots, a_{sk})/(\sum_{j=1}^{k} a_{sj})$, and one can chose the a_{sj} to reflect the fact that one knows that some categories have larger probabilities than others. One can also rely on the fact that the larger $\sum_{j=1}^{k} a_{sj}$, the smaller the prior variances of the probabilities θ_{sj}, and hence the more informative the prior will be about θ_s. In the implementation that follows all the $(a_{s1}, \ldots, a_{sk})$ are set to be equal to $(1, \ldots, 1)$, which corresponds to assuming a uniform distribution on the simplex and hence that $E[\theta_{sj}] = 1/k$ for all j, and that all the possible values for $\theta_s = (\theta_{s1}, \ldots, \theta_{sk})$ are equally likely, but more informative distributions have also been tried. In particular note that in the case of function words the categories are ordered from words appearing more frequently to words appearing less frequently, and hence it is also be natural to chose $(a_{s1}, \ldots, a_{sk})$ such that $a_{s1} \geq a_{s2} \geq \ldots \geq a_{sk}$, which lead to $E[\theta_{sj}]$ being decreasing with j.

As a prior distribution for the change-point, r, in the change-point model, one typically chooses a uniform distribution on $\{1, \ldots, n\}$, which assumes that the change in style could happen anywhere in the book equally likely. Nevertheless, if one suspects that the change-point is more likely to happen in certain chapters than in certain others, one should incorporate that information in a more informative prior.

In the cluster analysis model, as a prior for the cluster weight, ω, which is the probability that any chapter belongs to Cluster 1 and therefore takes values between 0 and 1, one typically assumes it to be Beta(b, c) distributed and independent of (θ_1, θ_2), which is a very flexible family of distributions supported on $[0, 1]$ with pdf:

$$\pi(\omega) = \frac{\Gamma(b + c)}{\Gamma(b)\Gamma(c)} \omega^{b-1}(1 - \omega)^{c-1}, \tag{3.6}$$

where, again, $\Gamma(\cdot)$ stands for the Gamma function. In the implementation (b, c) is set to be equal to $(1, 1)$, which is the same as assuming that ω takes a uniform distribution on $[0, 1]$, and hence that all possible values for ω are equally likely. For more details on the Dirichlet and Beta distributions, see Johnson, Kemp and Kotz (2005) and Johnson, Kotz and Balakrishnan (1997).

Note that beta and Dirichlet probability models are the default Bayesian choices as prior distributions when one needs to model proportions and vectors of probabilities, respectively. We also tried more informative priors, incorporating the fact that the categories in the second sub-table are ordered from more frequent to less frequent function words. More informative priors for r and ω were also tried, but sample sizes are large enough so that data is so much more informative than any of the prior distributions used and hence the posterior distributions were insensitive to the choice of prior distribution. Hence these distributional choices have very limited impact on the results of the analysis presented in Section 3.4. For more technical details on these multinomial change-point and cluster models, see Giron et al. (2005).

3.3.2 Multinomial cluster model with dependence

When carrying out a cluster analysis based on (3.4) one assumes that all rows and corresponding allocation variables, (y_i, ζ_i) for $i = 1, \ldots, n$, are conditionally independent and identically distributed. As a consequence, one is implicitly assuming that the two styles mix exchangeably along the text, without taking into consideration the order in which rows appear, which most often runs against what one anticipates to be happening.

One extension of the finite mixture model in (3.3) that corrects for that, first considered by Fernandez and Green (2002) in the context of Poisson mixtures for spatially indexed data, lets the weights in the mixture vary from row to row, $\omega = (\omega_1, \ldots, \omega_n)$, which leads to:

$$p(y|\omega, \theta_1, \theta_2) = \prod_{i=1}^{n} (\omega_i * \text{Mult}(N_i, \theta_1) + (1 - \omega_i) * \text{Mult}(N_i, \theta_2)), \tag{3.7}$$

where $\omega_i = (\omega_{i1}, \omega_{i2} = 1 - \omega_{i1})$ is such that $0 < \omega_{i1} < 1$, and hence to the rows of the table, $y = (y_1, \ldots, y_n)$, becoming conditionally independent but not identically distributed. As a consequence of that modification, the probability that the i-th row is allocated to the first cluster, ω_i, will be changing from row to row and the set of latent allocation variables, $\zeta = (\zeta_1, \ldots, \zeta_n)$, indicating whether each row belongs to cluster 1 or 2, will be conditionally independent but not identically distributed, with $\pi(\zeta_i = 1|\omega) = \omega_i$ and $\pi(\zeta_i = 0|\omega) = 1 - \omega_i$. The joint pdf of $y = (y_1, \ldots, y_n)$ and $\zeta = (\zeta_1, \ldots, \zeta_n)$ becomes:

$$p(y, \zeta|\omega, \theta_1, \theta_2) = \prod_{i=1}^{n} (\omega_i * \text{Mult}(N_i, \theta_1))^{\zeta_i} ((1 - \omega_i) * \text{Mult}(N_i, \theta_2))^{1-\zeta_i}, \tag{3.8}$$

and the allocation of the i-th row into either one of the two clusters will be done again based on point estimates of ζ_i. The posterior distribution of ω_i is closely related to the one of ζ_i, and it also helps determine the role of the two authors along the text.

A second feature of the basic cluster model in (3.3) that runs against what one anticipates

in most authorship attribution settings is that it does not consider rows (chapters) that are close to be more likely to belong to the same cluster (author) than rows (chapters) that are far apart. Here, certain degree of sequential dependence in chapter authorship is incorporated through a prior structured distribution of the weights, ω_i, making it more likely that rows in nearby locations have more similar allocation probabilities than rows that are located far apart. More specifically, here one will let ω_i be such that its log odds are:

$$\log \frac{\omega_i}{1-\omega_i} = \alpha_i + \beta_i, \tag{3.9}$$

where the α_i's and the β_i's for $i = 1, \ldots, n$ are terms playing a different role each, and are treated as random effects and hence linked by a hierarchical structure that lets their relative contributions be determined by data.

The term α_i is assumed to be conditionally independent and Normal$(\mu_\alpha, \sigma_\alpha^2)$ distributed, and hence with a contribution to the log odds of ω_i that is comparable for all i, thus capturing the global unstructured heterogeneity in ω_i induced by a likely large set of unobserved covariates. The term β_i is assumed to be conditionally independent and Normally distributed, with their mean and variance being equal to $(\beta_{i-1} + \beta_{i+1})/2$ and $\sigma_\beta^2/2$ for $i = 2, \ldots, n-1$, and with mean and variance being equal to β_2 and σ_β^2 for $i = 1$, and being equal to β_{n-1} and σ_β^2 for $i = n$. By relating the mean of β_i, corresponding to the i-th row (chapter) to the values taken by β_{i-1} and β_{i+1} corresponding to the $(i-1)$-th and the $(i+1)$-th rows (chapters), that term captures the local dependence effect that one expects to find when the degree of intervention of the authors shifts smoothly in the book.

The distribution for ω_i chosen here mimics the priors used by the disease mapping literature to obtain spatially smoothed estimates of Poisson means ever since Besag et al (1991) and Mollie (1996). The novelty is that here the prior is used on time and not space indexed data and that it is used to model dependence through the mixing weights of a cluster model and not through the mean parameter of a single cluster distribution. One can think of other ways of inducing sequentially dependent allocations of rows into clusters, but as long as they are flexible enough and use enough information about neighboring observations, they should all lead to similar results.

Fitting this model to the data through classical frequentist inference tools would be extremely difficult, and that is why here again the Bayesian inference approach is adopted. That requires one to chose a prior distribution on the parameters of the model to start with, and then compute the posterior distribution by incorporating the information in the data.

If the prior distributions chosen have little information compared with the information in the data, as it will be the case in our implementation, the choice of prior distribution barely has any influence on the posterior distribution, and hence on the inferences reached. Hence, in that case one can think of the choice of a prior distribution as a default technical step where one only needs to be careful to match the parameter set with the support of the priors chosen.

Here, as a prior distribution for μ_α, the expected value of the α_i, one assumes that it is $Normal(m, s)$ distributed, centered at the value expected for the average of the log odds for ω_i, which in our example will be $m = 0$, and with a large variance, that in our example will be set to be $s = 100$. By choosing a normal distribution with a large variance, one is assuming that one knows very little about the mean of the α_i and hence the inferences about these parameters will be very weakly influenced by the choice of that prior.

The inverse of σ_α^2 and of σ_β^2 are non-negative real valued, and by default they are typically assumed to be Gamma(c, d) distributed, and hence to have a pdf:

$$\pi(\sigma) = \frac{d^c}{\Gamma(c)}\sigma^{c-1}e^{-d\sigma}. \tag{3.10}$$

In the implementation that follows one chooses $c = 1$ and $d = .01$, which correspond to assuming that the distributions for σ_α^2 and for σ_β^2 have large variances, which is the standard choice when one wants to use prior distributions that assume that very little is known about σ. Hence, that choice barely influences the conclusions of the analysis.

As a prior distribution for the multinomial probabilities, (θ_1, θ_2), one assumes that they are independent and with each $\theta_s = (\theta_{s1}, \ldots, \theta_{sk})$ with $s = 1, 2$ having again a Dirichlet$(a_{s1}, \ldots, a_{sk})$ distribution with a pdf as in (3.5). In the actual implementation that follows the $(a_{s1}, \ldots, a_{sk})$ are also set to be equal to $(1, \ldots, 1)$, which corresponds to a reference uniform distribution on the simplex and hence to treating all k categories symmetrically and assuming that all possible values for $\theta_s = (\theta_{s1}, \ldots, \theta_{sk})$ are equally likely. For the details on this default choice as a distribution for (θ_1, θ_2), and for alternative choices that are more informative, we refer to the discussion at the end of Subsection 3.3.1. Even though the model in (3.7) and (3.8) is more general than the one in (3.3) and (3.4), the role played by these parameters is basically the same in both cases.

The whole Bayesian model, including both the statistical model as well as the prior distributions described above, can be found summarized in Table 3.2.

An extensive sensitivity analysis has been carried out by trying priors that incorporated different information about the parameters of the hyper prior and of the multinomial

$$
\begin{gathered}
(y_1,\ldots,y_n)|\theta_1,\theta_2,\zeta \sim \textstyle\prod_{i=1}^{n} \text{Mult}(N_i,\theta_1)^{\zeta_i}\text{Mult}(N_i,\theta_2)^{1-\zeta_i}, \\
(\theta_1,\theta_2) \sim \textstyle\prod_{j=1}^{2} \text{Dirichlet}(a_{j1},\ldots,a_{jk}), \\
(\zeta_1,\ldots,\zeta_n)|(\omega_1,\ldots,\omega_n) \sim \textstyle\prod_{i=1}^{n} \text{Bernoulli}(\omega_i), \\
\omega_i = e^{\alpha_i+\beta_i}/(1+e^{\alpha_i+\beta_i}), \quad i=1,\ldots,n \\
(\alpha_1,\ldots,\alpha_n)|\mu_\alpha,\sigma_\alpha^2 \sim \textstyle\prod_{i=1}^{n} \text{Normal}(\mu_\alpha,\sigma_\alpha^2) \\
\beta_1|\beta_2,\sigma_\beta^2 \sim \text{Normal}(\beta_2,\sigma_\beta^2) \\
\beta_i|\beta_{i-1},\beta_{i+1},\sigma_\beta^2 \sim \text{Normal}((\beta_{i-1}+\beta_{i+1})/2,\sigma_\beta^2/2), \qquad i=2,\ldots,n-1, \\
\beta_n|\beta_{n-1},\sigma_\beta^2 \sim \text{Normal}(\beta_{n-1},\sigma_\beta^2), \\
\mu_\alpha \sim \text{Normal}(m,s) \\
\sigma_\alpha^{-2} \sim \text{Gamma}(c_\alpha,d_\alpha) \\
\sigma_\beta^{-2} \sim \text{Gamma}(c_\beta,d_\beta)
\end{gathered}
$$

Table 3.2: Bayesian multinomial two-cluster model with dependence.

parameters. Here it is also found that data is so much more informative than the priors used, that the posterior distribution barely changes by changing the prior choices.

The posterior distribution for the parameters of these models are too complex to be computed analytically. Instead of that, to update the model and simulate from it the WinBugs MCMC implementation has been used (see, Lunn et al. 2000). The convergence of the chains has been assessed through the visual inspection of the sample traces and the monitoring of various diagnostic measures. The authors will provide the code and the data of the example to anyone that requests them.

3.3.3 Selection of the number of authors and testing

Under each one of the three models contemplated above, that is, the change-point model in (3.2), the cluster model in (3.4), and the cluster model with dependence in (3.8), one needs to chose between the single author (style) case and the two authors (styles) case. In all these situations, that issue can be posed as a choice between two models, and hence can be answered through a formal statistical hypothesis test.

In the change-point model, for example, one needs to test whether $r = n$ (single author) or $r \neq n$ (two authors), and in the basic cluster model, one needs to test whether $\omega = 1$ (single author) or $\omega \neq 1$ (two authors). Resorting to a Bayesian analysis has the advantage that one can select the model with the largest posterior probability. The posterior probability that the M_r model is the one generating the data is:

$$P(M_r|y) = \frac{P(M_r)P(y|M_r)}{\sum_{r=0}^{S} P(M_r)P(y|M_r)}, \tag{3.11}$$

where $P(M_r)$ is the prior probability of model r and where $P(y|M_r)$ is the marginal likelihood of M_r. When one is only interested in comparing models M_r and M_s, one resorts to:

$$\frac{P(M_r|y)}{P(M_s|y)} = \frac{P(M_r)}{P(M_s)} \frac{P(y|M_r)}{P(y|M_s)}. \tag{3.12}$$

In general, one will select the model with the largest posterior probability; when each model is considered equally likely a priori, the larger the marginal likelihood of a model, $P(y|M_S)$, the more attractive that model.

Most often, computing $P(y|M_S)$ exactly is too complicated to be attempted in practice, but one can estimate $P(y|M_S)$ through the MCMC simulations used to update the model, (see, e.g., Gelfand and Dey 1994 or Raftery and Newton, 1995), which is what will be used next to choose between single and multiple author hypotheses.

3.4 Results of the analysis of Tirant lo Blanc

Here the word length and the function word data in Table 3.1 is analyzed using the two-cluster model with dependence just presented, and the result of that analysis is compared with the results obtained using the change-point and basic cluster model in Section 3.3.1.

A single change-point analysis based on the model in (3.2) leads to a posterior distribution of the change-point, r, highly concentrated around Chapter 371 for the word length data, and highly concentrated around Chapter 382 for the function word data. That explains why the top panels of Figures 3.3 and 3.4 assign chapters to authors the way they do. Under both the word length as well as under the function words case, one finds that the posterior probability of the single author (no change-point) model is basically zero; As a consequence, Subsection 3.3.3 indicates that one should reject the single author hypothesis. Under both tables, the sequence of rows clearly have a change in distribution, indicating a change in style, somewhere between Chapters 371 and 382 of the book.

Under both the basic cluster model in (3.4) as well as the cluster model with dependence in (3.8), the posterior probability that y_i belongs to the first cluster, $E[\zeta_i|y]$, can be estimated through the MCMC simulated samples. Given that $E[\zeta_i|y]$ can be interpreted to be the probability that the i-th chapter belongs to cluster (author) 1, it is natural to allocate that chapter to cluster (author) 1 whenever $E[\zeta_i|y] > .5$, and to allocate that chapter to cluster (author) 2 otherwise.

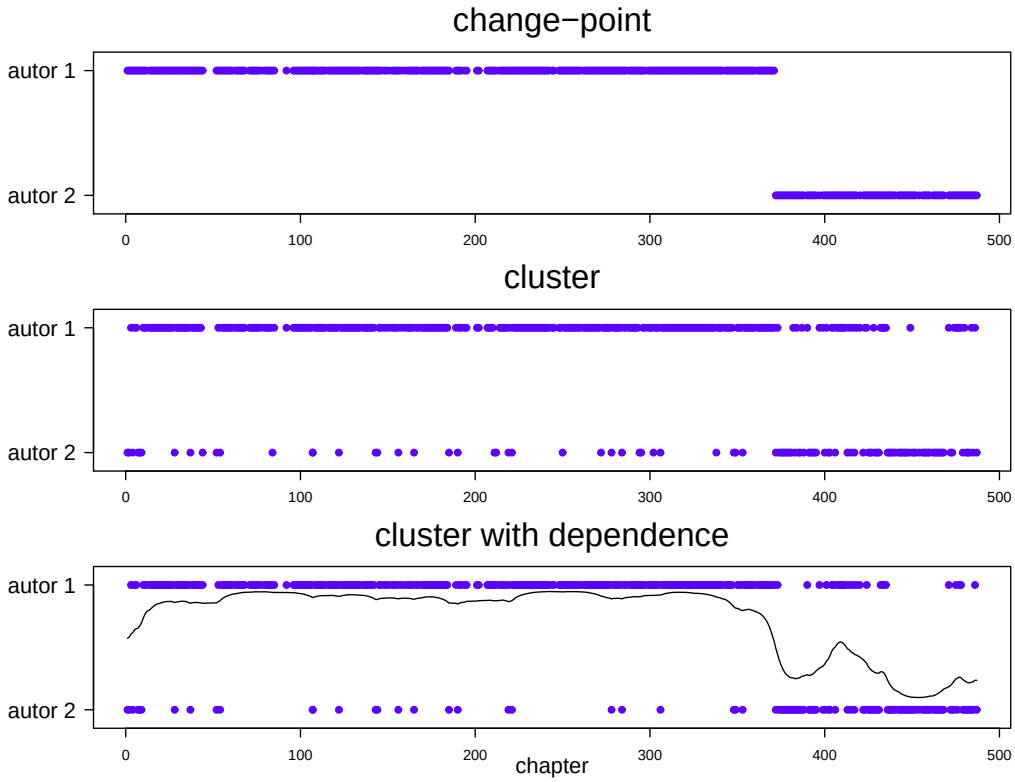


Figure 3.3: Chapter classification for word length under the single change-point model and under the two-cluster models with and without dependence. The curve on the bottom panel is the posterior expectation of ω_i, which helps describe the role of author 1 in that part of the book.

The second panel in Figures 3.3 and 3.4 presents the classification of chapters into authors according to this rule under the basic cluster model in (3.4). Using word length data, Figure 3.3 indicates that 319 chapters are attributed to the first author, which represents 75.06% of the 425 chapters considered, and only 75 chapters are classified differently than through the change-point model, of which 38 are attributed to the second author but are located before chapter 371, while 37 are attributed to the first author but are located after that chapter. For the function word data, in Figure 3.4 one finds 304 chapters attributed to the first author, which represents 71.53% of the total; in this case, 59 chapters are attributed to the second author but located before chapter 382, while 32 chapters are attributed to the first author but located after it. When one tests the single author hypothesis against the double author hypothesis, using the idea described in Subsection 3.3.3, one finds that under both tables the probability of the two-authors hypothesis is almost one, and therefore one again clearly rejects the single author hypothesis.

The third panel in Figures 3.3 and 3.4 presents the chapter classification based on the $E[\zeta_i|y]$ under the cluster model with dependence in (3.8). The classification under this

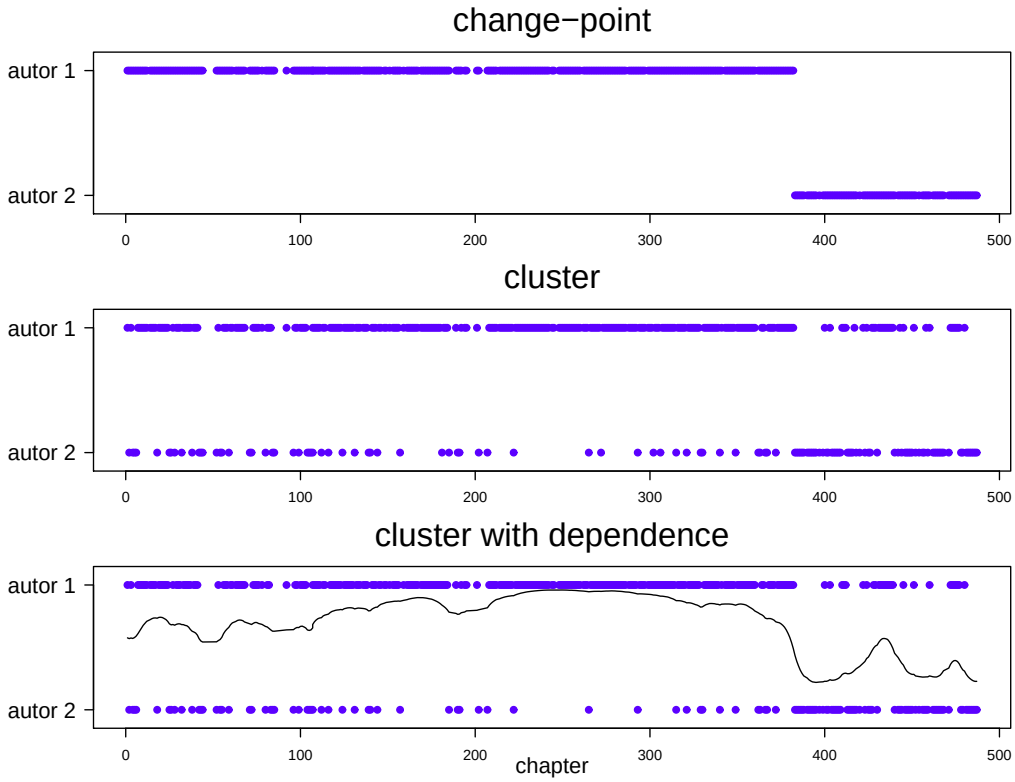


Figure 3.4: Chapter classification for the function word data under the single change-point model and under the two-cluster models with and without dependence. The curve on the bottom panel is the posterior expectation of ω_i, which helps describe the role of author 1 in that part of the book.

more sophisticated model is similar to the one obtained through the basic cluster model, and the corrections are in the direction of making the classification more similar to the one obtained through the change-point model. For the word length data here only 23 chapters are classified differently than through the basic cluster model, with only 27 chapters located before chapter 371 and yet attributed to the second author, and only 25 chapters located after that chapter and yet attributed to the first author. Using function word data only 9 chapters are classified differently than through the basic cluster model, with 56 chapters being attributed to the second author but located before chapter 382 and 28 chapters being located after that chapter but attributed to the first author.

According to the model with dependence, the chapters located before the $371 - 382$ change-points that are consistently allocated to Author 2 instead of Author 1 under both stylometric characteristics are chapters $2, 4, 28, 52, 54, 107, 144, 185, 190$ and 349 while the chapters located after these change-points that are consistently allocated to Author 1 are $410 - 412, 424, 432 - 435, 475$ and 477.

The posterior expected value of ω_i, in the third panel of Figures 3.3 and 3.4, also helps describe the role of each author along the book. Whether $E[\omega_i|y]$ is larger or smaller than .5 serves as an indication of which author plays the main role in that part of the book. Note the close agreement between $E[\omega_i|y]$ and the classification of chapters into authors according to the change-point model. This tool is unavailable under the basic two-cluster model.

Once the existence of two authors is settled and chapters are allocated into each one of the styles according to each one of the models, the question arises as to how do the components in $\theta_i = (\theta_{i1}, \ldots, \theta_{ik})$ change when one switches from one style to the other according to each one of the models. To address that, Figures 3.5 and 3.6 plot a sample of the posterior distribution of $\log(\theta_{bj}/\theta_{aj})$ under the change-point model in (3.2) and of $\log(\theta_{1j}/\theta_{2j})$ under the cluster models in (3.4) and in (3.8). Note the high degree of agreement between the three models, and specially between the cluster models with and without dependence, that follows from the agreement in the way these models allocate chapters into styles.

Figure 3.5 indicates that two, three, four and five lettered words are more abundant in the style of the author writing most of the book, while one, six, seven, eight, nine and ten or more lettered words are more abundant in the style of the author writing mostly at the end of the book. Figure 3.6 indicates that words *que, no, com, és, jo, si* and *dix* are more abundant in the part of the book written by the main author, while *e, de, la, l* and *molt* are more abundant in the parts of the book written by the second author.

3.5 Final comments

The statistical analysis identifies a change in style near chapters 371–382, with a few chapters being misclassified by that change-point. That agrees with the boundary detected in chapter 383 through the analysis of the diversity of vocabulary in Riba and Ginebra (2006), and it is in line with the hypothesis supported by experts attributing more credibility to the colophon of the book than to its dedicatory letter.

The change-point model, (3.2), is very strict in that it assumes that all consecutive chapters (except the r-th and the $(r+1)$-th chapters) belong to the same author, and that will not adapt to most practical settings. The cluster model that does not allow for dependence, (3.4), is more flexible in that it does not take order into consideration when allocating chapters to authors, and that will also fail to model many practical instances. Instead, the cluster model with dependence proposed in (3.8) strikes a compromise somewhere in between, allowing for neighboring chapters to be more likely by

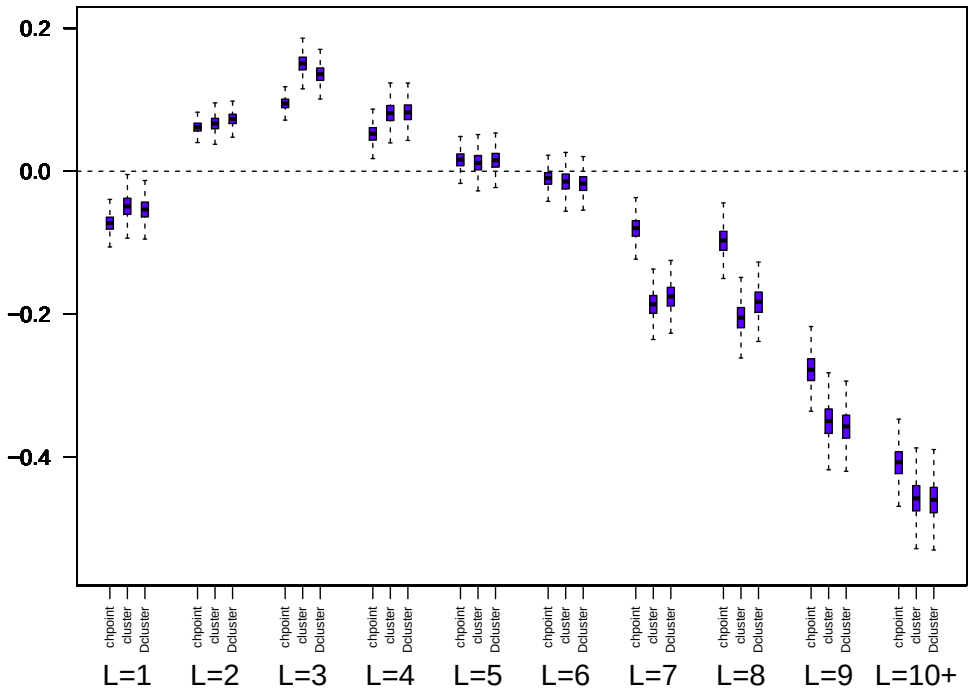


Figure 3.5: Boxplot of a sample of the posterior distribution of $\log(\theta_{bj}/\theta_{aj})$ under the change-point model, in (3.2), and of $\log(\theta_{1j}/\theta_{2j})$ under the clusters models with and without dependence, in (3.4) and (3.8), for the word length data.

the same author without imposing the restriction that they have to be so. Hence the model in (3.8) has the advantage of fitting better the scenarios typically faced in many authorship attribution settings.

As an alternative to the cluster model based on a mixtures of two multinomial models considered here, one could have started with a more flexible framework under which all rows belonging to the same cluster where multinomially distributed with a θ_i that varied from row to row, but with all these θ_i sharing a common distribution. If in particular one assumes that these θ_i are Dirichlet distributed, one would end up basing the analysis on mixtures of two Dirichlet-multinomial models and hence adding two parameters determining the degree of heterogeneity of the multinomial parameters in each cluster. We have tried that approach, but carrying out predictive checks to validate models has lead us to conclude that this type of data does not require these more sophisticated models.

Even though the presentation has focused on the use of word length and function words,

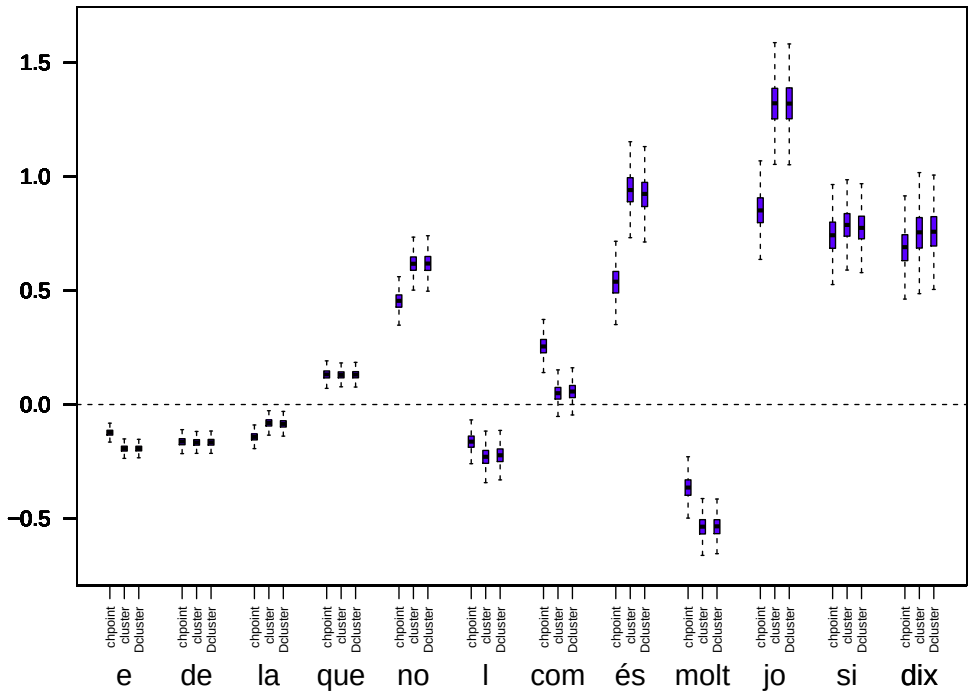


Figure 3.6: Boxplot of a sample of the posterior distribution of $\log(\theta_{bj}/\theta_{aj})$ under the change-point model, in (3.2), and of $\log(\theta_{1j}/\theta_{2j})$ under the cluster models with and without dependence, in (3.4) and (3.8), for the function word data.

and on the two-authors case, it all extends to other stylometric characteristics and to the authorship attribution of texts with more than two authors. A slight modification of the prior for the cluster weights, ω_i, can also accommodate for dependence structures other than the one used here for texts or corpus that are sequentially ordered.

Chapter 4

Bayesian Analysis of the Heterogeneity of Literary Style

A statistical analysis of the heterogeneity of literary style in a set of texts that simultaneously uses different stylometric characteristics, like word length and the frequency of function words, is proposed. Data consist of several tables with the same number of rows, with the i-th row of all tables corresponding to the i-th text. The analysis proposed clusters the rows of all these tables simultaneously into groups with homogeneous style, based on a finite mixture of sets of multinomial models. That has the advantage over the usual heuristic cluster analysis approaches that it naturally incorporates in the analysis the text size, the discrete nature of the data, and the dependence between categories. All this is illustrated through an analysis of the heterogeneity in the plays by Shakespeare and in *El Quijote*, and by revisiting the authorship-attribution of *Tirant lo Blanc*.

4.1 Introduction

The statistical analysis of literary style has often been used to characterize the style of texts and authors, and sometimes help settle authorship-attribution problems both in the academic as well as in the legal context. Work as early as Mendenhall (1887, 1901) and Yule (1938) already used word length and sentence length to characterize literary style. Other characteristics widely used for this purpose have been the proportion of nouns, articles, adjectives or adverbs, the frequency of use of function words, which are independent of the context, or of characters, and the richness and diversity of the vocabulary used by the author. Good reviews about the statistical analysis of literary style can be found in Holmes (1985, 94, 98, 99) and Stamatatos (2009).

The range of statistical methods used in this setting is wide, most often involving the use of classification tools. In typical authorship-attribution and verification problems one has a set of candidate authors and a list of known texts from each one of them that can be used as training texts, and one needs to assign texts of unknown author to one of the authors in the set by comparing their style to the one of the training texts. In these settings, one resorts to discriminant analysis, also recognized as supervised classification/learning.

Instead, in the analysis of the heterogeneity of literary style that is tackled in this paper, the setting is a lot less structured because one does not assume to have a reference set of candidate authors and of training texts, and one needs to resort to cluster analysis, also recognized as unsupervised classification/learning.

The goal in cluster analysis is to partition observations (texts) into meaningful subgroups, without assuming much about the number of subgroups and about the composition of the groups. Most of the literature on cluster analysis is devoted to continuous data and uses ad hoc heuristic partitioning algorithms that tend to be easy to apply and work well, but that do not allow one to assess cluster uncertainties and do not provide inference based methods to choose the number of clusters and allocate individual observations to clusters. Good introductions to that literature are Greenacre (1988) or Kaufman and Rousseeuw (1990).

Instead, model based clustering assumes that observations come from a population with several subpopulations, and one models the overall population through a finite mixture of the subpopulation models. Bayesian model based cluster analysis provides a complete probabilistic framework for the problem by assuming a finite mixture model under which observations belonging to the same cluster have the same distribution, and then estimating the mixed distributions and assigning observations to these component distributions. Model based approaches simultaneously group objects and estimate the component parameters, and that avoids the biases appearing whenever that is done separately. These methods also have the advantage of providing a measure of the uncertainty in the allocation of individual observations into clusters, and of casting the choice of the number of clusters and hence of component distributions as a statistical model selection problem.

For early examples of the use of Bayesian model based cluster analysis, mostly using mixtures of multivariate normal distributions, see Murtagh and Raftery (1993), Banfield and Raftery (1993), Fernandez and Green (2002) and Fraley and Raftery (2002).

To help settle the debate around the authorship of *Tirant lo Blanc*, Giron, Ginebra and Riba (2005) explored the heterogeneity of its style by carrying out a Bayesian model

based cluster analysis of word length and of the frequency of the most frequent words in its chapters. The data consisted of two contingency tables of ordered rows, with the i-th row in both tables corresponding to the i-th chapter of the book, and the cluster analysis of the rows of each one of these two tables was carried out separately based on a finite mixture of multinomial models. Resorting to these models allows one to implement a cluster analysis based on the whole vector of word length or of function word counts instead of basing it on individual counts. That also has the advantage over heuristic and/or normal based clustering approaches that it naturally incorporates in the analysis the text size, the discrete nature of the data and the dependence between categories.

This analysis based on finite mixtures of multinomial models is generalized here by:

1. carrying out a single cluster analysis using more than one stylometric characteristic at once, by treating a set of more than one vector of counts as an observation,

2. by incorporating a model-checking stage that compares the realization of statistics in the data with their realization in predictive simulations from the models, and

3. by providing closed form expressions for the exact calculation of the probabilities of the models considered being correct, to be used to select models.

The combination of the model-checking and model selection stages will help determine the number of mixture components required by the data, and hence the number of clusters. As a by product of the model-checking stage, the analysis allows one to check whether finite mixtures of a small number of purely multinomial models are flexible enough to capture all the variability in the data. If they were not, one would need to resort to more complicated finite mixtures of sets of continuous mixtures of multinomial models instead.

To illustrate the analysis, it is implemented on three examples, each dealing with the main work of a different literature. The first case study explores the heterogeneity of style in the plays in the *first folio edition* of Shakespeare's drama. In the second case study, the authorship-attribution problem of *Tirant lo Blanc* is revisited. Finally, the same type of heterogeneity analysis is implemented on the chapters of *El Quijote*.

In all the examples the analysis will be mostly exploratory, without attempting to assess whether the heterogeneities found are linked to differences in authorship or otherwise could be explained by differences in chronology, genre or topic. Some might question the legitimacy of limiting the approach to be exploratory in the Shakespeare case, which is the most structured of the three. Note though that, without making explicit a list of candidate authors and of training texts, there is no legitimate statistical way of going beyond proposing tentative explanations for the heterogeneities detected in the corpus.

4.2 Description of the data

The methodology advocated for here combines in the analysis as many stylometric characteristics as one needs to. All the characteristics considered will have to involve counting features that are categorical and have a fixed number of categories. That includes for example counting characters, words or sentences of certain lengths, function words, nouns or adjectives. As a consequence, data will consist of a set of tables with the same number of rows, with one table for each characteristic. We use *word length* and the count of the most frequent *function words* as illustrating examples.

Early uses of word length to help characterize style can be found in Mendenhall (1887, 1901), Mosteller and Wallace (1964, 84), Brinegaar (1963), Bruno (1974), Williams (1975), Morton (1978), Smith (1983), Hilton and Holmes (1993). Even though present day surveys on the use of stylometric variables in authorship attribution of texts written in English rarely find word length as a useful discriminating feature, Giron et al (2005) find that feature to be very useful in the authorship attribution of a text written in Catalan. Furthermore, note that in Figure 4.1 word length discriminates well between comedies on one side and histories and tragedies on the other, and therefore word length is useful to detect heterogeneities of style in English texts not necessarily linked to differences in authorship.

The frequency of use of function words has proved to be one of the best tools when it comes to discriminating styles. Early uses of function words can be found in some of the references already listed above, as well as in Burrows (1987, 92), Holmes (1992), Binongo (1994) or Oakes (1998). Recent discussions on the use of stylometric variables, and, in particular, of function words, can be found in Zhao and Zobel (2005), Miranda-Garcia and Calle-Martin (2007), Luyckx (2010), Hope (2010) and Rybicki and Eder (2011).

In those cases where the analysis of word length and word counts separately lead to very different results, their combination will be problematic. But when separately they lead to similar results, as was found to be the case in *Tirant lo Blanc* by Giron et al (2005), their combination in a single analysis is warranted. By combining them, the uncertainty in the classification of texts into clusters will be reduced.

When one decides to simultaneously analyze word length and function word counts in the example of *Tirant lo Blanc*, one is lead to the simultaneous analysis of the 487×10 table of word length counts and of the 487×12 table of counts of twelve of the most frequent function words partially presented in Table 4.1.

Word length counts												
Chapter	1	2	3	4	5	6	7	8	9	10+	N_i^1	$\overline{wl}_i$
1	21	59	44	19	33	20	16	17	9	17	285	4.47
2	53	113	80	49	52	33	28	36	16	16	476	4.14
...	...	...	...	...	...	...	...	...	...	...	...	...
487	48	49	62	53	41	36	21	9	16	13	348	4.20
Most frequent word counts												
Chapter	e	de	la	que	no	l	com	molt	és	jo	si	dix
1	12	15	9	8	1	7	2	1	6	0	3	0
2	26	28	19	9	3	2	3	8	3	1	3	1
...	...	...	...	...	...	...	...	...	...	...	...	...
487	29	13	8	10	2	10	3	9	0	0	0	0

Table 4.1: Part of the table of word length counts in the chapters of *Tirant lo Blanc*, and of the table of counts of twelve of the most frequent function words in them. N_i^1 is the total number of words and $\overline{wl}_i$ the average word length.

In general, for each chapter i in a book (or act of a play) with $i = 1, \ldots, n$, and each stylometric characteristic, r, with $r = 1, \ldots, R$, one has a vector valued categorical observation, $y_i^r = (y_{i1}^r, \ldots, y_{ik(r)}^r)$, where $k(r)$ denotes the number of categories of the r-th characteristic. This vector, y_i^r, becomes the i-th row of the r-th table considered.

In the *Tirant lo Blanc* example, y_i^1 is the ten dimensional vector of word length counts of its i-th chapter, and y_i^2 is the twelve dimensional vector of function word counts in that chapter. More generally that leads to a set of R different $n \times k(r)$ tables, one table for each characteristic. The set of all the n rows in the r-th table will be denoted by $y^r = (y_1^r, \ldots, y_n^r)$, and the set of all the R tables will be denoted by $y = (y^1, \ldots, y^R)$. The goal is to cluster the rows of all these tables simultaneously into S different groups with homogeneous style, assuming that the rows in a group are multinomially distributed.

One of the main shortcomings of the heuristic based cluster analysis approaches typically used in stylometry, like the ones based on PCA, k-means or hierarchical methods, is that they implicitly assume data to be continuous or are at least tailored to work best when data is continuous. But stylometric data is mostly categorical, and the methodology for it should move in the direction of addressing the specificities of that kind of data.

In particular, most of these mostly ad-hoc heuristic methods have a difficult time taking into account that texts of different length have different amount of information about the style of their author and hence they should be weighted differently in the analysis. These basic methods also have a hard time taking into consideration the dependence present between counts of categories of the same stylometric characteristic.

The cluster analysis proposed next, based on carefully modeling the data probabilistically using mixtures of multinomial models, avoids the continuity assumption and it naturally weights texts according to text size, which avoids the need to deal with texts of similar sizes to avoid biasing the results. Furthermore, by assuming the observations in each cluster to be multinomially distributed, one also naturally takes into account the dependence between counts of categories of the same characteristic.

4.3 Description of the Multinomial cluster model

The i-th row of the r-th table is assumed to be multinomially distributed, $\text{Mult}(N_i^r, \theta_i^r)$, where $\theta_i^r = (\theta_{i1}^r, \ldots, \theta_{ik(r)}^r)$ is such that $\sum_{j=1}^{k(r)} \theta_{ij}^r = 1$, where θ_{ij}^r is the probability of the j-th category for the i-th row and the r-th characteristic, and where $N_i^r = \sum_{j=1}^{k(r)} y_{ij}^r$. If all the chapters of the book or acts in the plays shared the same style, one might expect the distribution of all the n rows for any given characteristic to remain the same, in which case they could all be modeled as a random sample of a single $\text{Mult}(N_i^r, \theta^r)$ distribution.

Instead, if the style in the n chapters or acts was not homogeneous, but these chapters grouped themselves in S different styles, maybe because they had been written by S different authors, then the n rows of the r-th table, $y^r = (y_1^r, \ldots, y_n^r)$, could be considered to be conditionally independent and modeled through a finite mixture of S multinomial distributions, with probability density function (pdf):

$$p(y^r|\omega, \theta_1^r, \ldots, \theta_S^r) = \prod_{i=1}^{n} \sum_{s=1}^{S} \omega_s \text{Mult}(N_i^r, \theta_s^r), \tag{4.1}$$

where $\theta_s^r = (\theta_{s1}^r, \ldots, \theta_{sk(r)}^r)$ determines the distribution of the rows in the s-th cluster of the r-th table, and where $\omega = (\omega_1, \ldots, \omega_S)$ is a set of weights, with $0 \leq \omega_s \leq 1$ and $\sum_{s=1}^{S} \omega_s = 1$, determining the proportion of rows (chapters or acts) belonging to each cluster.

To be able to allocate rows into clusters, which is an essential feature in cluster analysis, one introduces a vector of unobserved (latent) categorical variables $\zeta = (\zeta_1, \ldots, \zeta_n)$, where ζ_i takes values in $\{1, \ldots, S\}$ and is such that $\zeta_i = s$ whenever the i-th row belongs to the s-th cluster. Here the ζ_i are assumed to be conditionally independent and hence:

$$p(y^r, \zeta|\omega, \theta^r) = \prod_{i=1}^{n} \omega_{\zeta_i} \text{Mult}(N_i^r, \theta_{\zeta_i}^r), \tag{4.2}$$

where $\theta^r = (\theta_1^r, \ldots, \theta_S^r)$ is the set of multinomial probabilities for the r-th table. The latent variable ζ assigning chapters or acts into clusters does not depend on r, and hence

it takes a common value for all the stylometric characteristics considered. That is, the i-th rows in all the tables are always allocated into the same cluster.

In Bayesian statistics, one needs to choose a distribution for the parameters of the model that captures what one knows about them before observing the data, which is denoted as the prior distribution. Here, that prior distribution will assume that all vectors of probabilities across clusters and tables, θ_s^r for $s = 1, \ldots, S$ and $r = 1, \ldots, R$, are independent, and that the θ_s^r are Dirichlet$(a_{s1}^r, \ldots, a_{sk(r)}^r)$ distributed. The weights ω determining the relative sizes of the clusters are assumed to be Dirichlet$(b_1, \ldots, b_S)$ distributed. In our examples all the $(a_{s1}^r, \ldots, a_{sk(r)}^r)$ and $(b_1, \ldots, b_S)$ are set to be equal to $(1, \ldots, 1)$, which corresponds to assuming a uniform distribution on the simplex. The $R = 1$ and $S = 2$ special case of this model is the one used in Giron et al. (2005).

In Bayesian statistics one combines the distribution chosen for the parameters before obtaining the data (the prior distribution) with the data, to compute an updated distribution that incorporates the information contributed by the data. That updated distribution for the parameters is called as the posterior distribution, and in our case it is too complicated to be computed analytically. Instead of that, one can update the model and simulate from it with the WinBugs implementation (see, Lunn et al. 2013).

4.4 The choice of the number of clusters

A difficulty of the heuristic clustering algorithms is that they often lack a statistically grounded method for determining the number of clusters. Instead, under model based clustering the choice of the number of clusters, S, coincides with the choice of model.

The safest way to build a model is through the iterative use of model checking tools that help discover aspects of reality not adequately captured by the models and suggest ways of improving them. To help support that model choice, one can also resort to formal model selection methods, based on the computation of the posterior probability that each one of the models considered is the one generating the data.

Cluster analysis is useful only when the answer contains a relatively small number of clusters, and hence it will typically be better to settle with an approximate model that has a small number of clusters but explains a large portion of the variability, than with a model that is "true" and captures all the variability but requires a large number of clusters.

4.4.1 Choice of s through model-checking

Building a Bayesian model is like building a data simulation model. Hence, they should be assessed and chosen based on whether it is plausible that they could simulate data like the one observed in reality or not. Following the lead of Gelman et al (2004), we will graphically compare the set of R observed tables, with analogous sets of tables simulated from the posterior predictive distribution of the models.

To compare the table with the word length data to the corresponding tables with the replicated data are summarized through the proportion of words of L letters in each chapter or act for $L = 1, \ldots, 9$ and for $L > 9$. We also summarize them through the average word length, through the ratio between the number of words with more than 5 and of less than 6 letters, and through the first correspondence analysis components of each table. To compare the table with the observed word counts with the corresponding simulated tables, they are summarized through the frequency of appearance of each one of these words separately, and through the first correspondence analysis components of each table.

A sampler of these predictive comparisons will be presented in the first case study. We do not report on the predictive checks for the other examples for the sake of brevity. For more examples of posterior predictive checks used to assess Bayesian models in the context of the analysis of literary style, see Font et al (2013), and for similar examples in the context of choosing the number of clusters, see Puig and Ginebra (2014a, b).

4.4.2 Choice of s through model selection

The formal way to select a model is through the posterior probability of each model, $P(M_S|y)$, which is the probability that the S-cluster model, M_S, is the one generating the data, assessed after the data has been observed. It can be computed through:

$$P(M_S|y) = \frac{P(M_S)P(y|M_S)}{\sum_{s=1}^{S_T} P(M_s)P(y|M_s)}, \tag{4.3}$$

where $P(M_S)$ is the prior probability assigned to M_S, (i.e., the probability that this model is correct, assessed before data is available), where$P(y|M_S)$ is the marginal likelihood of M_S, and where S_T is the largest number of clusters that one is willing to contemplate.

To select the number of clusters one needs to select a single model, and the most natural choice is the model with the highest posterior probability. If all models were considered equally likely a priori, the larger $P(y|M_S)$, the more attractive M_S would be. But there

is a big debate on how prior probabilities on model space should be chosen, due to the large difference in complexity between models (see, e.g., Casella et al, 2014).

Most often, computing $P(y|M_S)$ exactly is too complicated, and one approximates its logarithm through the BIC, as in Fraley and Raftery (2002). Alternatively, one can estimate $P(y|M_S)$ through the simulations used to update the model, as in Gelfand and Dey (1994). In our special multinomial mixture setting though, compute these marginal likelihoods exactly through the closed form expression given in an Appendix.

It is important to emphasize that adopting the formal Bayesian approach to model choice presented here does not help identify what are the shortcomings of the models, when they have them. Hence, computing the posterior probabilities of the models under consideration does not spare one having to check models on the side, the way described in Section 4.4.1.

4.5 Case study 1: Shakespeare's drama

William Shakespeare (1564-1616) is regarded by many to be the greatest writer in the English literature. Very little is known about his personal life, which has fueled a debate around the authorship of plays and poems attributed to him. Even though only a minority of the experts question his authorship, some claim that the true author of some or all of the works attributed to him could be Francis Bacon, Cristopher Marlowe, Ben Johnson, Sir Walter Raleigh or Edward de Vere. That debate has been going on for more than 150 years, and far too many people has contributed to it to be able to summarize it adequately here. For recent overviews of that debate see, for example, Hope (1994, 2010), Edmondson and Wells (2013) or Shahan and Waugh (2013).

The statistical analysis of the literary style in Shakespeare's drama also started a long time ago. Mendenhall (1901) is one of the earliest examples of the use of statistics to compare the style of Shakespeare's plays with the style of some of its contemporaries, like Marlowe and Bacon; He found that the word length distribution in Shakespeare's plays was extremely close to the one in plays by Marlowe. The list of contributions to the quantitative analysis of the style in texts linked to Shakespeare is very long, and it includes, for example, Smith (1990), Jackson (2003), Vickers (2004) and more recently Craig and Kinney (2009).

The type of statistical analysis carried out next is different of most of the statistical analysis carried out on Shakespeare's drama in two main regards. The first difference arises from the fact that here one is trying to identify any heterogeneities in the style of

Shakespeare's drama, irrespective of whether they are linked to authorship differences or not, while the literature on Shakespeare's drama has understandably focused mainly on authorship attribution issues. The second difference with other published statistical analysis of Shakespeare's drama, is that they heavily rely on the use of "training" groups of texts of undisputed authorship to help determine the authorship of the disputed texts, while we do not rely on any of such texts to start with. That explains that they mainly resort to the use of supervised classification (discriminant analysis) tools, while here we present a method to carry out unsupervised classification (cluster) analysis.

To explore the heterogeneity of style in Shakespeare's drama, here a cluster analysis is carried out on the 35 plays gathered in the first printing of the *first folio edition* of Shakespeare's plays published posthumously in 1623. That edition includes fourteen comedies, ten histories, and eleven tragedies, and it is the only reliable version for about twenty of these plays. Common wisdom supports the idea that some of the plays, and specially the early histories, might have been revised by other writers. *Troilus and Cressida* did not appear in the first printing of that edition and *Pericles and the two noble kinsmen* did not appear in any of its printings, and they have not been included in this study even though they are also attributed to Shakespeare.

In the analysis, plays are broken down into five acts each, and hence a total of 175 textual units are considered. The goal of the analysis is to check whether acts naturally cluster themselves together into more than one cluster when one takes into account word length and the frequency of the twenty most frequent function words in them. Hence data will consist of a 175×10 table with the word length counts, and of a 175×20 table with the twenty most frequent word counts. In this case study the analysis will be exploratory because a different style might be related to many different factors, such as the time of writing, the kind of play, or the author, and it is not easy to know which factors are at play.

To help choose the number of clusters, one needs to assess whether the models involved capture the relevant features in the data. As a sample of this exercise, Figure 4.1 compares the observed proportion of words of *one, two, three, nine* and of *more than nine* letters in these 175 acts, the average word length, the ratio of the number of long words and of short words with the ones corresponding to a sample simulated from the posterior predictive distribution under the one-, the two-, and the three-cluster model. The data plots on the left column of Figure 4.1 correspond to the actual plays by Shakespeare, while the data plots on the remaining three columns of that figure correspond to data replicates obtained from the three simplest multinomial mixture models.

Figure 4.2 compares the frequency of *the, and, I, you, it, your* and *his* actually observed

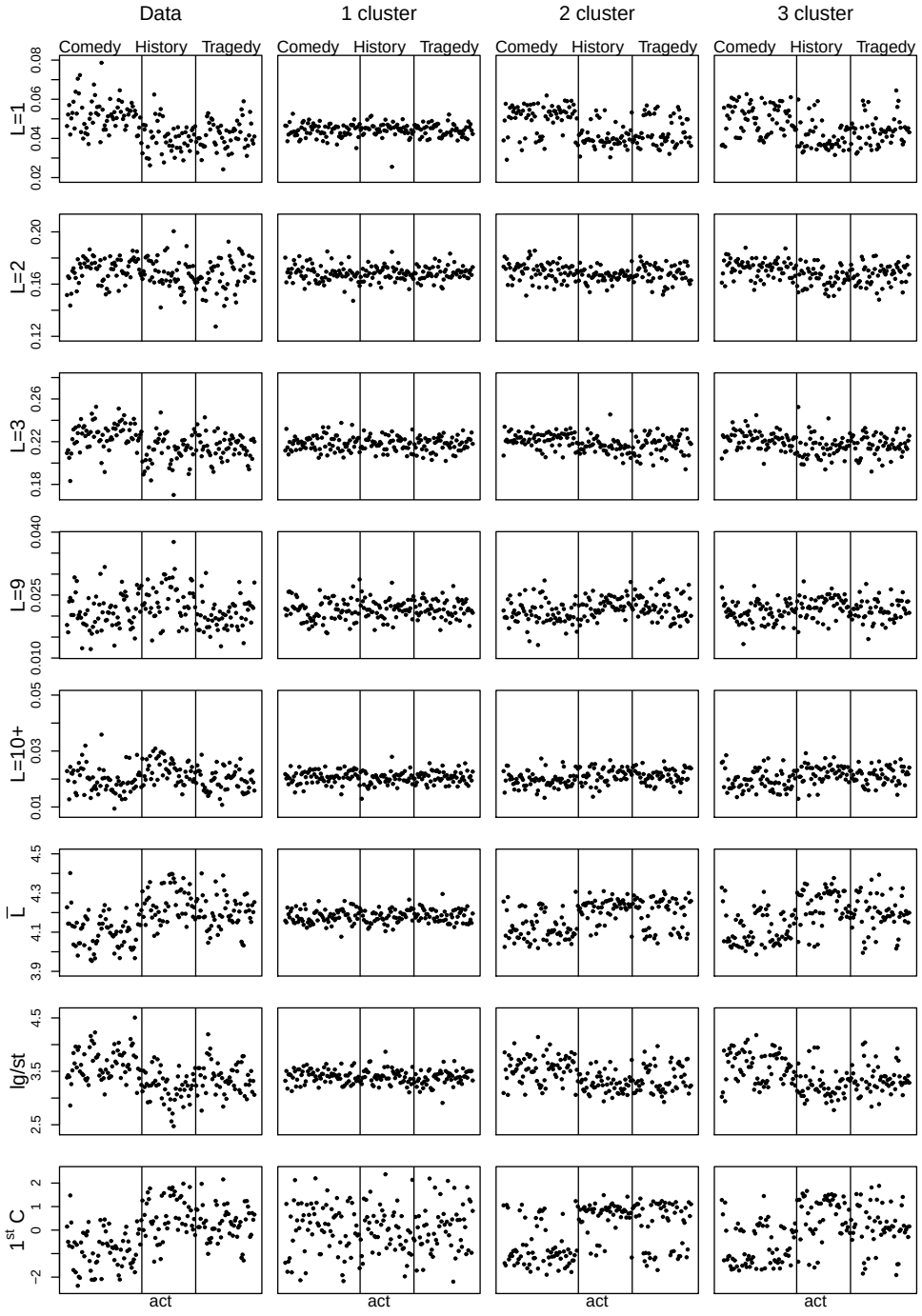


Figure 4.1: In the left column, proportion of words of one, two, three, nine and more than nine letters, average word lengths, ratio between the number of long and of short words in the acts of the plays in Shakespeare's drama, and first correspondence analysis component of the table of word lengths. Next to each of these plots, posterior predictive replicates under the one-, two- and three-cluster models.

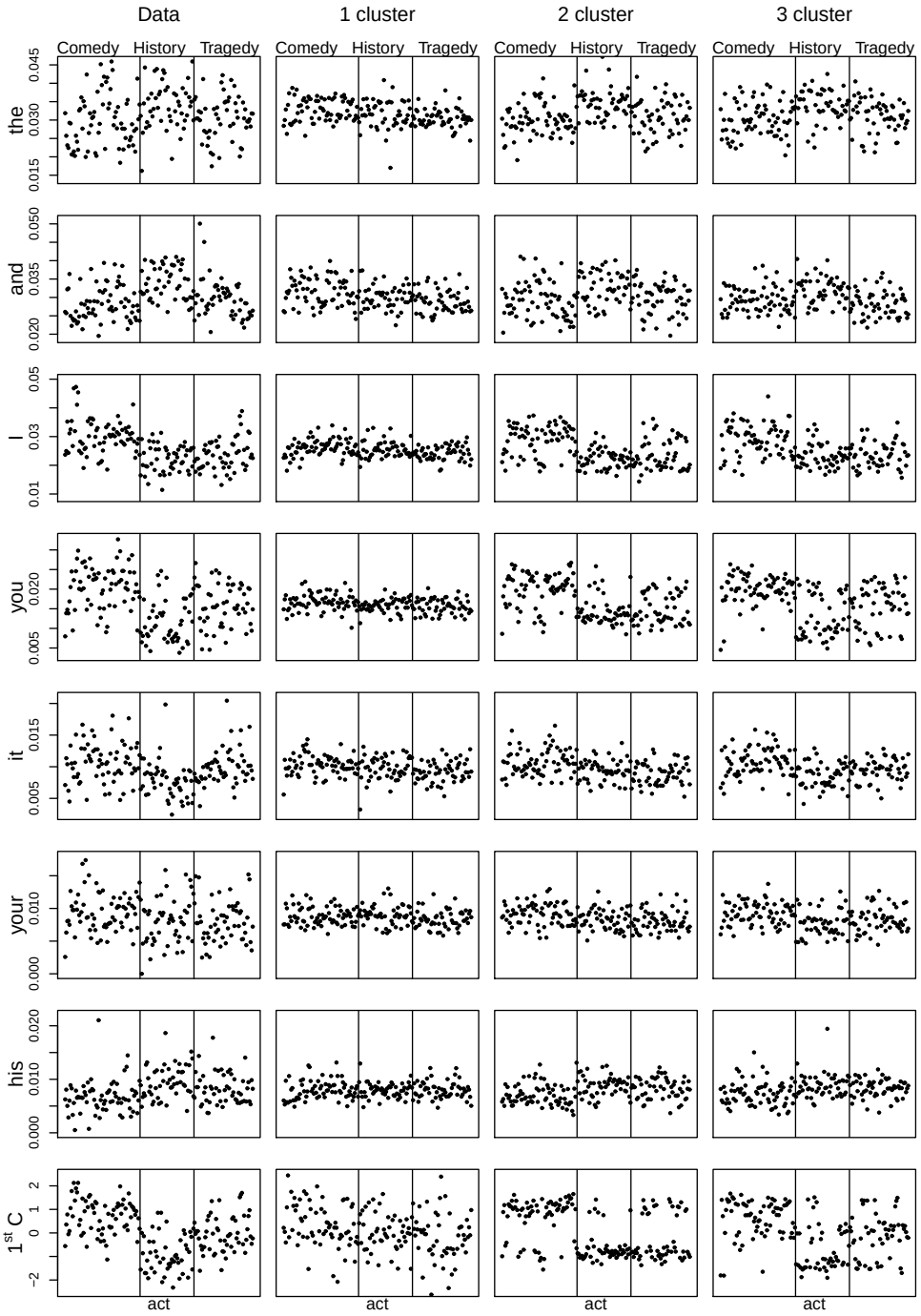


Figure 4.2: In the left column, frequency of appearance of *the, and, I, you, it, your* and *his* in the acts of the plays in the *first folio edition* of Shakespeare, and first correspondence analysis component of the table with the twenty most frequent word counts. Next to each of these plots, posterior predictive replicates under the one-, two- and three-cluster models.

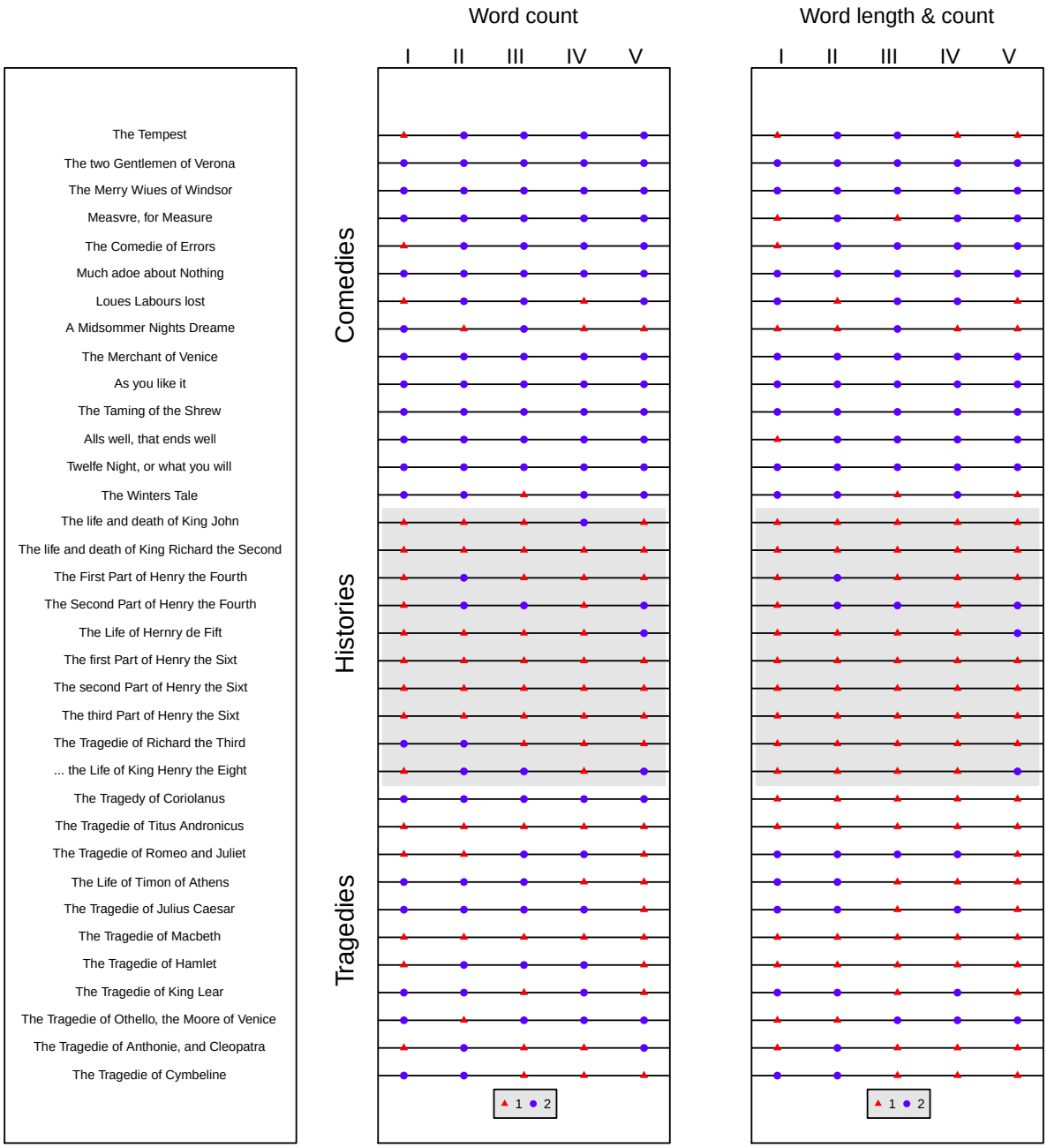


Figure 4.3: Classification of each one of the five acts of each of the plays in the *first folio edition* of Shakespeare under the two-cluster model, first using only word counts and second using both word length as well as word counts.

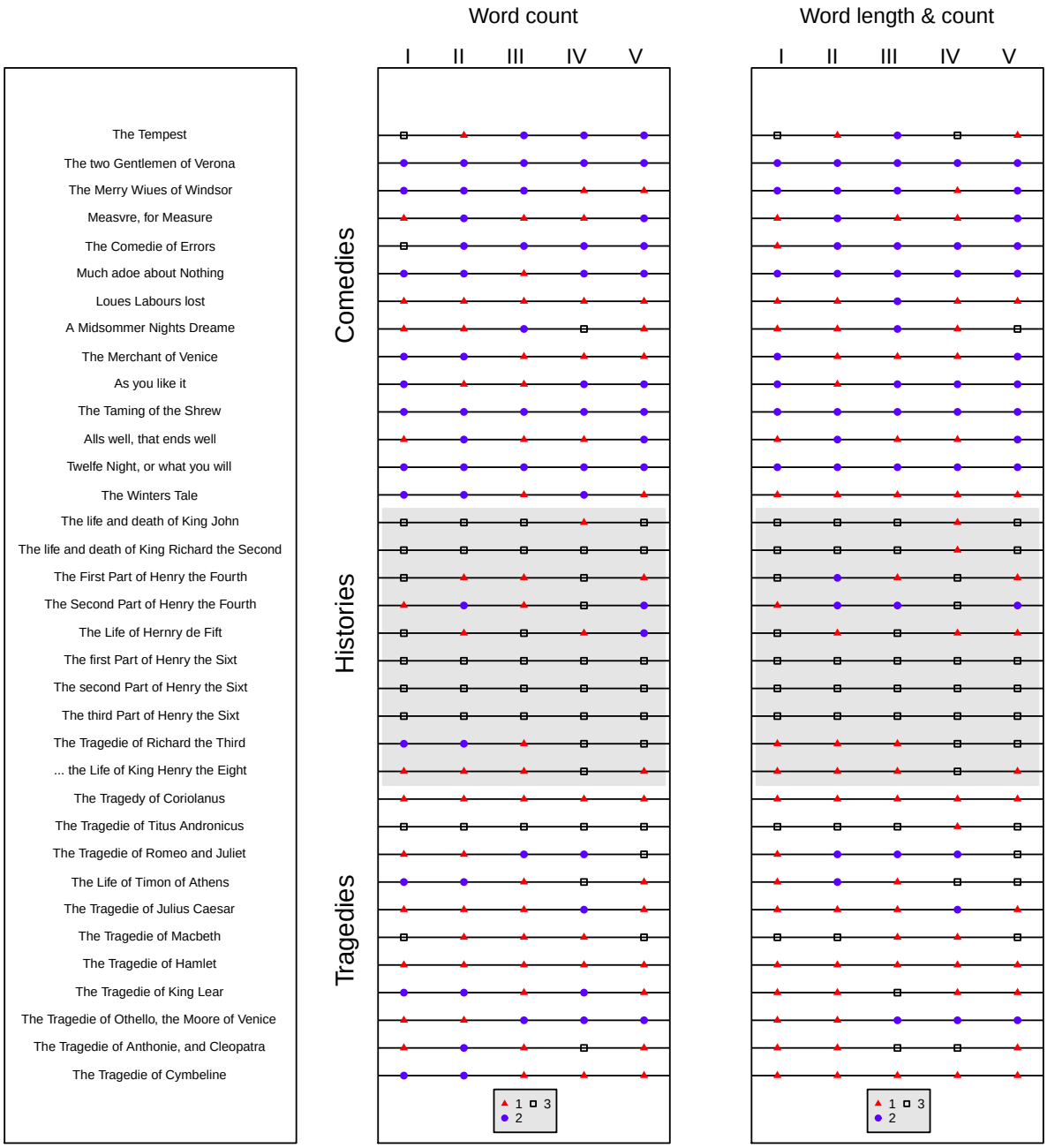


Figure 4.4: Classification of each one of the five acts of each of the plays in the *first folio edition* of Shakespeare under the three-cluster model, first using only word counts and second using both word length as well as word counts.

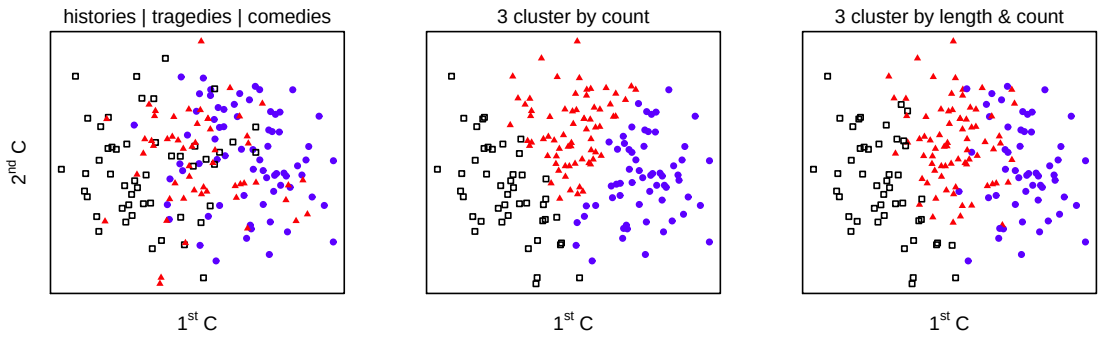


Figure 4.5: First correspondence analysis components of the table of word counts in the acts of Shakespeare drama, stratified according to genre, and according to the cluster to which the act belongs when using only word counts, and when using both word length as well as word counts.

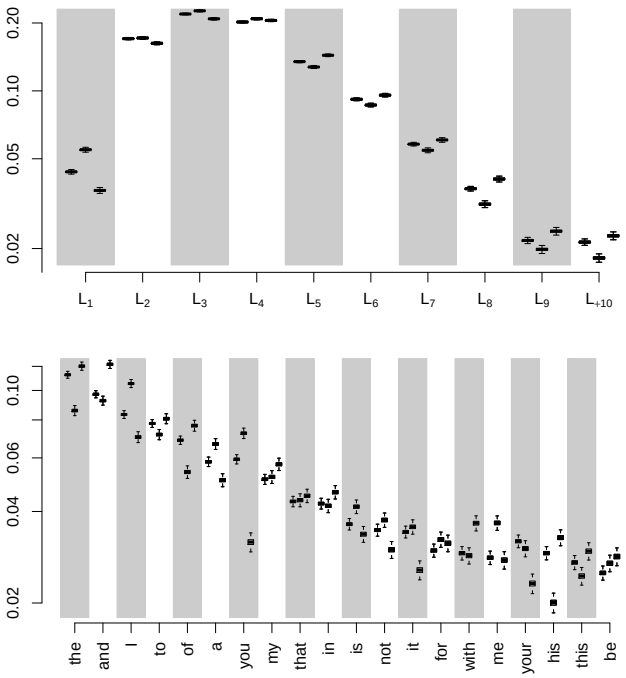


Figure 4.6: Box-plots of a sample of the probabilities for word length, $(\theta_1^{wl}, \theta_2^{wl}, \theta_3^{wl})$, and for word counts, $(\theta_1^{mf}, \theta_2^{mf}, \theta_3^{mf})$, in the three clusters of acts of plays in the *first folio edition* of Shakespeare, all in a logarithmic scale.

in the plays by Shakespeare, on the left column, with the corresponding frequencies in a sample simulated from the same multinomial mixture models, on the remaining three columns. Figures 4.1 and 4.2 also compare the first correspondence analysis component summarizing the two tables of data considered here with the components summarizing analogous tables obtained by simulating from these models.

Note for example that the average word length tends to be smaller and the proportion of one and three lettered words tends to be larger for comedies than for histories or tragedies, while for example the use of the words *I* and *you* tends to be more frequent for them. It is worth remarking the fact that, even though current common wisdom states that word length is not an effective stylometric variable when trying to discriminate the style of English authors (see, e.g., Mosteller and Wallace, 1984), word length does indeed help distinguish the style used in Shakespeare's comedies from the style used in his histories and tragedies.

One now has to check whether either one of the one-, two- or three-cluster models considered in Section 4.3 capture the patterns in Figures 4.1 and 4.2 adequately or not. Figures 4.1 and 4.2, and many other posterior predictive checks made on the side, not reported here, all indicate that here these finite mixtures of multinomial models are able to reproduce most of the variability in the data. To choose among the one-, the two- and the three-cluster models, several of the statistics in Figures 4.1 and 4.2 indicate that at least three clusters are needed to capture the variation in the levels of these statistics.

Here the natural logarithm of $P(y|M_S)$ under the one-, two-, three- and four-cluster models are equal to -25488.4, -23608.0, -22988.9 and -22677.3 respectively. If one computes the posterior probabilities that each one of these four cluster models is the correct one through (4.1), one chooses the four-cluster model. But if one penalizes models with more clusters by assigning them much smaller prior probabilities, as recommended by Casella et al (2014), one settles with the two- or three- cluster models. In fact, Figures 4.1 and 4.2 indicate that the two- and the three-cluster models already account for most of the variability in the data.

In order to compare the result of the cluster analysis combining the information of both word length and the use of word counts, with the results of the cluster analysis using only word counts, both analysis are carried out.

Figure 4.3 allocates acts into either one of two clusters using the posterior probabilities for ζ_i under the two-cluster model. It indicates that the two-cluster analysis classifies acts mostly along genre. Under this analysis, most of the acts in comedies fall into Cluster 1, most of the acts in histories fall into Cluster 2, while the acts in tragedies are

more or less evenly split across both clusters. As an exception to that rule, most of the acts of "A Midsommer Nights Dreame" are classified as a history instead of a comedy. Note also that all the acts of the tragedies of Titus Andronicus and of Machbeth are classified as histories, while the acts of all other tragedies are split between both clusters.

When one compares the result of the analysis combining word length and word counts, with the analysis based only on word counts, one finds that only a small number of acts change allocation. The results of both analysis are different and yet, similar enough, to justify the combination of both characteristics into a single analysis.

Figure 4.4 allocates acts into clusters under the three-cluster model, again first based only on word counts and second, based on both word counts as well as word lengths. Here it also appears that the classification of acts into clusters is mostly made along genre, with Cluster 1 being mostly formed by acts in tragedies, Cluster 2 mostly by acts in comedies, and Cluster 3 mostly by acts in histories. The result of the analysis combining word length and word counts and the analysis based only on word counts are again different, and yet, similar enough to justify the combination of both characteristics into a single analysis.

To help interpret the results, Figure 4.5 presents the first correspondence analysis components for the table of word counts in the acts of Shakespeare's drama. Correspondence analysis is analogous to PCA but tailored for categorical instead of continuous data (see, e.g., Greenacre, 2007). Acts are stratified first across genre, which helps emphasize that the heterogeneity of style found in Shakespeare's drama mostly relates to genre. Acts in Figure 4.5 are also stratified according to their three-cluster classification, which shows how clusters mostly group observations close together in the space of the first correspondence analysis components, and which helps appreciate what changes from combining word length and word counts in the analysis instead of just using word counts.

To help understand what distinguishes the style of clusters, Figure 4.6 presents a sample of the posterior distribution of the multinomial probabilities for word length counts and for the most frequent words under the three-cluster model. Cluster 2, mostly formed by comedies, has the largest proportion of words with *one, two* or *three* letters and the smallest proportion of words with *five, six, seven, eight, nine* or *more than nine* letters. Cluster 2 also has the largest frequencies of *I, a, you, it*, and of *me*, and the smallest frequencies of *and* and of *his*. Clusters 1 and 3 seem to be much more similar in terms of most of the categories considered, with Cluster 3 being special for having smaller frequencies of *I, you, it* and *your*, and larger frequencies of *the, of* and *with* than the other two clusters.

4.6 Case study 2: Tirant lo Blanc

Tirant lo Blanc is a chivalry book written in catalan and hailed to be "the best book of its kind in the world" by Miguel de Cervantes. The main body of the book was written between 1460 and 1464, but it was not printed until 1490, and there has been a long lasting debate around its authorship, originating from conflicting information given in its first edition. Where in the beginning of the book it is stated that "*So that no one else can be blamed if any faults are found in this work, I, Joanot Martorell, take sole responsibility for it,*" at the end of the book it is stated that "*Because of his death, Sir Joanot Martorell could only finish writing three parts of it. The fourth part, which is the end of the book, was written by the illustrious knight Sir Martí Joan de Galba.*" Over the years, experts have split between the ones favoring the single authorship hypotheses, and the ones backing the hypotheses of a change of author somewhere between chapters 350 and 400.

It is well accepted that the main (and maybe single) author died in 1465, and neither he nor the candidate to be the book finisher left any other texts comparable with this one. Different from the situation in the previous example, here the analysis is more structured because there are not as many factors that could explain differences in style other than differences in authorship, and hence the analysis is less of an exploratory nature.

An analysis of the diversity of the vocabulary in Riba and Ginebra (2006) finds that it becomes significantly less diverse after chapter 383. Giron et al (2005) and Riba and Ginebra (2005) carried out a change point and a two-cluster analysis first for word length and second for the most frequent words separately. In both cases a stylistic boundary is detected between chapters 371 and 382.

This agreement between the results reached through the analysis of word counts and through the analysis of word lengths was what triggered our interest in combining the information in word length with the information in word counts in a single combined analysis. Different from what happens for English texts, it turns that in other languages word length might be useful when discriminating between authors.

These papers formally tested for the existence of more than one cluster under each characteristic, by computing the probabilities in (4.1) under each one of the two tables separately, and it was decided that there were two clusters, but it was also conjectured that finite mixtures of Dirichlet-multinomials might be better able to capture the variability in the data than finite mixtures of multinomials.

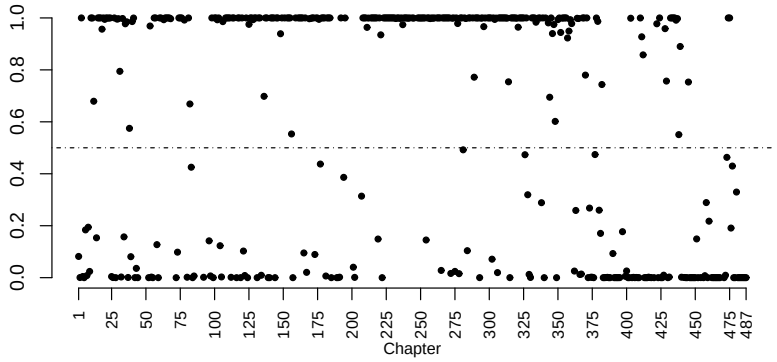


Figure 4.7: Probability that chapters in *Tirant lo Blanc* belong to Cluster 1.

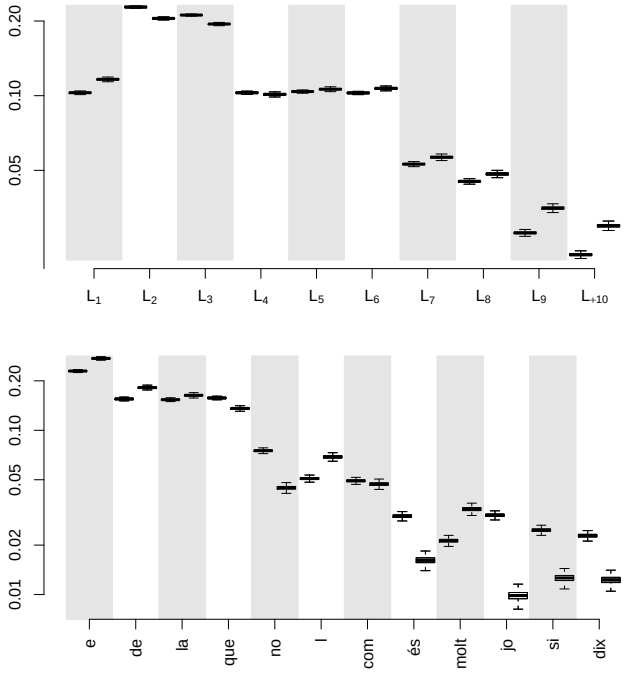


Figure 4.8: Box-plots of a sample of the multinomial probabilities for word length, $(\theta_1^{wl}, \theta_2^{wl})$, and for word counts, $(\theta_1^{mf}, \theta_2^{mf})$, for the two clusters in *Tirant lo Blanc*, all in a logarithmic scale.

Here a cluster analysis is carried out simultaneously based on both the 425×10 table of word length counts as well as on the 425×12 table with the count of the twelve words chosen in Giron et al (2005) based on their discrimination power between the beginning of the book and its ending. As in that paper, only chapters with more than 200 words are considered. Posterior predictive model checks carried out here similar to the ones in Figures 4.1 and 4.2 for the plays of Shakespeare indicate that here one can also rely on a finite mixture of sets of purely multinomial models. Hence the conjecture that one might need mixtures of sets of Dirichlet-multinomial models instead is not called for.

Figure 4.7 presents the posterior probability that the i-th row (chapter) belongs to Cluster 1, $\zeta_i = 1$, which is what one needs to classify the chapters of *Tirant lo Blanc* into either one of the two clusters. Cluster 1 mostly includes chapters previous to chapters 375-385, while Cluster 2 mostly includes chapters that come after that boundary, but there are a fair amount of chapters misclassified by that boundary. This partition of chapters into clusters is similar to the partitions obtained through the analysis carried out in Giron et al (2005) considering the two characteristics separately.

The distribution of the multinomial probabilities under the two-cluster model presented in Figure 4.8 indicate that *two* and *three* lettered words are more abundant in Cluster 1, while *one, six, seven, eight, nine* and *more than nine* lettered words are more abundant in Cluster 2. That figure also indicates that the words *que, no, com, és, jo, si* and *dix* are significantly more abundant in Cluster 1, mostly in the first part of the book, while *e, de, la, l'* and *molt* are more abundant in Cluster 2, mostly at the end of the book.

Note that the results presented in this case study are based on the analysis of the counts of twelve words that were selected by Giron et al (2005) based on their discriminating power between the style at the beginning and at the ending of that book. They first did the analysis with a larger set of words and realized that the main difference in style as between the first four fifths of the book and the last one fifth, and then they repeated the analysis with the twelve most discriminating subset of words that we have also used here. This sequential approach that starts with about twenty words and then repeats the analysis with the most discriminating words among them is useful, because it helps sharpen the classification power of the method.

Finally, note that different from the previous case study, in this example texts (chapters) are ordered sequentially, and that order is not taken into consideration in the cluster analysis model used here. Puig, Font and Ginebra (2014) proposes an alternative analysis that treats the two stylometric variables separately, but incorporates the fact that chapters close together are more likely to belong to the same author than chapters that are far apart. In that way, one strikes a compromise between change-point analysis, as-

suming all neighboring chapters to belong to the same cluster except the boundary ones, and the kind of cluster analysis considered here, that treat all chapters exchangeably, as if order did not matter whatsoever. In this case the results of the analysis are similar.

4.7 Case study 3: el Quijote

El Quijote, written by Cervantes (1547-1616), is considered to be the most important book in the Spanish literature. It was published in two parts, with the first part having 52 chapters and appearing in 1605, and the second part having 74 chapters and appearing in 1615. The cluster analysis of this book, broken down into its 126 chapters, is carried out to check how our approach fares when it is used on a text that is considered to have a rather homogeneous style. Given that no one disputes the single authorship of this book and the contents in the two parts of the book are similar, this exercise allows one to check whether there are any differences in the style of the two volumes that could be explained by the ten year lapse between them.

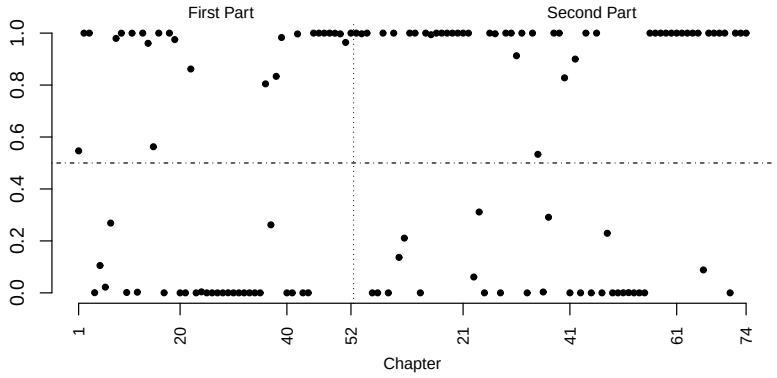


Figure 4.9: Probability that the chapters in *El Quijote* belong to Cluster 1.

Here the analysis is based on the 126×10 table of word length counts and on the 126×20 table of counts of the twenty most frequent function words. Here, the posterior predictive checks that compare the actual data with simulations from the multinomial based S-cluster models already indicate that there is not much to be gained from going beyond one- or two-cluster models in terms of the variability explained by the models. That, and the lack of any meaningful reason why one should expect to find more than one style in *El Quijote*, explains why we only report the result for the two-cluster analysis next.

Figure 4.9 indicates that Cluster 1 is formed by 47 chapters in the second part and 24

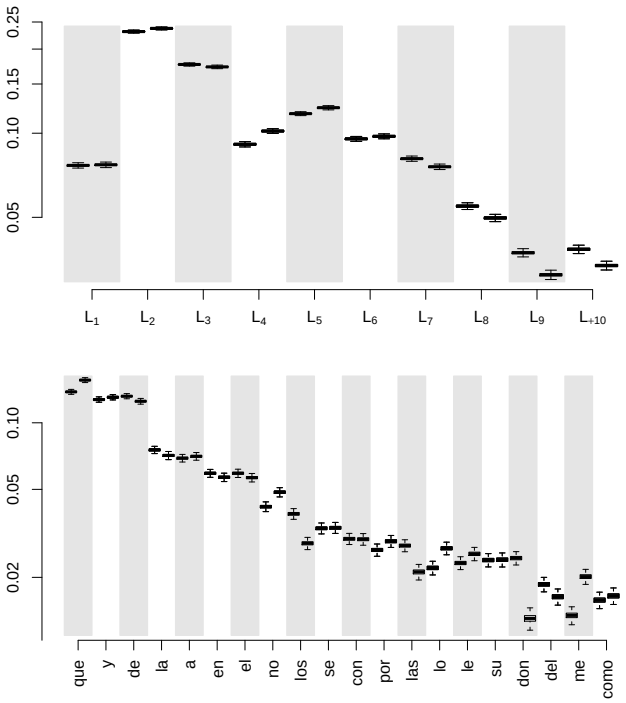


Figure 4.10: Box-plots of a sample of the multinomial probabilities for word length, $(\theta_1^{wl}, \theta_2^{wl})$, and for word counts, $(\theta_1^{mf}, \theta_2^{mf})$, for the two clusters in *El Quijote*, all in a logarithmic scale.

chapters in the first part of the book, while Cluster 2 is evenly split between the two parts of the book. Hence, there does not seem to be any significant differences in style between the first and second parts of the book. Figure 4.10 describing the stylometric characteristics of the two clusters indicates that they are a lot more similar than the two clusters found for *Tirant lo Blanc*, specially when it comes to the word length distribution, in line with the fact that in *El Quijote* there is a single author. The two clusters seem to be mainly distinguished by the frequency in the use of the words *que, no, los, las, lo, don, del* and *me*, but one should not try to make too much out of it since that variation can be most likely explained through the variation in the contents of the respective chapters.

4.8 Final comments

The paper deals with the analysis of the heterogeneity of literary style, which is different from authorship attribution in that one does not have a list of candidate authors and of training texts of known authorship to help build the list of best discriminating words needed to determine authorship of disputed texts. Without them, there is no statistical ground on which to determine whether the heterogeneities detected are due to authorship, chronology, genre, topic or otherwise.

When the original problem is unstructured, because there do not exist any training texts on which to test specific authorship hypothesis, one can only proceed in a way similar to the one used here. That is the case of *Tirant lo Blanc*.

In settings like the one of Shakespeare's drama, that are a lot more structured, one will typically want to use discriminant analysis tools to help determine authorship, instead of the approach taken here. If one is provided with lists of Shakespeare's preferred words, and of words that are more used by his contemporaries than by him, like the ones used in Craig and Kinney (2009), one could analyze the heterogeneity of style based on them. That would be similar to carrying out a discriminant analysis to attribute authorship. We intend to work on a paper presenting a more formal Bayesian discriminant analysis framework tailored to deal with authorship attribution and verification problems.

In the first and third case studies the results presented are based on twenty of the most frequent words. In both of these studies we also repeated the analysis using only the subset of these words that better discriminate between clusters according to what is found in Figures 4.6 and 4.10. We consider this sequential approach to selecting the list of words, starting with about twenty words and then repeating the analysis with the most discriminating words among them, to be very useful. Using far more than twenty

words to start with is usually problematic, because that includes in the analysis many words that do not distinguish between styles and hamper the classification power of the algorithm.

On a more technical level, note that when one bases heterogeneity analysis of style on word length and word counts, our predictive checks in case studies covering three different literatures indicate that finite mixtures of multinomial models capture most of the variability in the data. That settles the issue raised in Giron et al (2005) on whether or not this kind of models were flexible enough for typical stylometric data. In this setting, one does not need to resort to hierarchical models, like the finite mixtures of Dirichlet-multinomial models used in Puig and Ginebra (2014a), to account for any extra variability in the data.

Appendix: Computation of marginal likelihoods, $P(y|M_s)$

Under the single cluster model, M_1, the marginal likelihood is:

$$p(y|M_1) = \prod_{r=1}^{R} \frac{\prod_{i=1}^{n} N_i^r!}{\prod_{j=1}^{k(r)} \prod_{i=1}^{n} y_{ij}^r!} \frac{\prod_{j=1}^{k(r)} (\sum_{i=1}^{n} y_{ij}^r)!}{(\sum_{i=1}^{n} N_i^r)!} \text{Dir-Mult}(y_r; \sum_{i=1}^{n} N_i^r, a^r), \tag{4.4}$$

where y_r is the vector of aggregated counts of the r-th table, $y_r = (\sum_{i=1}^{n} y_{i1}^r, \ldots, \sum_{i=1}^{n} y_{ik}^r)$, and where Dir-Mult$(x; N, a)$ denotes the pdf of a Dirichlet-multinomial distribution with parameters N and $a = (a_1, \ldots, a_k)$ evaluated at $x = (x_1, \ldots, x_k)$,

$$\text{Dir-Mult}(x; N, a) = \frac{N! \Gamma(\sum_{j=1}^{k} a_j)}{\Gamma(N + \sum_{j=1}^{k} a_j)} \prod_{j=1}^{k} \frac{\Gamma(x_j + a_j)}{x_j! \Gamma(a_j)}. \tag{4.5}$$

The marginal likelihood under the S-cluster model, M_S, is

$$p(y|M_S) = \prod_{r=1}^{R} \frac{\prod_{i=1}^{n} N_i^r!}{\prod_{j=1}^{k(r)} \prod_{i=1}^{n} y_{ij}^r!} \prod_{s=1}^{S} \frac{\prod_{j=1}^{k(r)} (\sum_{i=1}^{n} y_{ij}^r I_{[\hat{\zeta}_i = s]})!}{(\sum_{i=1}^{n} N_i^r I_{[\hat{\zeta}_i = s]})!} \text{Dir-Mult}(y_r^{[\hat{\zeta}_i = s]}; \sum_{i=1}^{n} N_i^r I_{[\hat{\zeta}_i = s]}, a_s^r), \tag{4.6}$$

where $I_{[\hat{\zeta}_i = s]}$ denotes the indicator function that is 1 when the i-th observation is estimated to belong to the s-th cluster and it is 0 otherwise, and where $y_r^{[\hat{\zeta}_i = s]}$ denotes the vector of aggregated counts of all the observations estimated to belong to the s-cluster, $y_r^{[\hat{\zeta}_i = s]} = (\sum_{i=1}^{n} y_{i1}^r I_{[\hat{\zeta}_i = s]}, \ldots, \sum_{i=1}^{n} y_{ik}^r I_{[\hat{\zeta}_i = s]})$.

Chapter 5

Unified Approach to Authorship Attribution and Verification

In authorship attribution problems one needs to assign a text or a set of texts from an unknown author to either one of two or more candidate authors on the basis of the comparison of the disputed texts with texts known to have been written by the candidate authors. In authorship verification problems one needs to decide whether a text or a set of texts could have been written by a given single author or not. These two problems are usually treated separately. By assuming an open-set classification framework for the attribution problem, contemplating the possibility that neither one of the candidate authors is the unknown author, the verification problem becomes a special case of attribution problem. Here both problems are posed as a formal Bayesian multinomial model selection problem and are given a closed form solution, tailored for categorical data and naturally incorporating text length in the analysis. The approach to the verification problem is illustrated by exploring whether a court ruling sentence could have been written by the judge that signs it or not, and the approach to the attribution problem is illustrated by revisiting the authorship attribution of the Federalist papers and through a simulation study.

5.1 Introduction

The statistical analysis of literary style has long been used to characterize the style of texts and authors, and to help settle authorship attribution problems. Early work (see, e.g., Mendelhall, 1887, or Yule, 1938) used word length and sentence length to

characterize literary style. Other characteristics widely used for this purpose have been the proportion of nouns, articles, adjectives or adverbs, the frequency of use of function words, which are independent of the context, or of characters, and the richness and diversity of vocabulary.

Early applications involved the study of literary, religious or legal texts, but recently lots of new challenging problems have appeared due to widespread availability of electronic texts, leading for example to new applications in homeland security, computer forensics or spam detection. Good reviews about the statistical analysis of literary style can be found in Holmes (1985, 94, 98, 99). The range of statistical methods used in this setting is wide, but they most often involve various approaches to classification.

In the analysis of the heterogeneity of the style in a given text or set of texts, one does not always know how many authors might have contributed to the text, and one typically does not have a reference set of candidate authors and training texts. In these settings one needs to resort to cluster analysis techniques, also recognized as unsupervised classification/learning. A Bayesian approach to the analysis of the heterogeneity of style using mixtures of multinomial models is presented in Giron et al (2005).

Instead, in this manuscript one deals with authorship attribution problems, where one has a set of S candidate authors, and for each one of these authors one has a set of texts known to have been written by him or her, recognized as training texts. With the help of these training texts, one needs to assign a text or several texts by an unknown author to either one of the authors in the set. As a consequence, in these settings one needs to resort to the use of discriminant analysis techniques, also recognized as supervised classification/learning.

In most authorship attribution applications one adapts a closed-set classification framework, assuming that one knows with certainty that the unknown author is one of the S hypothesized candidates. Instead, nothing is lost by adopting a more prudent and flexible open-set classification framework contemplating as an extra hypothesis the possibility that the unknown author is not included among the list of S candidate authors. By adopting this open-set classification framework, the authorship verification problem that requires one to decide whether a text or a set of texts of unknown author have been written by a known author with comparable texts, becomes a special case of authorship attribution with $S = 1$.

In this paper we address the open-set authorship attribution and the verification problems using stylometric characteristics that involve counting features that are categorical and have a fixed number of categories, and are frequently observed. That covers count-

ing word lengths, sentence lengths, letters, function words, nouns or adjectives. Our approach excludes the analysis based on the word frequency counts used in vocabulary richness and diversity analysis, because the number of categories in such type of data is the frequency of the most frequent word, which typically grows with text size.

By restricting attention to such stylometric features, data will consist of a contingency table with as many rows as texts under consideration. The "training rows" will correspond to the texts that are known to belong to one of the S candidate authors, and the remaining rows will correspond to the texts of unknown author.

A huge variety of statistical tools have been used to tackle authorship attribution and verification problems. Even though Mosteller and Wallace (1964, 84) used probability models to drive the authorship attribution in one of the earliest seminal authorship study, most of that literature resorts to ad-hoc heuristic classifiers using linear or quadratic discriminant analysis (Stamatatos et al, 2000, Tambouratzis et al, 2004), support vector machines (Joachims, 1998, Diederich et al, 2003, Li et al, 2006), decision trees (Zheng et al, 2006), neural networks (Matthews and Merriam, 1993, Merriam and Matthews, 1994, Tweedie et al, 1996) or other machine learning based feature selection algorithms (Forsyth and Holmes, 1996, Forman, 2003, Binongo, 2003, Koppel et al, 2006). Recent applications of these supervised classification tools in authorship problems can be found, for example, in Stamatatos et al (2001), Holmes et al (2001), Burrows (2002, 2007), Hoover (2001, 2004), Abbasi and Chen (2005), Chaski (2005), Grant (2007), Argamon (2008), or Holmes and Crofts (2010). Recent reviews can be found in Stamatatatos (2009) and in Sebastiani (2002), and recent comparisons of some of these classification approaches in Zhao and Zobel (2005), Juola et al (2006), Yu (2008), Jockers et al (2008), Jockers and Witten (2010)

One of the shortcomings of most of these algorithmic based approaches is that they implicitly assume data to be continuous, or at least are tuned to work best when data is continuous. But the data in authorship attribution problems is mostly categorical, and one should adapt to the specificities of that kind of data. In particular, one needs to adequately take into account the length of texts and to accommodate for the dependence between the counts of different categories of a given stylometric characteristic, which is not easy to do in the framework of most of the classifiers typically used in authorship attribution.

Another shortcoming of the algorithmic based approaches advocated for in machine learning is that they are tailored to work with large training samples, and hence do not tend to fare well when one has a small number of training texts as it is often the case in authorship attribution practice. Furthermore, they can not accommodate for the

classification of disputed texts to unknown authors, without training texts, and therefore they can not be used in an open-set classification framework.

Here we adopt a formal Bayesian model based approach, in the spirit of Mosteller and Wallace (1984), that addresses all these shortcomings and it allows one to assess the uncertainty in the classification of the disputed texts as belonging to each one of the candidate authors. Adopting the Bayesian framework allows one to assign the disputed texts to either one of the S candidate authors or to neither one of them based on the posterior probability that disputed texts were written by each one of the candidate authors. Note also that Bayesian models are probabilistic models, and building them is like building a data simulation model. Hence, resorting to them allows one to check the assumptions on which the analysis is based by comparing the data observed with the data simulated from the selected model. That is in stark contrast with alternative algorithmic approaches that do not make explicit the stochastic assumptions on which they are grounded.

One of the strengths of the specific Bayesian approach adopted here is that, different from the approach taken by Mosteller and Wallace, here the whole vector of counts is analyzed simultaneously, instead of analyzing the count for each category separately. A second strength of our approach is that it provides closed form expressions for the posterior probabilities used to assign the disputed texts to an author, and hence they can be evaluated without the need to resort to iterative algorithms or to heuristic approximations to these posterior probabilities, as in other solutions to these classification problems.

To illustrate our approach, an authorship verification case study involving a court ruling sentence is presented, and the authorship attribution of the Federalist papers is revisited. There is a growing agreement that the frequency of high frequency function words is one of the most reliable features in authorship attribution (see, e.g., Hoover, 2003, Zhao and Zobel, 2005, Uzuner and Katz, 2005, Grieve, 2007). Even though word length has rarely proven to be useful in the authorship attribution of texts written in English, it has been found to be useful for texts in other languages (see, e.g., Giron et al, 2005). In the verification case study involving court rulings written in Spanish, the problem will be tackled through the analysis of word lengths and of the use of the most frequent function words, while in the Federalist papers case study, one focuses on the use of frequent function word counts.

A small simulation experiment is also carried out to help assess the performance of our Bayesian model driven approach under repeated use and to compare it to three of the main alternative approaches available for the authorship attribution problem.

5.2 Bayesian model building

5.2.1 Description of the model

In authorship attribution problems one starts with n^0 disputed texts that are assumed to have been written by the same unknown author, and with S potential authors for these texts. One also has n^s texts that are comparable to the disputed ones and are known to belong to the s-th candidate author, for $s = 1, \ldots, S$. In order for texts to be comparable, ideally they all should have been written at around the same time, belong to the same genre and deal with a similar topic, even though in practice that might be difficult to attain.

Given a stylometric characteristic that involves counting features that are categorical with a fixed number of categories, k, like counting the appearance of the $k = 25$ most frequent function words, the i-th text of the unknown author will become a vector valued categorical observation, $y_i^0 = (y_{i1}^0, \ldots, y_{ik}^0)$, for $i = 1, \ldots, n^0$, where y_{ij}^0 is the number of counts of the j-th category (the j-th most frequent word) in the i-th disputed text. Analogously, the i-th text known to be by the s-th author will yield the vector of counts $y_i^s = (y_{i1}^s, \ldots, y_{ik}^s)$, for $i = 1, \ldots, n^s$. Table 5.1 presents two examples of the kind of data that one will be dealing with in this paper, with each row of the table corresponding to either one of the training or one of the disputed texts, and playing the role of a y_i^s or a y_i^0 observation.

The set of all the n^0 vector valued observations corresponding to the n^0 disputed texts, denoted $y^0 = (y_1^0, \ldots, y_{n^0}^0)$, are assumed to be conditionally independent and multinomially distributed, $\prod_{i=1}^{n^0} \text{Mult}(y_i^0; N_i^0, \theta^0)$, where $N_i^0 = \sum_{j=1}^k y_{ij}^0$ is the total count for the i-th disputed text, and where $\theta^0 = (\theta_1^0, \ldots, \theta_k^0)$ with θ_j^0 being the probability of the j-th category for all the disputed texts, and hence with $\sum_{j=1}^k \theta_j^0 = 1$. Analogously, the set of all the n^s observations that correspond to the n^s texts known to be by the s-th author, $y^s = (y_1^s, \ldots, y_{n^s}^s)$, are assumed to be $\prod_{i=1}^{n^s} \text{Mult}(y_i^s; N_i^s, \theta^s)$ distributed, with $N_i^s = \sum_{j=1}^k y_{ij}^s$ and $\theta^s = (\theta_1^s, \ldots, \theta_k^s)$, where $\sum_{j=1}^k \theta_j^s = 1$.

When one is willing to assume that all the n^0 disputed texts share the same multinomial parameter θ^0, which is an assumption that will have to be checked, nothing is lost by combining all these n^0 texts into a single text and work with the vector of aggregated counts, $y_0 = (\sum_{i=1}^{n^0} y_{i1}^0, \ldots, \sum_{i=1}^{n^0} y_{ik}^0)$, that is known to follow a $\text{Mult}(y_0; N^0, \theta^0)$ distribution, where now $N^0 = \sum_{i=1}^{n^0} N_i^0$ is the total count of words in the texts by the disputed author. Analogously, if all the observations that correspond to texts by the s-th author are indeed conditionally independent and multinomially distributed, and do share the

same θ^s, which again is an assumption that should be checked, nothing is lost by working with the corresponding vector of aggregated counts, $y_s = (\sum_{i=1}^{n^s} y_{i1}^s, \ldots, \sum_{i=1}^{n^s} y_{ik}^s)$, that follows a $\text{Mult}(y_s; N^s, \theta^s)$ distribution, where $N^s = \sum_{i=1}^{n^s} N_i^s$.

If the author of the disputed texts was the s-th candidate author for some $s \in \{1, \ldots, S\}$, then one expects that the distribution of the aggregated counts in the disputed texts, y_0, will be distributed as the aggregated counts of texts by that author and hence have a $\text{Mult}(y_0; N^0, \theta^0 = \theta^s)$ distribution. If one further assumes that the sample counts of all texts are conditionally independent, then the probability density function of the whole set of data, $y = (y_0, y_1, \ldots, y_S)$, will be:

$$p_s(y|\theta^1, \ldots, \theta^S) = \text{Mult}(y_0; N^0, \theta^s)\text{Mult}(y_s; N^s, \theta^s) \prod_{r=1, r\neq s}^{S} \text{Mult}(y_r; N^r, \theta^r), \tag{5.1}$$

which will be recognized from now on as the M_s model.

In most authorship attribution studies one adopts a closed-set classification framework, where one acts as if one had the certainty that the unknown author was one of the S candidates. In that case, one would only consider the $M_1, \ldots, M_S$ models.

Instead, in the open-set classification setting adopted here one also contemplates the possibility that the author of the disputed texts might not be included in the set of S candidate authors. That requires one to consider an extra $(S+1)$-th sub-model, M_0, under which $\theta^0 \neq \theta^s$ for $s = 1, \ldots, S$, and hence with pdf:

$$p_0(y|\theta^0, \theta^1, \ldots, \theta^S) = \text{Mult}(y_0; N^0, \theta^0) \prod_{s=1}^{S} \text{Mult}(y_s; N^s, \theta^s). \tag{5.2}$$

In this open-set classification framework, determining whether the disputed texts were written by either one of the S candidate authors and hence share his or her style, or by someone else, becomes the problem of choosing one model among $M_0, M_1, \ldots, M_S$, in the light of data.

The framework covered by the $S = 1$ case corresponds to the authorship verification problems, requiring one to choose between the model M_1, indicating that the single candidate author has written both the disputed texts as well as the training texts, and the model M_0, indicating that the disputed texts were written by someone else.

In a Bayesian setting, one needs to choose a distribution for the parameters of the model that captures what one knows about them before observing the data, which is denoted as the prior distribution. As a prior distribution for the multinomial probabilities, θ^r, for

$r = 0, 1, \ldots, S$, it will be assumed that they are independent and Dirichlet$(a_1^r, \ldots, a_k^r)$ distributed, where $a^r = (a_1^r, \ldots, a_k^r)$ is such that $a_j^r > 0$. Depending on the values chosen for a^r, the prior will capture different types of information and it will be more or less informative. In particular, the expected value of θ^r will be $(a_1^r, \ldots, a_k^r)/(\sum_{j=1}^k a_j^r)$, and one can choose the a_j^r to reflect the fact that some categories might be known to appear with larger probabilities than others. That is the often the case, for example, when one is modeling word frequencies. Also, the larger $\sum_{j=1}^k a_j^r$ the smaller the variances of θ_j^r and the more informative the prior chosen for θ^r.

Choosing this prior distribution is convenient, because it leads to closed form expressions for the posterior probabilities of each one of the $S + 1$ sub-models, which will be key in selecting a model and hence an author for the disputed texts. In the examples that follow all the $a^r = (a_1^r, \ldots, a_k^r)$ are set to be equal to $(1, \ldots, 1)$, which corresponds to assuming a uniform distribution on the simplex for θ^r. The amount of information in this prior distribution is equivalent to the one in a sample text with a count total of $N = k$. Given that the total number of words (counts) in the texts analyzed will always be a lot larger than k, by choosing the uniform prior distribution the influence of the prior on the posterior distribution will always be a lot weaker than the influence of the data on the posterior through the likelihood function. As a consequence, varying the parameters of the prior distribution around the chosen $(1, \ldots, 1)$ does not alter the conclusions of the analysis.

It will also be assumed that all $S + 1$ sub-models, and hence all $S + 1$ authorship hypotheses, are equally likely a priori, and hence that their prior probabilities are $P(M_r) = 1/(S + 1)$ for $r = 0, 1, \ldots, S$, but that can be trivially set to be otherwise.

5.2.2 Author selection through model selection

A difficulty of the heuristic algorithms is that they often lack a statistically well grounded method for selecting an author for the disputed texts. Here that problem is tackled first through the use of a formal model selection method, based on the posterior probability that each one of the models considered is the one active. Model checks will also be used to help support the choice of model, and hence of author.

Resorting to a Bayesian analysis has the advantage that one can update the prior probabilities and select the model (author) with the largest posterior probability. The posterior probability that the M_r model is the one generating the data is:

$$P(M_r|y) = \frac{P(M_r)P(y|M_r)}{\sum_{r=0}^{S} P(M_r)P(y|M_r)}, \text{ for } r = 0, 1, \ldots, S, \tag{5.3}$$

where $P(M_r)$ is the prior probability of model r and where $P(y|M_r)$ is the density function of the prior predictive distribution under model M_r evaluated at the observed data, also recognized as the marginal likelihood of M_r. Hence, the posterior probability of M_r is proportional to $P(M_r)$ and $P(y|M_r)$. One will select the model (author) with the largest posterior probability, and when each model (author) is considered equally likely a priori, that means picking the M_r with the largest marginal likelihood, $P(y|M_r)$.

Often, computing $P(y|M_r)$ exactly is too complicated to be attempted in practice, and one approximates its logarithm through the BIC, or through the MCMC simulations used to update the model. But in our case, by choosing a Dirichlet prior for the multinomial probabilities one has a closed form expressions for $P(y|M_r)$, that can be easily evaluated. In particular, when $y = (y_0, y_1, \ldots, y_S)$ one has that:

$$p(y|M_0) = \text{Dir-Mult}(y_0; N^0, a^0) \prod_{s=1}^{S} \text{Dir-Mult}(y_s; N^s, a^s), \tag{5.4}$$

where $\text{Dir-Mult}(x; N, a)$ denotes the pdf of a Dirichlet-multinomial distribution with parameters N and $a = (a_1, \ldots, a_k)$ evaluated at $x = (x_1, \ldots, x_k)$,

$$\text{Dir-Mult}(x; N, a) = \frac{N!\Gamma(\sum_{j=1}^{k} a_j)}{\Gamma(N + \sum_{j=1}^{k} a_j)} \prod_{j=1}^{k} \frac{\Gamma(x_j + a_j)}{x_j!\Gamma(a_j)}. \tag{5.5}$$

The marginal likelihood under M_r for $r \in \{1, \ldots, S\}$ becomes:

$$p(y|M_r) = \frac{N^0!N^r!}{(N^0 + N^r)!} \frac{\prod_{j=1}^{k}(\sum_{i=1}^{n^0} y_{ij}^0 + \sum_{i=1}^{n^r} y_{ij}^r)!}{\prod_{j=1}^{k}(\sum_{i=1}^{n^0} y_{ij}^0)! \prod_{j=1}^{k}(\sum_{i=1}^{n^r} y_{ij}^r)!} \times \tag{5.6}$$

$$\text{Dir-Mult}(y_0 + y_r; N^0 + N^r, a^r) \prod_{s=1, s\neq r}^{S} \text{Dir-Mult}(y_s; N^s, a^s). \tag{5.7}$$

In this way, one can compute $P(y|M_r)$, and hence $P(M_r|y)$, exactly, and select the model (author) with the largest $P(M_r|y)$. That allows one to classify the disputed texts as either belonging to the r-th author, when $P(M_r|y)$ is the largest with $r \in \{1, \ldots, S\}$, or as having an author not in the list, when $P(M_0|y)$ is the largest.

Note that here one is computing the exact posterior probabilities, $P(M_r|y)$, conditional on the training as well as the disputed texts, $y = (y_0, y_1, \ldots, y_S)$, based on the simultaneous use of all these texts counts. That is different from taking an approximate two-stage approach, first "estimating" the posterior distribution of the multinomial probabilities θ^r of the r-th author for $r = 1, \ldots, S$, based only on the counts in the training texts by that author, y_r, and using (2.3) with $y = y_0$ and replacing $P(y = y_0|M_r)$ by an approximation

$P(y = y_0|\hat{\theta}^r)$, where $\hat{\theta}^r$ is an estimate of θ^r. One often uses the maximum likelihood estimate of θ^r, which is also the posterior mode under a uniform prior. Examples of the use of this approximate Bayesian approach can be found in Gale et al (1993), McCallum and Nigan (1998), Lewis (1998), Schneider (2003), or Peng et al (2004). Note that this two-stage approximation can not be used in the open-set classification framework adopted here.

5.2.3 Model checking

The solution given here to the authorship attribution and verification problems relies on the model comparison just described, which in turn relies on the assumption that the model considered is correct. Before standing by the conclusions reached, one should check whether that model does indeed capture all the relevant features in the data or not.

The main model assumption is that all the vectors with the counts of the texts by the same author, s, are conditionally independent and distributed as a $\text{Mult}(N_i, \theta^s)$, where the multinomial parameter θ^s is identical for all the texts by that author. Even though inference is made after aggregating all texts by the same author in a single text, to check that assumption one needs to resort back to the sample of n^s vectors of counts, $y_1^s, \ldots, y_{n^s}^s$, or the texts available for each author before aggregation. The two most likely deviations from that assumption, and the way to check them, are:

1. The style of one or several of the texts attributed to the s-th author might not be comparable to the style of the other texts by him, or might not even be by that author. In such a situation, some of the observation(s) assumed to be from the s-th author, y_i^s, for $i = 1, \ldots, n^s$, might be independent and multinomially distributed but with different and unrelated multinomial parameter values.

   To verify whether all the n^s texts assumed to be comparable and by the same author are indeed so, one can verify whether each one of them is by that author by treating the other $n^s - 1$ texts as training set. That is, one would go author by author, and resort to the $S = 1$ special case of the model in Section 5.2.1 to test whether each training text shares the same style (model) as the other training texts by that author. This use of the solution to the verification problem to check this model assumption will be illustrated in the two case studies that follow.

2. The vectors of counts y_i^s, for $i = 1, \ldots, n^s$, corresponding to the training texts from the s-th author, might be multinomially distributed with similar but not identical values of θ_i^s. That leads to the count data from the s-th author being more dispersed than anticipated by (2.1) or (2.2). If these θ_i^s can be assumed

to be exchangeable and follow a given distribution, one can improve the model by switching from the purely multinomial models considered here to multinomial mixtures instead.

To check whether the vector of counts for the texts of a given author are identically distributed as a multinomial or not, one can assess whether it is plausible that one could simulate data like the data observed through the predictive distributions (see, e.g., Gelman et al, 2004). We do not report on the predictive checks carried out in the examples that follow, but in them it was found that the purely multinomial based models in Section 5.2.1 match closely the variability of the counts observed.

5.3 Authorship verification case study

Here, one compares the style of a Spanish patent court ruling sentence, denoted by D, with the style of four other patent court ruling sentences written at around the same time and dealing with similar issues, denoted by S_1, S_2, S_3 and S_4. Even though all the five sentences considered were signed by the same judge, there is grounded suspicion that the disputed sentence was actually written by someone else. The goal is to verify whether the style of the disputed sentence is similar enough to the style of the other four sentences to back the single authorship hypothesis or not.

In order to verify whether that is the case, the comparison will be based both on word length distribution, as well as on the frequency with which the twenty most frequent function words are used in these sentences. Before counting the number of l-lettered words and the number of times function words appear in the sentences, we have excluded from the text all citations, acronyms, capital lettered words, numbers, dates and names of persons and of cities. On top of that, we have only considered the factual, the legal basis and the final verdict, excluding from the analysis the formal paragraphs that are always repeated at the end of all sentences. These twenty most frequent function words are: *de, la, que, el, en, y, a, los, se, por, del, las, no, una, con, es, o, para, su* y *al*.

Note that, different from what happens in the authorship attribution problem case, with $S > 1$, in the authorship verification case, with $S = 1$, one can not choose the list of words or features based on their discriminating power, because one only has a single candidate author. This is, in fact, the only feature that distinguishes verification studies from attribution studies, other than the number of candidate authors involved.

The resulting data, on which the statistical analysis will be based, are partially presented in Table 5.1. The first row of the first sub-table for example indicates that in the disputed

word length counts											
court ruling	1	2	3	4	5	6	7	8	9	10+	N_i
D	**598**	**4069**	**1882**	**673**	**707**	**689**	**1145**	**997**	**737**	**1554**	**13051**
S_1	158	942	397	149	249	191	220	196	200	318	3020
S_2	629	2587	1200	450	690	573	631	579	680	1070	9089
S_3	186	978	413	160	257	192	241	198	224	316	3165
S_4	560	3049	1257	499	810	582	705	629	683	1126	9900
Function word counts											
court ruling	de	la	que	el	en	y	a	los	se	por	...
D	**1269**	**851**	**568**	**437**	**480**	**240**	**277**	**229**	**260**	**204**	**...**
S_1	310	184	107	129	85	67	39	34	54	56	...
S_2	806	509	392	297	289	236	192	144	147	116	...
S_3	320	202	115	143	77	77	58	36	62	61	...
S_4	1067	642	376	312	317	214	147	164	157	137	...

Table 5.1: Number of l-lettered words for $l = 1, 2, \ldots, 9$ and for $l > 9$, and number of times that the ten most frequent words appear in the sentences. D is the disputed sentence, and S_1, S_2, S_3 and S_4 is a training set of comparable sentences signed by the same judge that also signed D.

sentence, D, there are 598 one-lettered words, 4069 two-lettered words and so on, and that one has considered a total of 13051 words. The first row of the second sub-table indicates that the most frequent word in that disputed sentence is *de*, appearing 1269 times, the second most frequent word is *la*, appearing 851 times and so on. The remaining rows of that table have the counts for the four training sentences, known to have been written by the judge signing the disputed one. Note that if all the texts had been written by the same author, one might expect all the rows in each sub-table to come from the same multinomial distribution, and hence the model M_1 in (2.1) holds. If instead, the distribution of the first row is different from the distribution of the other four rows and hence the model M_0 in (2.2) holds, it indicates that its style is different and hence the disputed sentence might very well have been written by a different person.

Figure 5.1 compares the proportion of l-lettered words observed in the disputed sentence D with the proportion observed in the other four sentences, S_1 to S_4. It indicates that the proportion of words of 3, 4, 7, 8 and more than nine letters in the D sentence is the largest, and the proportion of words of 1, 5, 6 and 9 letters in D is the smallest of all the five sentences considered. Figure 5.2 compares the frequency of appearance of the twenty most frequent words in sentence D with the one observed in the other four sentences. Note that the frequency of appearance of *que, en, a, los, las* and *no* in D is the highest, and the frequency of *y, con, o* and *su* in D is the lowest among all the five

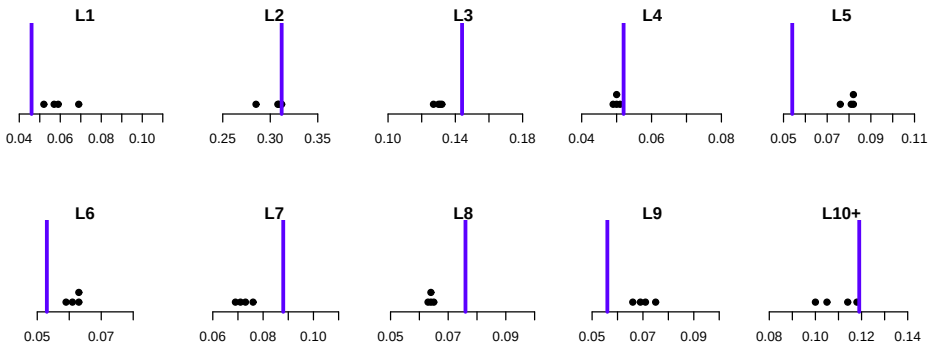


Figure 5.1: Dots indicate the proportion of l-lettered words, Ll, observed in the four training sentences, S_1 to S_4. Lines indicate the proportions observed in the disputed sentence, D.

Sentence	word length	function words
S_1	1.00	1.00
S_2	0.99	1.00
S_3	1.00	1.00
S_4	1.00	1.00
D	0.00	0.00

Table 5.2: Posterior probability that the style of a sentence is the same as the style in the other ones, $P(M_1|y)$. D is not used in the first four rows, checking whether S_1 to S_4 share style.

sentences considered.

In order to check first whether all the four sentences used as a training sample of the style of the known judge, S_1 to S_4, are comparable and do indeed have a similar style and hence can all be safely attributed to that judge, we compare each one of them with the other three sentences in that sample, excluding the disputed sentence D.

The first four rows of Table 5.2 present $P(M_1|y) = 1 - P(M_0|y)$, which is the probability that the counts for the corresponding S_i sentence shares the same multinomial distribution as the counts obtained by adding up the other three rows of the sub-table that correspond to the remaining training texts, S_j with $j \neq i$, and hence that all the four training sentences share the same style. Note that the probability that the distribution observed in each of the undisputed training sentences is the same as in the other undisputed training sentences is basically equal to one. This is consistent with the hypotheses that these four sentences all share the same style, and hence that a single author wrote

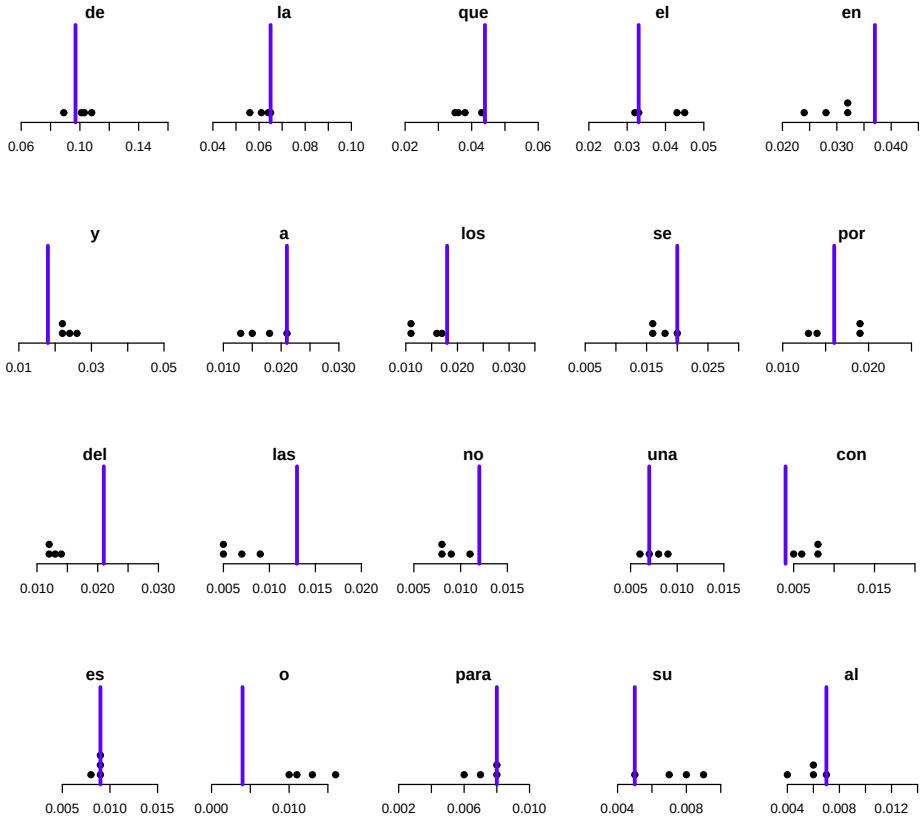


Figure 5.2: Dots indicate the frequency of appearance of the twenty most frequent function words in the four training sentences, S_1 to S_4. Lines indicate the frequency of appearance observed in the disputed sentence, D.

them, the judge signing them.

But the style in S_1 to S_4 seems to be very different from the style for the disputed sentence D. The word length and the word count distributions of D is compared with the corresponding distributions of the four training sentences, S_1 to S_4, by computing the probability that the counts for D in Table 5.1 share the same multinomial distribution as the counts obtained by adding up the other four rows of the sub-table, $P(M_1|y)$. According to the last row in Table 5.2, that probability is zero under both features considered.

That indicates that these distributions are clearly different, and hence that the style of the disputed sentence is very different from the style of the remaining sentences. That is consistent with what is observed in Figures 5.1 and 5.2, comparing the actual counts in D, with the counts in S_1 to S_4. Hence it is likely that the disputed sentence was actually

written by someone other than the one signing it.

5.4 Authorship attribution case study

The Federalist papers were published anonymously between 1787 and 1788 by Alexander Hamilton, John Jay, and James Madison to persuade New Yorkers to adopt a new constitution of the United States. Of the seventy seven essays, having somewhere between 900 and 3500 words each, it is generally agreed that Jay wrote five, Hamilton wrote forty three, Madison wrote fourteen, and three papers are known to be the joint work of Madison and Hamilton. That leaves the twelve papers, numbered 49 to 58, 62 and 63, which is not clear whether were written by Hamilton or by Madison.

Mosteller and Wallace (1964, 84) carried extensive comparisons of the frequencies of a carefully chosen set of common words in writings known to be by Hamilton and by Madison, with the frequencies of these words on the twelve disputed papers. That seminal case study involves a clearly defined set of candidate authors, with a clear set of texts known to be by them and which are comparable to the disputed ones. That explains that the federalist papers soon became a benchmark on which alternative authorship attribution approaches test themselves. Recent studies re-visiting that problem are, for example Holmes and Forsyth (1995), Martindale and McKenzie (1995), Tweedie et al. (1996), Bosch and Smith (1998), Khmelev and Tweedie (2001), Collins et al (2004), and Jockers and Witten (2010).

Our approach to authorship attribution is Bayesian, as the one taken by Mosteller and Wallace, but it is different from the one taken by them in that we model the whole vector of counts jointly, using multinomial distributions instead of modeling each count separately assuming that they were independent with a Poisson or a negative binomial distribution. Analyzing the whole vector of counts simultaneously, instead of the individual counts of each category separately, allows one to take into consideration the dependency that one always has between the counts of different categories. A second difference with respect to the analysis by Mosteller and Wallace is that we take the open-set classification approach described in Section 5.2, instead of a closed-set approach.

Mosteller and Wallace (1964, 84) tentatively explores the use of word length as a way to help determine authorship, but concludes that this feature is of no use when distinguishing Hamilton and Madison styles. Our analysis have confirmed that fact.

Hence, in this case study we focus the analysis on word counts. Different from what happens in authorship verification studies, where there is a single candidate author and

a single set of training texts, when one has more than one candidate author one has the privilege of picking up a list of words that best discriminate among them. Mosteller and Wallace (1964, 84) base their main analysis on the counts of thirty frequent words that are assessed to discriminate best between the style of Madison and the style of Hamilton using both federalist papers as well as external texts known to have been written by these authors.

Besides carrying out our authorship attribution analysis based on the thirty words used by Mosteller and Wallace, we have also carried out parallel analysis based on two new lists of words. The first list contains the twenty function words that are most frequent in the federalist papers, without taking into consideration their discriminating power. The second list consists of thirty function words that we found to be most discriminant between the forty three federalist papers known to be by Hamilton and the fourteen federalist papers known to be by Madison, without using any texts external to the federalist papers.

In order to select our list of the thirty most discriminant words, we started with the list of 200 most frequent words in the papers known to be by Hamilton and the 200 most frequent words in the papers known to be by Madison. By merging these two lists, one obtains a set of 240 different words. In order to assess the discriminating power of each one of these words, we modeled the 240-dimensional vector with the counts of these words in the papers by Hamilton, y^H, and the corresponding vector with the counts in the papers by Madison, y^M, as:

$$p(y^H, y^M|\theta^H, \theta^M) = Mult(y^H; N^H, \theta^H)Mult(y^M; N^M, \theta^M) \tag{5.8}$$

where θ^H and θ^M are the multinomial probability vectors modeling the relative frequency of these words in the papers by Hamilton and by Madison, and where N^H and N^M are the sum of the counts of these words in these papers. As a prior distribution on θ^H and θ^M, one uses the same one used for θ^r in Section 5.2. Words are then ranked from having better discriminating power to having worse discriminating power based on the statistic:

$$T_i = \left| \frac{E(\log \frac{\theta_i^H}{\theta_i^M}|y^H, y^M)}{\sqrt{Var(\log \frac{\theta_i^H}{\theta_i^M}|y^H, y^M)}} \right| \tag{5.9}$$

where i is the index identifying each word in the list of 240 words.

The thirty words with the largest T_i were selected, after discarding the ones that clearly depended on context. The list of words selected in this manner, together with the value of the corresponding T_i between brackets, were: *on (10,73), would (8,16), upon (7,69), there (7,54), by (7,47), to (6,94), and (6,81), the (5,42), these (4,82), in (4,39), at*

(4,19), latter (4,16), several (3,96), I (3,8), if (3,69), might (3,62), any (3,51), kind (3,48), had (3,46), between (3,45), those (3,34), an (3,2), he (3,19), this (3,19), very (3,17), against (3,12), no (2,95), were (2,9), into (2,89) and *same (2,88).*

Only eight of our thirty most discriminating words obtained based only on Federalist papers, (*an, by, kind, on, there, this, to* and *upon*), appear also in the list of Mosteller and Wallace thirty most discriminating words obtained based on texts by Hamilton and Madison different from the Federalist papers. Figure 5.3 compares the frequencies of appearance of our thirty most discriminating words in the federalist papers by Hamilton and by Madison, and the corresponding frequencies of appearance in the twelve disputed Federalist papers.

In order to check whether all the forty three federalist papers used as a training sample of the style of Hamilton are comparable and do indeed have a similar style, one verifies whether each one of these papers has a style that is similar to the style of the other forty two papers by Hamilton. Using the same approach as the one in the case study in Section 5.3 on each one of these papers separately, one classifies all of them as belonging to Hamilton, with probability one. When one repeats the same verification exercise on each one of the fourteen federalist papers used as training samples of Madison, one also classifies all of them as belonging to Madison with probability one.

text	Unknown	Hamilton	Madison
49	0.	0.	1.
50	0.	0.	1.
51	0.	0.	1.
52	0.	0.	1.
53	0.	0.	1.
54	0.	0.	1.
55	0.	.59	.41
56	0.	0.	1.
57	0.	0.	1.
58	0.	0.	1.
62	0.	0.	1.
63	0.	0.	1.

Table 5.3: Posterior probabilities of the three authorship hypotheses considered for each one of the disputed papers, based on the analysis of the vector with the counts of our set of thirty most discriminant words.

To settle the authorship attribution of the twelve disputed texts, we carried out the analysis described in Section 5.2 on each one of these twelve papers separately, considering as

tentative hypothesis that each one of then had been authored by Hamilton, by Madison, or by an unknown someone else. Table 5.3 presents the posterior probabilities of each one of these three hypothesis for each one of the twelve disputed papers based on our set of thirty most discriminating words. From these probabilities it is clear that all the disputed papers, except paper 55, should be clearly attributed to Madison. Figure 5.3 indicates what is that makes the style of paper 55 different from the style of the rest of disputed papers by Madison, and closer to the style of the papers by Hamilton. This might indicate the collaboration of Madison and Hamilton in the writing of that paper.

When one repeats the same type of analysis based on the use of the thirty most discriminating words used by Mosteller and Wallace, the only difference is that the posterior probability that paper 55 follows Hamilton style is .05 instead of .59. When one bases the same analysis on the use of the twenty most frequent function words instead, without filtering out the words that do not discriminate between Hamilton and Madison, one finds that all the disputed papers except papers 49 and 55 are again clearly attributed to Madison with probability close to one. All these findings are in close agreement with the ones in Mosteller and Wallace (1964, 84), and in the other studies looking into this authorship problem.

5.5 Simulation study

To assess the performance of the Bayesian multinomial model driven classification method proposed above, and to compare it to alternative supervised classification techniques, two perfectly known simulation scenarios are designed. In the first scenario, word length data from five training texts by Author 1 and from five training texts by Author 2 are simulated, to be used to help settle the authorship attribution of three disputed texts, D1, D2 and DU. In the second simulation scenario, word length data from fifty texts by Author 1 and from fifty texts by Author 2 are simulated, to be used to settle the authorship of texts D1, D2 and DU. All texts in the simulation exercise are set to have $N = 500$ words.

The multinomial probabilities used to simulate the word length data by Author 1 are $\theta^1 = (.04, .17, .22, .20, .14, .09, .06, .04, .02, .02)$, while the probabilities used for Author 2 are $\theta^2 = (.035, .16, .23, .19, .15, .095, .065, .045, .015, .015)$. The disputed text D1 is simulated to be by Author 1, and hence with $\theta^0 = \theta^1$, the disputed text D2 is simulated to be by Author 2, and hence with $\theta^0 = \theta^2$, and the disputed text DU is simulated to neither be by Author 1 nor by Author 2, with $\theta^0 = (.07, .13, .17, .15, .13, .11, .09, .06, .05, .04, .07)$.

Under each one of these two simulation scenarios, one first checks how our authorship

attribution method behaves under repeated use. Second, one compares the performance of our method with the performance of three popular methods being used in supervised classification. In both cases, the assessment will be based on repeating the two simulation experiments described above 1000 times, each time simulating the word length data of all the training texts as well as the word length data of the three disputed texts.

To assess how the Bayesian multinomial approach fares under repeated use, Figure 5.4 presents the histograms of the 1000 posterior probabilities of the three authorship hypotheses, (Author is 1, Author is 2, and Author is neither 1 nor 2 and hence unknown), for each one of the three disputed papers under the two simulation scenarios.

In the case of the disputed text D1, which we know it to be by Author 1, in 733 (824) of the 1000 realizations for the 5 training texts (50 training texts) scenario one finds that the posterior probability that it is by Author 1 is the largest one, while in 267 (176) of these realizations one finds that the probability that it is by Author 2 is the largest one. In almost all these 1000 sample realizations, these two posterior probabilities are far from 0 or 1, due to the fact that the styles of Authors 1 and 2 are set to be similar, which makes the classification problem significantly more difficult than the ones in the case studies in Sections 5.3 and 5.4. In contrast, Figure 5.4 indicates that all 1000 realizations lead to a posterior probability close to 0 that D1 is by an unknown author, and hence that it is neither by Author 1 nor by Author 2. Something similar is observed through the histograms of the posterior probabilities for the disputed D2 text.

Instead, the style of the disputed DU text is purposely set to be very different from the styles of Authors 1 and 2, and therefore in most (but not in all) the 1000 realizations our multinomial model driven method assigns a posterior probability close to 1 that the author is neither 1 nor 2, and hence close to 0 that it is by Author 1 or by Author 2. The scenario with 50 training texts per author is a bit more conclusive than the one with 5 training texts, as one would expect it to be.

Next, our Bayesian multinomial model driven method is compared to a decision tree classification method, to a support vector machine method and to a logistic regression method. To do that, the three alternative methods together with the method proposed in this manuscript are used to classify each one of the 1000 realizations of the D1, D2 and DU disputed texts based on each one of the corresponding 1000 realizations of the training texts. And that is done again under both simulation scenarios.

For a description on how the alternative classification methods work, see Chapters 4, 8 and 9 of Gareth et al (2014). To implement the decision tree method, the tree() function from the tree library in R has been used, to implement the support vector machine

method, the svm() function from the e1071 library has been used, and to implement the logistic regression method, the glm() function has been used. The optimal level of model complexity under each one of these three approaches has been determined through cross validation.

By restricting consideration to texts that have 500 words, one avoids the need to decide how to incorporate text length in these three alternative analysis, which is an issue not adequately settled in authorship attribution practice. Note also that these alternative approaches are tailored to work with large training samples and hence with many training texts, which is not what one has in our first simulation scenario, with only five training texts per author. In contrast, the Bayesian multinomial model driven approach advocated for in this manuscript naturally incorporates text size in the analysis, and it works well with any number of training samples, including instances with a single training text.

Table 5.4 presents the proportion of times each one of the three disputed texts is correctly attributed to the author that actually wrote it. These proportions are estimates of the long run (frequentist) probability that the method correctly classifies the disputed text to the actual author. The first row of that table, for example, indicates that the decision tree approach correctly classifies D1 to be by Author 1 in 639 out of the 1000 realizations, the support vector machine approach does that 588 times and the logistic regression approach does that 653 times, all compared to the 733 times that the Bayesian multinomial approach correctly classifies D1. Different from the Bayesian multinomial method, the three top-of-the-counter alternative supervised classification approaches considered here do not allow for an open-set classification framework, because they can not handle the hypothesis that neither Author 1 nor Author 2 wrote a text. Hence, no proportion of correct classifications can be provided for DU under these alternative approaches.

Table 5.4 indicates that the Bayesian multinomial method implemented with a uniform prior for the multinomial parameters performs better than the logistic regression based approach and that, in turn, the logistic regression approach performs better than the decision tree and the support vector machine based approaches. The performance of the three alternative methods considered is specially poor in the five training texts per author scenario, because they are designed to work with many training samples and not just a few.

When text length, N_i, and/or the number of training samples increase, the authorship attribution problem becomes easier, and one finds that the performance of the logistic regression and of the support vector machine methods becomes closer to the performance of the Bayesian multinomial model driven method. We have repeated this kind of sim-

ulation exercise under many other simulation scenarios, and using different alternative classification methods, reaching similar conclusions.

5 training texts per author				
text	BM	DT	SVM	LR
D1	0.733	0.639	0.588	0.653
D2	0.717	0.577	0.584	0.616
DU	0.946	–	–	–
50 training texts per author				
text	BM	DT	SVM	LR
D1	0.824	0.671	0.784	0.793
D2	0.816	0.674	0.704	0.793
DU	0.989	–	–	–

Table 5.4: Estimated probability of correct classification under the Bayesian multinomial method (BM), under a decision tree method (DT), under a support vector machine method (SVM), and under a logistic regression method (LR). The first three rows correspond to the five training texts per author scenario and the last three to the fifty training texts per author scenario.

5.6 Final Comments

Different from the algorithmic based supervised classification methods typically used for authorship attribution, the Bayesian multinomial model driven approach advocated for here has the advantage of being tailored for categorical data, of naturally incorporating text size in the analysis, of adequately dealing with settings with a small number of training texts, and of easily adapting to an open-set classification context. On top of that, it also comes with the scientific advantage of making explicit the list of distributional assumptions on which the conclusions of the analysis are based; by checking whether those assumptions are adequate, one can check the validity of the analysis carried out.

Even though the presentation has focused on the use of word length and of word counts, and it has only been illustrated with examples with at most two candidate authors, our approach naturally extends to any stylometric characteristic with a fixed number of categories, and to any number of candidate authors. In the authorship attribution (verification) analysis proposed here, one carries out as many separate Bayesian discriminant analysis as stylometric characteristics used. Instead, one could also implement a single discriminant analysis combining the information of all the characteristics at once, by extending the models in Section 5.2 to apply to the analysis of several contingency tables at once.

Even though the main goal in authorship attribution is to classify the disputed texts by making inference about M_r, one can also benefit from exploring the posterior distributions for $(\theta^0, \theta^1, \ldots, \theta^S)$, to help characterize what distinguishes the style of authors.

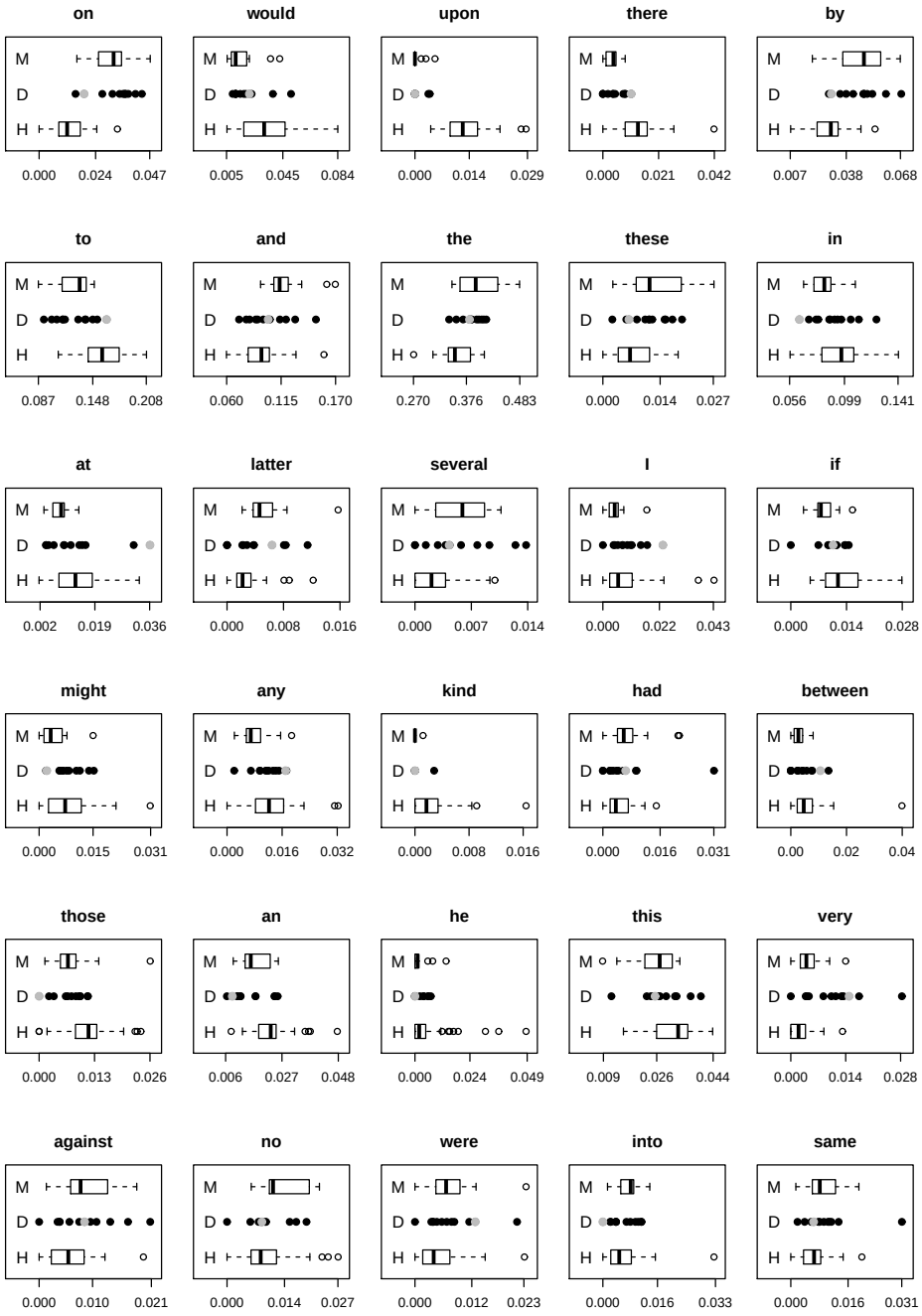


Figure 5.3: Comparison of the frequencies of appearance of the thirty most discriminating words in the papers known to be by Hamilton and by Madison, and in the twelve disputed papers. The counts for the disputed paper 55, with a style closer to Hamilton than to Madison are shaded lighter.

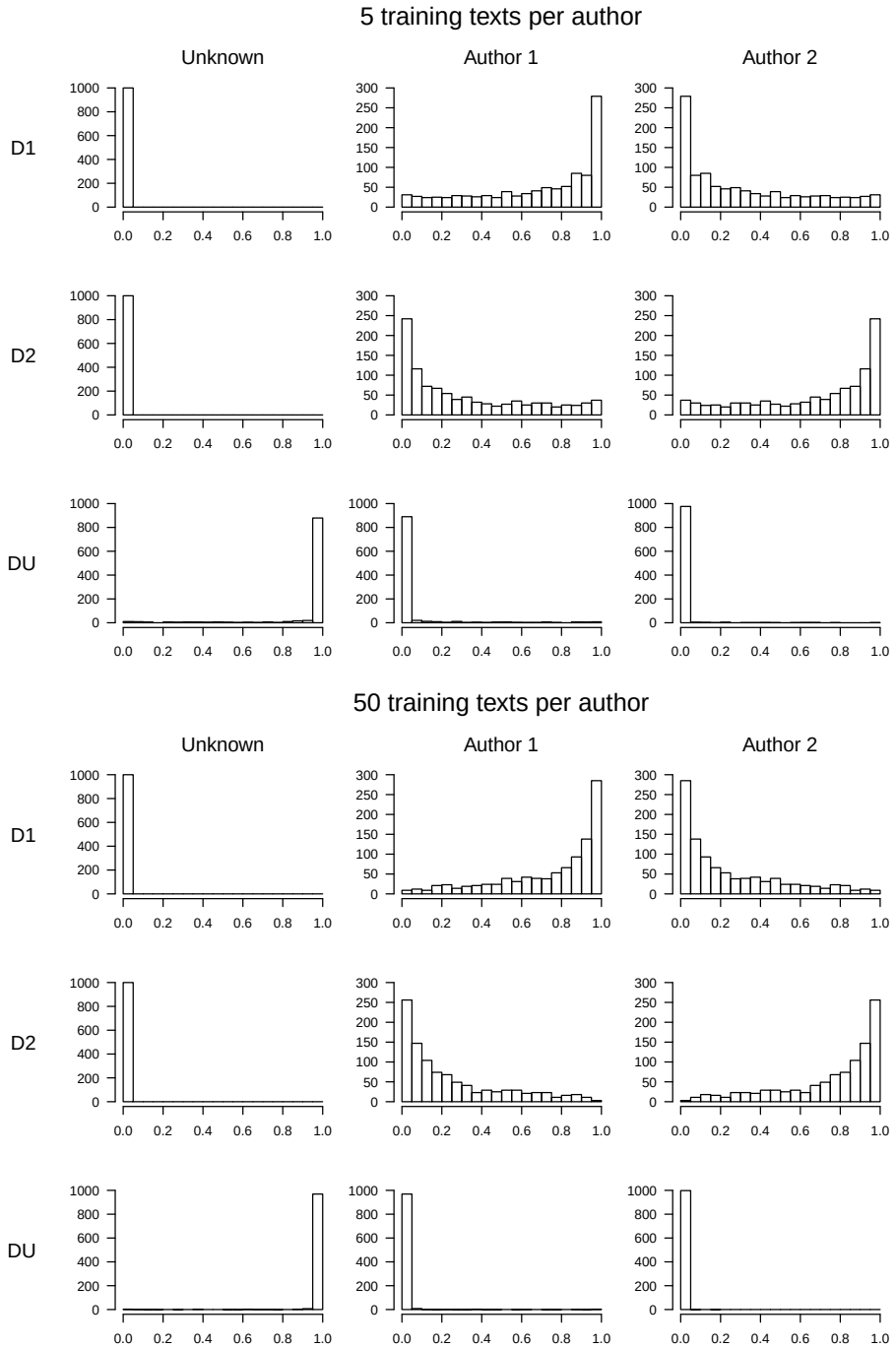


Figure 5.4: Histogram of the sample of 1000 posterior probabilities of the three authorship hypotheses, with D1 being by Author 1 and thus having $\theta^0 = \theta^1$, with D2 being by Author 2 and thus having $\theta^0 = \theta^2$, and with DU being by an unknown author.

Chapter 6

Future Work

Listed below are some topics related to this thesis on which we have done some ground work and on which we intend to continue working in the future.

6.1 Extension of the methods in Chapter 2 by using a three parameter mixing distributions

1. Extend of the IG mixing distribution used in Chapter 2 to three parameters mixing distributions, as the Generalized Inverse Gausian (GIG) and Tweedie distribution, that include the IG as a special case. This extension is called for in these instances where texts are large, because we find the IG based models fail to fit data properly

   Sichel (1975,1986a,1986b,1997) developes a very complete and useful non Bayesian methodology for the analysis of frequency count data based on the IG- and the GIG-Poisson mixture models. Many authors, like Pollatschek and Radday (1981), Holmes (1992), Holmes and Forsyth (1995), Baayen (2001), Riba and Ginebra (2006), Puig, Ginebra and Perez-Casany (2009) and Puig, Ginebra and Font(2010) build on that methodology.

   About Tweedie and the resulting Tweedie-Poisson, the framework for the analysis of frequency count data has not yet been developed but one expects that switching from using the GIG to using the Tweedie as mixing distribution might have some advantages. A complete characterization of this distribution would be needed, which would require to:

a) Isolate the role of the parameters of the Tweedie mixing model from the role of the text size, to be able to estimate the probability density of the word frequencies of the author through estimate of that mixing distribution.

b) Provide an interpretation of the parameters of the Tweedie mixing model in terms of the size, evenness and diversity of the vocabulary of the author and in terms of the overdispersion in the data. One of the main advantages of using the Tweedie-Poisson model instead of the GIG-Poisson model lies in the interpretation of its third parameter.

c) Find eficient ways to estimate the parameters of both the Tweedie-Poisson model as well as the one for the zero truncated Tweedie-Poisson model, and to find efficient ways of estimating the uncertainty of those estimates.

d) Find the way to estimate and represent the density of the Tweedie mixing distribution, which is not as trivial as it might seem because there is no analytic closed form expression for that density and one has to rely on the Fourier inversion of its characteristic function (see Dunn(2008)). This is extremely useful in stylometry (ecology) because this density can be used as an estimate of the density of the word (species) frequencies distribution which can be used as a fingerprint of the style of the author (cosystem) in his texts (samples).

e) Word frequency count data are zero-truncated. Aspects like the extension of the parameter space due to truncation and the effect of switching the mixing and the detruncation stages will have to be taken into consideration.

2. Perform a Bayesian frequency count data analysis based on the GIG-Poisson and on the Tweedie-Poisson models.

   Chapter 2 shows a whole methodology for the Bayesian analysis based on the truncated IG-Poisson model. In the future we intend to implement a Bayesian analysis based on the GIG-Poison model and on the Tweedie-Poisson model.

   a) For the Bayesian analysis based on the GIG-Poisson mixture model, one can take advantage of the fact that the generalized inverse Gaussian distribution can be seen as a model playing the role of the prior distribution of the parameter of the Poisson and as a prior it is a conjugate one. Our first goal is to obtain a closed form expression for the posterior distribution of the parameter, taking advantage of the fact that the posterior predictive distribution in the GIG-Poisson model that can be obtained in closed form. Note that having a closed form for the density of the Poisson mixture allows one to get the posterior distribution when we are using the mixing distribution as prior distribution for the parameter of the Poisson.

For the bayesian analysis based on the Tweedie-Poisson model, the scenario is more complex because the lack of a closed form expression for the distribution function of the Tweedie model means that it is not possible to obtain analytical closed form of the posterior distribution. Hence we will need to develop MCMC algorithms that simulate from it, and from the correspondent posterior predictive distribution.

These analysis will provide a generalization of the classic conjugate bayesian analysis which use a Gamma as prior distribution. Note that gamma is a limiting case of the Generalized Inverse Gaussian distribution and a particular case of Tweedie family of distributions. The extra flexibility of the GIG and of the Tweedie together with large degrees of skewness allowed by them make them excellent candidates for a non-informative reference Bayesian analysis.

b) In practice one will have to choose a prior on the parameters of the GIG model and on the Tweedie model because the goal of the frequency count data analysis is to estimate them, and not the Poisson parameter. To implement this Bayesian analysis we would have to go to the R and WinBUGS computational tricks learned when implementing the Bayesian analysis based on the IG-Poisson model. WinBUGS is no longer developed though, and hence other alternatives might need to be considered.

Implementing this Bayesian analysis will also require that we enhance all the Bayesian model checking techniques that we already developed for the IG-Poisson Bayesian model, so that they better fit the analysis based on the GIG- and Tweedie-Poisson models. We also intend to find the ways to compute the DIC of these models, and friendly graphical ways to present the results of our analysis.

c) Finally we also plan to implement Bayesian hierarchical generalization of the non-hierarchical approach.

3. Performance comparison between Poisson mixture Models.

a) Compare the performance of the three parameter (truncated) Tweedie-Poisson model with the performance of some of its two parameter submodels, like the (truncated) negative binomial and the (truncated) IG-Poisson models, on a wide array of sample texts.

b) Compare the performance of the (truncated) GIG-Poisson model with the performance of the (truncated) Tweedie-Poisson model on a wide spectrum of word frequency count data.

c) Explore the performance of the untruncated Tweedie-Poisson model on untruncated frequency count data, like insurance claims frequency count data, and compare it with the performance of the untruncated GIG-Poisson model.

6.2 Cluster analysis of frequency count data

1. Bayesian cluster analysis of frequency of word frequency data.

   Giron, Ginebra and Riba (2005) implements a Bayesian cluster analysis of multinomial data based on the non-hierarchical Dirichlet-Multinomial model and Puig (2009) extends that analysis basing it on a hierarchical Dirichlet-Multinomial model. Here we use these models for word length counts and more frequent function words counts but not for the frequency of word frequency that can be modeled by IG-Poison mixture models and their proposed three parameter extensions.

   We intend to implement Bayesian cluster analysis of frequency count data that mimic the work already done for multinomial type data. To do that we will take advantage of all the tools developed for the IG-Poisson and planned for the GIG-Poisson and the Tweedie-Poisson models under the homogeneous single population case.

   To implement this cluster analysis of frequency count data we will have to learn how to:

   a) Simulate from the posterior and from the predictive posterior distribution,

   b) Implement useful posterior predictive checks,

   c) Find ways to present the results in a friendly graphical maner, usually through clever graphs.

2. Here we use Dirichlet-Multinomial cluster models for simultaneous analysis of word length counts and most frequent function words counts, but this idea of simultaneous analysis of more than one contingency table is not limited to the use of a single reference model like the Dirichlet-Multinomial.We can extent it to the Poisson mixtures described above. Then the frequency of word frequency, word length counts and most frequent function word counts could be analyzed simultaneously.

3. A typical problem when simulating from a Bayesian cluster model is label switching, which occurs as a result of the symmetry in the likelihood of the model parameters. Recent studies have focused in this problem, trying to remove the symmetry by using artificial identifiability constraints, but that does not solve the problem. This problem makes interpretation the MCMC chains difficult. Here we reject all the simulations with label switching problems that could not be fixed through simple relabeling. But that problem becomes harder to solve with more than 3 cluster and that opens a way for new identifiability constraints research.

6.3 Extend the authorship attribution analysis

1. The authorship attribution (verification) analysis, proposed in Chapter 5, carries out as many separate Bayesian discriminant analysis as stylometric characteristics used. Instead, one could also implement a single discriminant analysis combining the information of all the characteristics at once, by extending the models to apply to the analysis of several contingency tables at once in a way analogous to the one used in Chapter 4 for cluster analysis.

Appendix A

Bayesian Computation with WinBUGS

To do our computations we use WinBUGS, a free software for Bayesian analisys of complex statistical models using Markov chain Monte Carlo (MCMC) methods. WinBUGS can be executed from R by means of R2WinBUGS library. The combination of WinBUGS and R, becomes a perfect platform to update our bayesian models.

WinBUGS has a powerful and flexible way to define models, which speeds up the process of building and refining an appropriate model. Really useful in modeling mixture multinomial models, as the one used in Chapters 3, 4 and 5. Unfortunately this ease of modeling is absent in the case of zero truncated Inverse Gaussian mixtures of poison distribution, used for modelling frequency count data in Chapter 2. These models are not easy to define. The problem is that these models are not in the list of WinBUGS models available by default.

All the simulations in this thesis were obtained with the last version of WinBUGS (1.4.3), released in August 2007. Unfortunately, WinBUGS is no longer updated. Although this version still remains available, it is expected that in the future other MCMC implementations will take among Bayesian data analysis practitioners.

A.1 Simulations on IG-Poisson mixture models

The main difficulty to perform simulations on models related withInverse Gaussian - Posisson mixture models was that this model is not among the list of WinBUGS models available and that on top of that we need a truncated version of an existing one. The list of distribution we need is:

- the zero truncated IG-Poisson model needed to update the bayesian model presented in Section 3 of Chapter 2.
- the IG-zerotruncated Poisson model needed to update the bayesian model presented in Section 4 of Chapter 2.

WinBUGS allow the user to define new sampling distributions by means of an advanced use of the BUGS language called "Zeros Triks". This method produces high autocorrelation, poor convergence and high MC error, and so it is computationally slow and long runs are necessary.

A harder but more precise way to solve this problem is to take advantage of WinBUGS Development Interface (WBDev), that allows restricted access to areas of the WinBUGS source that have been used for defining elements of the BUGS language. One can implement one's own sampling distributions, 'hard-wiring' them into the WinBUGS framework via compiled Pascal code. WBDev 'hard-wired' components can be computed much more quickly and can lead to more simplified, clearer and better interpretable WinBUGS code which reduces the possibility of making coding errors.

In this way, we implemented three WBDev components;

- dIGP.zerotrunc(b,c,N) for the the zero truncated IG-Poisson mixture
- dpoisson.zerotrunc(l) for the zero truncated Poisson
- dinverse.gaussian (b,c) for the inverse Gausian

For more details on definition of these modules see Section 3, 4 and 5 of Appendix A. With the model complexity hidden in the WBDev 'hard-wired' components, we obtain clear WINBUGS models for the zero truncated IG-Poisson and for the IG-zerotruncated Poisson.

- WinBUGS model for the bayesian zero truncated IG-Poisson:

```
model {
    for (i in 1:V) {
        y[i] ~ dIGP.zerotrunc (b,c,N)
    }
    b ~ dgamma(0.001,0.001)
    c ~ dgamma(0.001,0.001)
}
```

- WinBUGS model for the bayesian IG-truncated Poisson:

```
model {
    for (i in 1:V) {
        y[i] ~ dpoisson.zerotrunc (Npi[i])
        Npi[i]<-N*pi[i]
        pi[i] ~ dinverse.gaussian (b,c)
    }
    b ~ dgamma(0.001,0.001)
    c ~ dgamma(0.001,0.001)
}
```

We take special care with the selection of reference prior distributions. The usual prior for real positive parameters is the Gamma distribution. We select it as priori for the parameters b and c and we chose the hiperparameters alpha=beta=0.001, so that the priori does not impact significantly on the posteriori. We made a sensitivity study to make sure that they really were little informative.

Figure A.1 shows an example of trace plots of the sample values of two independent Markov chains. The convergence looks reasonable after only 500 warming iterations are needed. The initial values for the two chains are (b=0.01, c=0.01) and (b=0.5, c=0.1).

In all simulations three chains of simulations were run. No convergence problems were found. Convergence was quickly obtained after a few warming iterations. The use of 'hard-wiring' distributions slowed down the simulation speed. One simulation takes over 6h for 1000 iterations for the zero truncated IG-Poisson and over 8h for the IG-truncated Poisson.

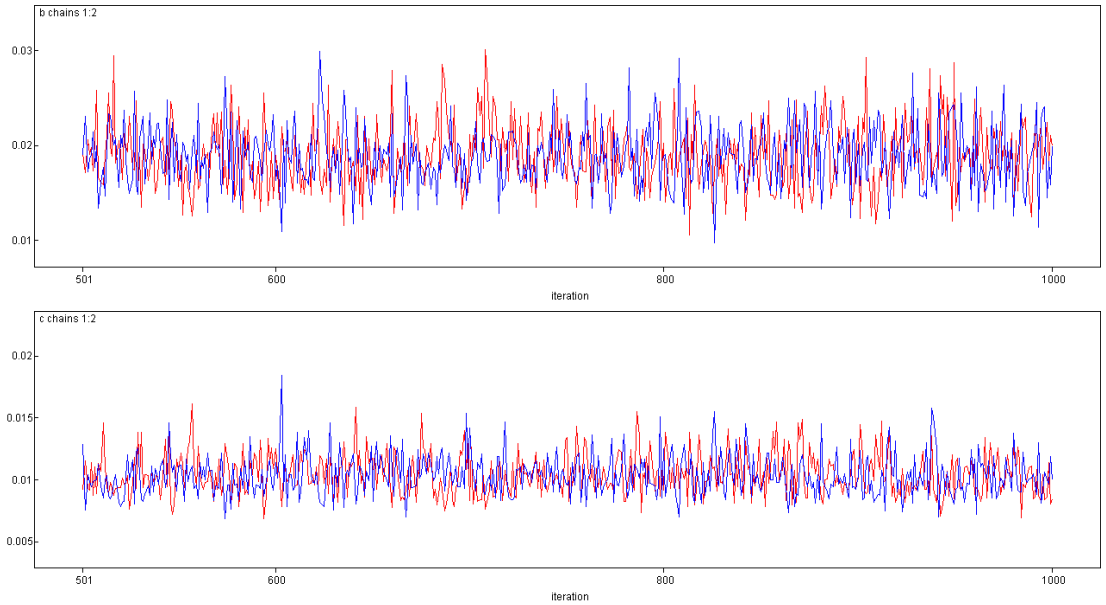


Figure A.1: Convergence check: trace for the parameters b and c, of the two computed Montecarlo chains, for the the zero truncated IG-Poisson mixture and for the frequency count data of Alice in Wonderland

Node stat.	mean	sd	MC error	2.5%	median	97.5%	start	sample
b	0.01896	0.003324	1.134E-4	0.01292	0.01882	0.02596	501	1000
c	0.01036	0.00164	6.123E-5	0.007779	0.01019	0.01406	501	1000

Table A.1: Summary for the parameters b and c, of the two computed Montecarlo chains, for the the zero truncated IG-Poisson mixture and for the frequency count data of Alice in Wonderland. They are based on 1000 simulations folowing 500 iterations of the warming period

One way to assess the accuracy of the posterior estimates, in Table A.1, is by calculating the Monte Carlo error for each parameter. This is an estimate of the difference between the mean of the sampled values (which we are using as our estimate of the posterior mean for each parameter) and the true posterior mean. As a rule of thumb, the simulation should be run until the Monte Carlo error for each parameter of interest is less than about 5% of the sample standard deviation. We can see in the example that the MC error fulfills this rule.

WinBUGS automatically implements the DIC model comparison criterion that trades off goodness-of-fit against model complexity by means of an effective number of parameters pD. This information has not been useful due to the different structure of the two models analyzed lead to very different values of pD and non comparable values of DIC. It happens because in WinBUGS one can not indicate in which level of the hierarchical model are the parameters on study.

A.2 Simulations on Multinomial cluster models

Because the distributions required to define these models are included in WinBUGS, the model definition was a simple task. A particular model have been established for the case of one cluster. Models for two or more clusters has the same structure.

- For the model of 1 cluster:

```
model {
    thetaL[1, 1:KL] ~ ddirch(alphaL[])
    thetaP[1, 1:KP] ~ ddirch(alphaP[])

    for (i in 1 : I) {
        z[i]<- 1

        L[i,1:KL] ~ dmulti( thetaL[z[i], 1:KL] , NL[i]  )
        NL[i] <- sum(L[i,])
        P[i,1:KP] ~ dmulti( thetaP[z[i], 1:KP] , NP[i]  )
        NP[i] <- sum(P[i,])
    }
}
```

- For the model of 2 cluster:

```
model {
    thetaL[1, 1:KL] ~ ddirch(alphaL[])
    thetaL[2, 1:KL] ~ ddirch(alphaL[])
    thetaP[1, 1:KP] ~ ddirch(alphaP[])
    thetaP[2, 1:KP] ~ ddirch(alphaP[])

    p[1:2] ~ ddirch(alpha2[])

    for (i in 1 : I) {
        z[i]~ dcat(p[1:2])

        L[i,1:KL] ~ dmulti( thetaL[z[i], 1:KL] , NL[i]  )
        NL[i] <- sum(L[i,])
        P[i,1:KP] ~ dmulti( thetaP[z[i], 1:KP] , NP[i]  )
        NP[i] <- sum(P[i,])
    }
}
```

It is also important to pay attention to the convergence of the chains resulting from the simulation. One must always verify that convergence has been achieved. If one has not achieved it, one needs to increase the number of warming simulations. A quick way to assess convergence is visual inspection of multiple chains ran in parallel with initial values randomly taken. This procedure also allowed us to highlight problems of identifiability which are frequent when the number of clusters increases.

In Figure B.1 a typical example of the identifiability is given. In one of the chains (in blue) labels of clusters 1 and 2 are switched with respect to the other two chains (in red and green). In this case the identifiability problem is easy to fix, because the index assignation to the clusters is stable inside each chain, and it is possible to relabel clusters. When problems of identifiability happen inside a chain, resulting in cluster labels switching, the simulations were rejected and one started trying again from other initial conditions.

We used 5000 warming iterations. After that convergence was generally obtained. Then the model was run for 20.000 iteration/chain x 3 chains = 60.000 iterations more, with a thinning parameter of 4 (only 1 of each 4 iteration was saved to the results file). As

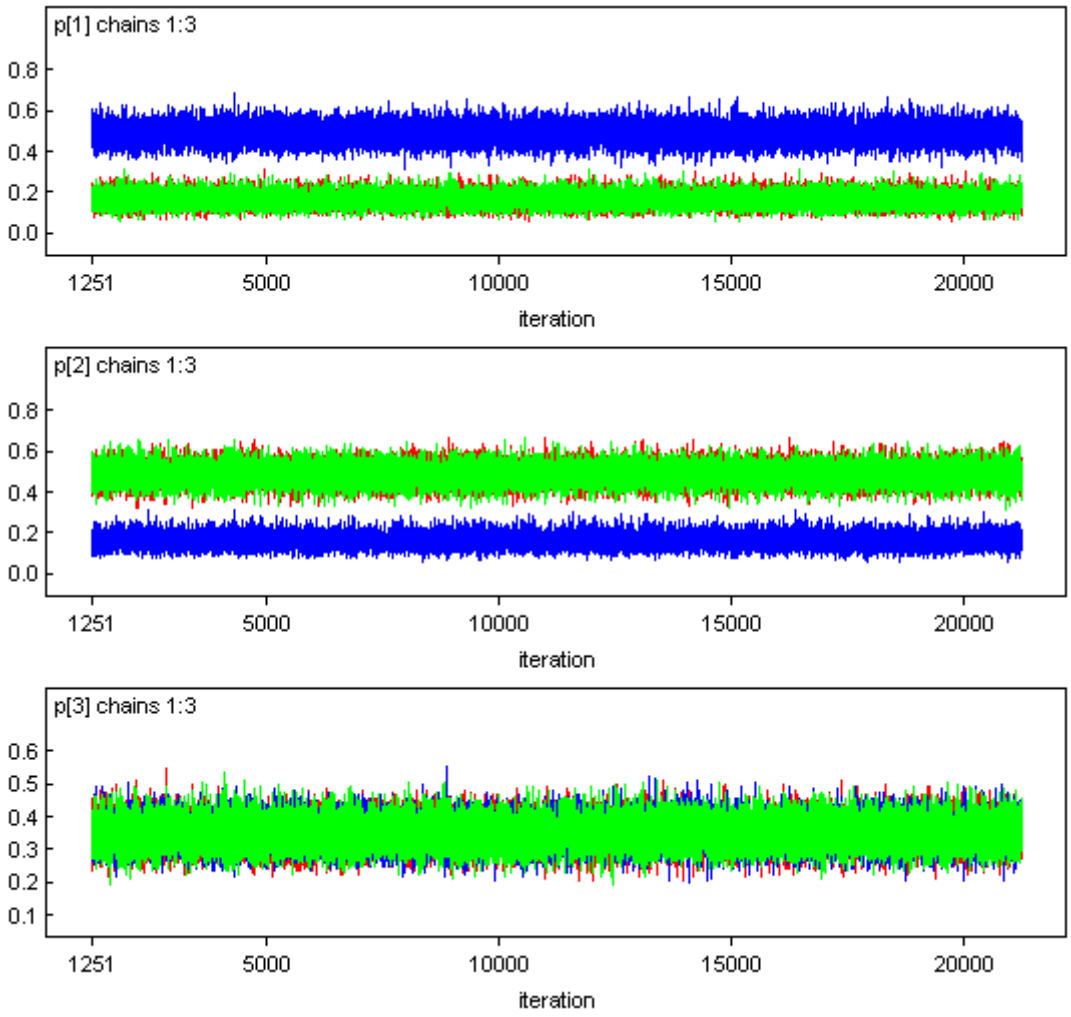


Figure A.2: Convergence check: trace for the parameters p[i], of the three computed Montecarlo chains, for the three cluster model for Don Quijote

initial values for the three chains we use a non informative dirichlet with initial values; $alphaL[i] = 1$ and $alphaP[i] = 1$ for the multinomial prior, an equiprobable distribution $p[i] = 1/s$ where s is the number of clusters for the relative size of each cluster and a random assignation to clusters 1 to s for the categorical variable that carries the assignation to the label of one cluster, $z[i]$.

As an example, the summary for the parameters are shown in Table A.2 for the two cluster model from Don Quijote chapters. The accuracy of the posterior estimates is assessed again by calculating the Monte Carlo error for each parameter and checking that it is less than about 5% of the sample standard deviation.

Node stat.	mean	sd	MC error	2.5%	median	97.5%	start	sample
p[1]	0.5679	0.04836	2.265E-4	0.4721	0.5683	0.6617	1251	60000
p[2]	0.4321	0.04836	2.265E-4	0.3384	0.4317	0.5279	1251	60000
thetaL[1,1]	0.07671	6.307E-4	2.697E-6	0.07547	0.07672	0.07796	1251	60000
thetaL[1,2]	0.2311	9.878E-4	4.076E-6	0.2292	0.2311	0.2331	1251	60000
thetaL[1,3]	0.1762	9.055E-4	3.868E-6	0.1744	0.1762	0.178	1251	60000
...	...	...	...	...	...	...	...	...
thetaL[2,1]	0.07708	6.478E-4	2.648E-6	0.0758	0.07708	0.07835	1251	60000
thetaL[2,2]	0.2371	0.001056	4.301E-6	0.2351	0.2371	0.2392	1251	60000
thetaL[2,3]	0.1727	9.163E-4	3.832E-6	0.1709	0.1727	0.1745	1251	60000
...	...	...	...	...	...	...	...	...
thetaP[1,1]	0.138	0.001355	5.728E-6	0.1354	0.138	0.1407	1251	60000
thetaP[1,2]	0.1274	0.0013	5.702E-6	0.1249	0.1274	0.13	1251	60000
thetaP[1,3]	0.1319	0.001408	6.608E-6	0.1292	0.1319	0.1347	1251	60000
...	...	...	...	...	...	...	...	...
thetaP[2,1]	0.156	0.001554	7.073E-6	0.1531	0.156	0.1592	1251	60000
thetaP[2,2]	0.1305	0.001356	5.325E-6	0.1278	0.1305	0.1332	1251	60000
thetaP[2,3]	0.1252	0.001392	6.5E-6	0.1225	0.1252	0.1279	1251	60000
...	...	...	...	...	...	...	...	...
z[1]	1.446	0.4971	0.0022	1.0	1.0	2.0	1251	60000
z[2]	1.0	0.0	2.357E-13	1.0	1.0	1.0	1251	60000
z[3]	1.0	0.0	2.357E-13	1.0	1.0	1.0	1251	60000
...	...	...	...	...	...	...	...	...

Table A.2: Summary of the three computed Montecarlo chains, for the two cluster multinomial model, both for word length counts and for the most frequent function words counts data from Don Quijote

Finally, the WinBUGS model for the Multinomial cluster model with dependence used in Chapter 3 is:

```
model{
    for (i in 1 : I) {
        Y[i,1:K] ~ dmulti(theta[index[i], 1:K] , N[i]  )
        N[i] <- sum(Y[i,])

        index[i] <- z[i]+1
        z[i] ~ dbern(p[i])

        logit(p[i]) <- h0 + h[i] + b[i]
        h[i] ~ dnorm(0, tau.h)

        for (k in 1:K) {
            AUX[i,k] <- Y[i,k]*log(theta[index[i],k])-logfact(Y[i,k])
        }

        LL[i] <-  logfact(N[i]) + sum(AUX[i,])
    }

    b[1:I] ~ car.normal(adj[], weights[], num[], tau.b)

    for(k in 1:sumNumNeigh) {
        weights[k] <-1
    }

    pz <- mean(z[])

    #  PRIORIS
    theta[1, 1:K]~ddirch(alpha[])
    theta[2, 1:K]~ddirch(alpha[])

    h0 ~ dnorm(0,0.1)

    tau.h ~ dgamma(20, 0.1)I(0.0001,100000) # taula01
    tau.b ~ dgamma(3, 0.1)I(0.01,1000) # taula01

    L <- sum(LL[])
}
```

A.3 WinBUGS Development Interface (WBDev) Implementing new univariate distributions

With WebDev (Winbugs Development Interface) one can implement new custom distributions. This section summarizes the steps required to do it. For complete instructions on how to add new univariate distributions to WinBUGS by "hardwiring" them into the system, see the document "WinBUGS Development Interface (WBDev) Implementing your own univariate distributions" available on the WBDev website (*http://winbugs-development.mrc-bsu.cam.ac.uk/*).

Computer code for a new distribution have to be defined by a Component Pascal module .odc. Then one have to set up the system so that Component Pascal code can be compiled with the source code of WinBUGS. For it, one needs to install the BlackBox Component Builder (*http://www.oberon.ch/blackbox.html*). Once it is installed it is included a module, named *UnivariateTemplate.odc*, that could be use as a template. As an example of adding a new distribution, this template defines the zero truncated normal distribution. We have started from this template to define our new distributions, changing only the necessary parts of code on it.

The following instructions should be followed when defining a new WinBUGS distribution via the template:

1. Choose a name for the new component, *NewDistribution*. Then save the template under the new name, *WBDev/Mod/NewDistribution.odc*

2. Now modify the code in the new module according to the desired distributional form, declare the types of arguments required and redefine this procedures:

   - DeclareProperties(.), this procedure is used to specify two important pieces of information about the new distribution. First, whether the distribution is discrete or continuous (isDiscrete = "TRUE"/"FALSE"); and, second, whether or not we can evaluate its cumulative distribution function (canIntegrate = "TRUE"/"FALSE")
   - NaturalBounds(.), this procedure should specify the natural bounds of the new distribution.
   - LogFullLikelihood(.),LogPropLikelihood(.),LogPrior(.), these procedures all return the natural logarithm of a number that is proportional to the probability density function evaluated at the current value. The reason for having

three procedures that all do essentially the same thing is that WinBUGS doesnt always require the same level of "exactness". Sometimes WinBUGS needs the log-pdf specifying exactly, in which case the LogFullLikelihood(.) procedure is called by the core software. Other times, normalizing constants can be ignored, in which case LogPropLikelihood(.) is called. Often, however, only those factors of the pdf that are functions of the value are needed. Then the software calls the LogPrior(.) procedure. Of course, as there is no harm done in including normalizing constants when they are not actually required, one can always simply call LogFullLikelihood(.) from within both LogPropLikelihood(.) and LogPrior(.) to save coding. However, considerable gains in efficiency can often be made by avoiding unnecessary calculations, especially in cases where normalizing constants are cumbersome to calculate.

- Cumulative(.), this procedure should be used to return the value of the new distributions cumulative distribution function at the real-valued input parameter. In cases we can not evaluate it, one could specify *"canIntegrate := FALSE;"* in the *DeclareProperties(.)* procedure to skip it.
- DrawSample(.), this procedure should return, a sample from the new distribution

3. Once the new module has been successfully compiled (and saved) then it can be linked into the WinBUGS software by modifying the file *WBDev/Rsrc/Distributions.odc*. The first line of this file contains the required entry for the truncated normal distribution defined in the *WBDevUnivariateTemplate* module:

   $s \sim$ "*dnew.distribution*"$(s, s)I(s, s)$ "$WBDevNewDistribution.Install$"

A.4 WBDev implementation of the Inverse Gaussian (IG) model

There are many different parameterizations of the inverse Gaussian distribution. For this implementation we use the one given by Tweedie (1956) with two parameters $\nu \in (0, \infty)$ and $\lambda \in (0, \infty)$. The probability density function of the inverse Gaussian distribution, $IG(\nu, \lambda)$, is:

$$f(x|\nu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} e^{-\lambda \frac{(x-\nu)^2}{2\nu^2 x}}.$$

The log-likehood of one observation x is:

$$logL(\nu, \lambda|x) = 0.5\left(\ln(\lambda) - \ln(2\pi) - 3\ln(x)\right) - \frac{\lambda(x-\nu)^2}{2\nu^2 x},$$

this expression is needed for the definition of *LogFullLikelihood(.)*,*LogPropLikelihood(.)* and *LogPrior(.)* procedures.

The cumulative distribution function is:

$$F(x|\nu, \lambda)) = \Phi\left[\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\nu} - 1\right)\right] + e^{2\lambda/\nu}\Phi\left[\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\nu} + 1\right)\right],$$

where $\Phi[]$ is the standard normal distribution function, that is available as *WBDevSpecfunc.Phi(.)* function in the WBDev environment. This expression is used for the definition of the *Cumulative(.)* procedure.

Seshadri V (1993) gives a method for simulate from a Inverse Gaussian, which it is based on a general procedure for sampling that starts in finding a $Y = \Psi(X)$ that follows a well known distribution. In this case the used distribution is a Chi-square of one degree of freedom:

$$Y = \Psi(X) = \frac{\lambda(X-\nu)^2}{\nu^2 X} \sim \chi^2_1$$

Applying this methodology the steps for generating random numbers distributed as an Inverse Gaussian, $IG(\nu, \lambda)$, are:

1. Sample a random value y from a chi-square distribution of one degree of freedom; $Y \sim \chi^2_1$
2. Calculate $x_1 = \nu + \frac{\nu^2 y}{2\lambda} - \frac{\nu}{2\lambda}\sqrt{4\nu\lambda y + \nu^2 y^2}$
3. Sample a random value u from an uniform $[0, 1]$; $U \sim uniform[0, 1]$
4. If $u \leq \frac{\nu}{\nu + x_1}$ then $x = x_1$, otherwise $x = \frac{\nu^2}{x_1}$

Then x is a a random value from a $X \sim IG(\nu, \lambda)$

A.4.1 Source code for the odc module for the Inverse Gaussian model

```
(*1*)   MODULE WBDevInversaGaussianaMF;

        IMPORT
            WBDevUnivariate,
(*2*)           WBDevRandnum, WBDevSpecfunc,
(*3*)           Math;
        CONST
(*4*)           location = 0; inverseScale = 1;
        TYPE
            StdNode = POINTER TO RECORD (WBDevUnivariate.StdNode) END;
            Left = POINTER TO RECORD (WBDevUnivariate.Left) END;
            Right = POINTER TO RECORD (WBDevUnivariate.Right) END;
            Interval = POINTER TO RECORD (WBDevUnivariate.Interval) END;
            Factory = POINTER TO RECORD (WBDevUnivariate.Factory) END;
        VAR
(*5*)           log2Pi: REAL;
            fact-: WBDevUnivariate.Factory;

(*6*)       PROCEDURE DeclareArgTypes (OUT args: ARRAY OF CHAR);
(*7*)       BEGIN
(*8*)           args := "ss";
(*9*)       END DeclareArgTypes;

(*10*)      PROCEDURE DeclareProperties (OUT isDiscrete,canIntegrate: BOOLEAN);
(*11*)      BEGIN
```

```
(*12*)          isDiscrete:= FALSE;
(*13*)          canIntegrate := TRUE;
(*14*)      END DeclareProperties;

(*15*)      PROCEDURE NaturalBounds (node: WBDevUnivariate.Node; OUT lower, upper: REAL);
(*16*)      BEGIN
(*17*)          lower := 0;
(*18*)          upper := INF;
(*19*)      END NaturalBounds;

(*20*)      PROCEDURE LogFullLikelihood (node: WBDevUnivariate.Node; OUT value: REAL);
(*21*)      VAR
(*22*)          x, nu, lam: REAL;
(*23*)      BEGIN
(*24*)          x := node.value;
(*25*)          nu := node.arguments[location][0].Value();
(*26*)          lam := node.arguments[inverseScale][0].Value();
(*27*)          value :=0.5*(Math.Ln(lam)-log2Pi-3*Math.Ln(x))
                - lam*((x-nu)*(x-nu)/(2*nu*nu*x));
(*28*)          value := value;
(*29*)      END LogFullLikelihood;

(*30*)      PROCEDURE LogPropLikelihood (node: WBDevUnivariate.Node; OUT value: REAL);
(*31*)      BEGIN
(*32*)          LogFullLikelihood(node, value);
(*33*)      END LogPropLikelihood;

(*34*)      PROCEDURE LogPrior (node: WBDevUnivariate.Node; OUT value: REAL);
(*35*)      VAR
(*36*)          x, nu, lam: REAL;
(*37*)      BEGIN
(*38*)          x := node.value;
(*39*)          nu := node.arguments[location][0].Value();
(*40*)          lam := node.arguments[inverseScale][0].Value();
(*41*)          value := 0.5*(Math.Ln(lam)-log2Pi-3*Math.Ln(x))-
lam*((x-nu)*(x-nu)/(2*nu*nu*x));
(*42*)      END LogPrior;

(*43*)      PROCEDURE Cumulative (node: WBDevUnivariate.Node; x:
REAL; OUT value: REAL);
(*44*)      VAR
(*45*)          nu, lam, v1, v2, v3: REAL;
(*46*)      BEGIN
(*47*)          (*      HALT(126);*)
(*48*)          nu := node.arguments[location][0].Value();
(*49*)          lam := node.arguments[inverseScale][0].Value();
(*50*)          v1 := Math.Sqrt(lam/x)*((x/nu)-1);
(*51*)          v2 := Math.Sqrt(lam/x)*((x/nu)+1);
```

```
                v3 := Math.Exp(-2*lam/nu);
(*52*)          value := WBDevSpecfunc.Phi(v1) + v3* WBDevSpecfunc.Phi(v2);
(*53*)      END Cumulative;

(*54*)      PROCEDURE DrawSample (node: WBDevUnivariate.Node; censoring:
INTEGER; OUT sample: REAL);
(*55*)      VAR
(*56*)          nu, lam, left, right,y,y2,sqrt1,x1,u: REAL;
(*57*)      BEGIN
(*58*)          nu := node.arguments[location][0].Value();
(*59*)          lam := node.arguments[inverseScale][0].Value();
(*60*)          node.Bounds(left, right);
(*61*)          CASE censoring OF
(*62*)          |WBDevUnivariate.noCensoring:
                y := WBDevRandnum.Normal(0, 1);
                y2 := y*y;
                sqrt1 := Math.Sqrt(4*nu*y2*lam + nu*y2*nu*y2);
                x1 := nu + 0.5*nu/lam*nu*y2 - 0.5*nu/lam*sqrt1;
                u := WBDevRandnum.Uniform(0,1);
                IF (u > nu/(x1+nu)) THEN;
                    sample:= nu*nu/x1;
                ELSE;
                    sample:= x1;
(*63*)              END;
(*64*)          |WBDevUnivariate.leftCensored:
(*65*)              sample := WBDevRandnum.NormalLeftTruncated(nu,lam, left);
(*66*)          |WBDevUnivariate.rightCensored:
(*67*)              sample := WBDevRandnum.NormalTruncated(nu, lam, lam, right);
(*68*)          |WBDevUnivariate.intervalCensored:
(*69*)              sample := WBDevRandnum.NormalTruncated(nu, lam, left, right);
(*70*)          END;
(*71*)      END DrawSample;

        PROCEDURE (f: Factory) New (option: INTEGER): WBDevUnivariate.Node;
        VAR
            node: WBDevUnivariate.Node;
            stdNode: StdNode; left: Left; right: Right; interval: Interval;
        BEGIN
            CASE option OF
            |WBDevUnivariate.noCensoring:
                NEW(stdNode);
                node := stdNode;
            |WBDevUnivariate.leftCensored:
                NEW(left);
                node := left;
            |WBDevUnivariate.rightCensored:
                NEW(right);
                node := right;
```

```
            |WBDevUnivariate.intervalCensored:
                NEW(interval);
                node := interval;
            END;
            node.SetCumulative(Cumulative);
            node.SetDeclareArgTypes(DeclareArgTypes);
            node.SetDeclareProperties(DeclareProperties);
            node.SetDrawSample(DrawSample);
            node.SetLogFullLikelihood(LogFullLikelihood);
            node.SetLogPropLikelihood(LogPropLikelihood);
            node.SetLogPrior(LogPrior);
            node.SetNaturalBounds(NaturalBounds);
            node.Initialize;
            RETURN node;
        END New;

        PROCEDURE Install*;
        BEGIN
            WBDevUnivariate.Install(fact);
        END Install;

        PROCEDURE Init;
        VAR
            f: Factory;
        BEGIN
(*5*)           log2Pi := Math.Ln(2 * Math.Pi());
            NEW(f); fact := f;
        END Init;

    BEGIN
        Init;
(*1*)   END WBDevInversaGaussianaMF.
```

A.5 WBDev implementation of the Truncated IG-Poisson model

The truncated IG-Poisson model $p^{tigp}_{r:n}(b,c)$ is defined in (2.3)

$$p^{tigp}_{r:n}(b,c) = \frac{1}{(1+cn)^{-1/4}K_{-1/2}(b) - K_{-1/2}(b\sqrt{1+cn})} \frac{(\frac{1}{2}\frac{bcn}{\sqrt{1+cn}})^r}{r!} K_{r-1/2}(b\sqrt{1+cn})$$

,

for $r = 1, 2, \cdots, +\infty$, where $K_\alpha()$ is the modified Bessel function of the third kind of order α. This function is not available in the WBDev enviroiment, but this function has a recursive property :

$$K_{\nu+1}(z) = (2\nu/z)K_\nu(z) + K_{\nu-1}(z)$$

It makes that $p^{tigp}_{r:n}(b,c)$ can be calculated recursively:

$$p^{tigp}_{r:n}(b,c) = \left[\frac{cn}{1+cn}(1 - \frac{3}{2r})\right] p^{tigp}_{r-1:n}(b,c) + \left[\frac{(bcn)^2}{4r(r-1)(1+cn)}\right] p^{tigp}_{r-2:n}(b,c)$$

for $r = 3, 4, \cdots, +\infty$

where first two probabilities are:

$$p^{tigp}_{1:n}(b,c) = \frac{bcn}{2(1+cn)^{\frac{1}{2}}(e^{b((1+cn)^{\frac{1}{2}}-1)} - 1)}$$

$$p^{tigp}_{2:n}(b,c) = \frac{cn(1 + b(1+cn)^{\frac{1}{2}})}{4(1+cn)} p^{tigp}_{1:n}(b,c)$$

Once the value $p^{tigp}_{r:n}(b,c)$ is achieved, the log-likelihood $logp^{tigp}_{r:n}(b,c)$ is directly calculated.

The cumulative distribution function has no closed form and should be calculated by summing:

$$F^{tigp}_{r:n}(b,c) = \sum_{i=1}^{r} p^{tigp}_{i:n}(b,c) \tag{A.1}$$

As the definition of the cumulative distribution function is optional, we decided not to incorporate it. We must use this expression for generating random numbers distributed as Zero-Truncated IG-Poisson model:

1. Sample a random value u from an uniform $[0, 1]$; $U \sim uniform[0, 1]$
2. Get the minimum value of r that accomplish:

$$\sum_{i=1}^{r} p_{i:n}^{tigp}(b, c) > u$$

A.5.1 Source code for the odc module for the Truncated IG-Poisson model

```
(*1*)   MODULE WBDevSichelMF;

        IMPORT
            WBDevUnivariate,
(*2*)           WBDevRandnum, WBDevSpecfunc,WBDevBesselKMF,
(*3*)           Math;

        CONST
(*4*)           alpha = 0; theta = 1; tamany=2;

        TYPE
            StdNode = POINTER TO RECORD (WBDevUnivariate.StdNode) END;
            Left = POINTER TO RECORD (WBDevUnivariate.Left) END;
            Right = POINTER TO RECORD (WBDevUnivariate.Right) END;
            Interval = POINTER TO RECORD (WBDevUnivariate.Interval) END;
            Factory = POINTER TO RECORD (WBDevUnivariate.Factory) END;

        VAR
(*5*)           log2Pi: REAL;
            fact-: WBDevUnivariate.Factory;

(*6*)       PROCEDURE DeclareArgTypes (OUT args: ARRAY OF CHAR);
(*7*)       BEGIN
(*8*)           args := "sss";
(*9*)       END DeclareArgTypes;

(*10*)      PROCEDURE DeclareProperties (OUT isDiscrete, canIntegrate: BOOLEAN);
(*11*)      BEGIN
(*12*)          isDiscrete := TRUE;
```

```
(*13*)          canIntegrate := FALSE;
(*14*)      END DeclareProperties;

(*15*)      PROCEDURE NaturalBounds (node: WBDevUnivariate.Node; OUT lower, upper: REAL);
(*16*)      BEGIN
(*17*)          lower := 0;
(*18*)          upper := INF;
(*19*)      END NaturalBounds;

(*20*)      PROCEDURE LogFullLikelihood (node: WBDevUnivariate.Node; OUT value: REAL);
(*21*)      VAR
(*22*)          r, a, t,value1,value2: REAL;
            r_int,j: INTEGER;
(*23*)      BEGIN
(*24*)          r := node.value;
            r_int := SHORT(ENTIER(r));
(*25*)          a := node.arguments[alpha][0].Value();
(*26*)          t := node.arguments[theta][0].Value();
            value2:= Math.Ln((a*t/2)/(Math.Exp(a*(1-Math.Sqrt(1-t)))-1));
            value1:= value2+Math.Ln(t/4)+Math.Ln(1+a);
            IF r_int=1 THEN;
                value:=value2;
            ELSE
                IF r_int=2 THEN;
                    value:=value1;
                ELSE
                    j:=3;
                    REPEAT
                        value:= t*(1 -(3/(2*j))) * Math.Exp(value1);
                        value:= value + (Math.IntPower(a*t,2)
                        / ((4*j)*(j-1))) * Math.Exp(value2);
                        value:=Math.Ln(value);
                        j:=j+1;
                        value1:=value;
                        value2:=value1;
                    UNTIL j>r_int;
                END;
            END;
(*29*)      END LogFullLikelihood;

(*30*)      PROCEDURE LogPropLikelihood (node: WBDevUnivariate.Node; OUT value: REAL);
(*31*)      BEGIN
(*32*)          LogFullLikelihood(node, value);
(*33*)      END LogPropLikelihood;

(*34*)      PROCEDURE LogPrior (node: WBDevUnivariate.Node; OUT value: REAL);
(*37*)      BEGIN
(*38*)          LogFullLikelihood(node, value);
```

```
(*42*)      END LogPrior;

(*43*)      PROCEDURE Cumulative (node: WBDevUnivariate.Node; x: REAL; OUT value: REAL);
(*44*)      VAR
(*45*)          mu, tau, sqrtTau: REAL;
(*46*)      BEGIN
(*47*)                  HALT(126);
(*53*)      END Cumulative;

(*54*)      PROCEDURE DrawSample (node: WBDevUnivariate.Node; censoring: INTEGER; OUT sample: REAL)
(*55*)      VAR
(*56*)          a, t, left, right,r,u,N, prob1, prob2, prob, probacc: REAL;
(*57*)      BEGIN
(*25*)          a := node.arguments[alpha][0].Value();
(*26*)          t := node.arguments[theta][0].Value();
            N:=node.arguments[tamany][0].Value();
            node.Bounds(left, right);
            CASE censoring OF
(*62*)          |WBDevUnivariate.noCensoring:
                u := WBDevRandnum.Uniform(0,1);
                prob1:=0.5*a*t/(Math.Exp(a*(1-Math.Sqrt(1-t)))-1);
                prob2:=0.25*t*(1+a)*prob1;
                r:=1;
                probacc:=prob1;
                IF probacc>u THEN;
                    sample:=r;
                ELSE
                    r:=2;
                    probacc:=probacc+prob2;
                    IF probacc>u THEN;
                        sample:=r;
                    ELSE
                        r:=3;
                        REPEAT
                            prob:= t*((r -1.5)/r)*prob2
                            + (a*a*t*t)/(4*r*(r-1))*prob1;
                            probacc:=probacc + prob;
                            r:=r+1;
                            prob1:=prob2;
                            prob2:=prob;
                        UNTIL probacc>u;
                        sample:=r-1;
                    END;
                END;


(*64*)          |WBDevUnivariate.leftCensored:
(*65*)              sample := 1;
```

```
(*66*)          |WBDevUnivariate.rightCensored:
(*67*)              sample := 1;
(*68*)          |WBDevUnivariate.intervalCensored:
(*69*)              sample := 1;
(*70*)          END;
(*71*)      END DrawSample;

        PROCEDURE (f: Factory) New (option: INTEGER): WBDevUnivariate.Node;
        VAR
            node: WBDevUnivariate.Node;
            stdNode: StdNode; left: Left; right: Right; interval: Interval;
        BEGIN
            CASE option OF
            |WBDevUnivariate.noCensoring:
                NEW(stdNode);
                node := stdNode;
            |WBDevUnivariate.leftCensored:
                NEW(left);
                node := left;
            |WBDevUnivariate.rightCensored:
                NEW(right);
                node := right;
            |WBDevUnivariate.intervalCensored:
                NEW(interval);
                node := interval;
            END;
            node.SetCumulative(Cumulative);
            node.SetDeclareArgTypes(DeclareArgTypes);
            node.SetDeclareProperties(DeclareProperties);
            node.SetDrawSample(DrawSample);
            node.SetLogFullLikelihood(LogFullLikelihood);
            node.SetLogPropLikelihood(LogPropLikelihood);
            node.SetLogPrior(LogPrior);
            node.SetNaturalBounds(NaturalBounds);
            node.Initialize;
            RETURN node;
        END New;

        PROCEDURE Install*;
        BEGIN
            WBDevUnivariate.Install(fact);
        END Install;

        PROCEDURE Init;
        VAR
            f: Factory;
        BEGIN
(*5*)           log2Pi := Math.Ln(2 * Math.Pi());
```

```
            NEW(f); fact := f;
        END Init;

    BEGIN
        Init;
(*1*)   END WBDevSichelMF.
```

A.6 WBDev implementation of the Zero Truncated Poisson model

One may be under the impression that this distribution could be specified straightforwardly in Win- BUGS by applying the I(0,.) construct to dpois(.). Whilst in some circumstances this may lead to the same results, the I(.,.) construct was originally designed only to denote censored observations and shouldnt really be used in an attempt to model truncation in which the likelihood expression changes;

$$p_r^{tp}(\lambda) = \frac{\lambda^r}{r!}\frac{e^{-\lambda}}{(1-e^{-\lambda})}$$

for $r = 1, 2, \cdots, +\infty$

Then the log-likehood of the zero truncated Poisson model is;

$$logL^{tp}(\lambda|r) = -\lambda - \ln(1-e^{-\lambda}) + r\ln(\lambda) - \ln(r!)$$

for $r = 1, 2, \cdots, +\infty$

The cumulative distribution function should be expressed by a sum of the probabilities up to a given value r, It would be cumbersome and slow to calculate. As the definition of this function is optional, we decided not to incorporate the cumulative procedure into our module.

There is a function *WBDevRandnum. PoissonTruncated(.)*, to simulate from a zero truncated poisson, available in the WBDev environment. We directly use this function in DrawSample procedure to generate a random sample of the zero truncated Poisson.

A.6.1 Source code for the odc module for the Truncated Poisson model

```
(*1*)   MODULE WBDevTrPoissonMF;

        IMPORT
            WBDevUnivariate,
(*2*)           WBDevRandnum, WBDevSpecfunc,WBDevBesselKMF,
```

```
(*3*)           Math;

        CONST
(*4*)           lambda = 0;

        TYPE
            StdNode = POINTER TO RECORD (WBDevUnivariate.StdNode) END;
            Left = POINTER TO RECORD (WBDevUnivariate.Left) END;
            Right = POINTER TO RECORD (WBDevUnivariate.Right) END;
            Interval = POINTER TO RECORD (WBDevUnivariate.Interval) END;
            Factory = POINTER TO RECORD (WBDevUnivariate.Factory) END;

        VAR
(*5*)           log2Pi: REAL;
            fact-: WBDevUnivariate.Factory;

(*6*)       PROCEDURE DeclareArgTypes (OUT args: ARRAY OF CHAR);
(*7*)       BEGIN
(*8*)           args := "s";
(*9*)       END DeclareArgTypes;

(*10*)      PROCEDURE DeclareProperties (OUT isDiscrete, canIntegrate: BOOLEAN);
(*11*)      BEGIN
(*12*)          isDiscrete := TRUE;
(*13*)          canIntegrate := FALSE;
(*14*)      END DeclareProperties;

(*15*)      PROCEDURE NaturalBounds (node: WBDevUnivariate.Node; OUT lower, upper: REAL);
(*16*)      BEGIN
(*17*)          lower := 1;
(*18*)          upper := INF;
(*19*)      END NaturalBounds;

(*20*)      PROCEDURE LogFullLikelihood (node: WBDevUnivariate.Node; OUT value: REAL);
(*21*)      VAR
(*22*)        x,lam: REAL;
            x_int: INTEGER;
(*23*)      BEGIN
(*24*)          x := node.value;
                x_int := SHORT(ENTIER(x));
(*25*)          lam := node.arguments[lambda][0].Value();
                value:= -lam - Math.Ln(1-Math.Exp(-lam)) + x*Math.Ln(lam)
                - WBDevSpecfunc.LogFactorial(x_int);
(*26*)
(*29*)      END LogFullLikelihood;

(*30*)      PROCEDURE LogPropLikelihood (node: WBDevUnivariate.Node; OUT value: REAL);
(*31*)      BEGIN
```

```
(*32*)          LogFullLikelihood(node, value);
(*33*)      END LogPropLikelihood;

(*34*)      PROCEDURE LogPrior (node: WBDevUnivariate.Node; OUT value: REAL);
(*37*)      BEGIN
(*38*)          LogFullLikelihood(node, value);
(*42*)      END LogPrior;

(*43*)      PROCEDURE Cumulative (node: WBDevUnivariate.Node; x: REAL; OUT value: REAL);
(*44*)      VAR
(*45*)          mu, tau, sqrtTau: REAL;
(*46*)      BEGIN
(*47*)                  HALT(126);
(*53*)      END Cumulative;

(*54*)      PROCEDURE DrawSample (node: WBDevUnivariate.Node; censoring: INTEGER; OUT sample: REAL)
(*55*)      VAR
(*56*)          lam, left, right: REAL;
            right_int: INTEGER;
(*57*)      BEGIN
(*25*)          lam := node.arguments[lambda][0].Value();
(*26*)          node.Bounds(left, right);
            CASE censoring OF
(*62*)          |WBDevUnivariate.noCensoring:
                right_int:=SHORT(ENTIER(right));
                sample := WBDevRandnum.PoissonTruncated(lambda,1,right_int);
(*64*)          |WBDevUnivariate.leftCensored:
(*65*)              sample := 1;
(*66*)          |WBDevUnivariate.rightCensored:
(*67*)              sample := 1;
(*68*)          |WBDevUnivariate.intervalCensored:
(*69*)              sample := 1;
(*70*)          END;
(*71*)      END DrawSample;

        PROCEDURE (f: Factory) New (option: INTEGER): WBDevUnivariate.Node;
        VAR
            node: WBDevUnivariate.Node;
            stdNode: StdNode; left: Left; right: Right; interval: Interval;
        BEGIN
            CASE option OF
            |WBDevUnivariate.noCensoring:
                NEW(stdNode);
                node := stdNode;
            |WBDevUnivariate.leftCensored:
                NEW(left);
                node := left;
            |WBDevUnivariate.rightCensored:
```

```
                NEW(right);
                node := right;
            |WBDevUnivariate.intervalCensored:
                NEW(interval);
                node := interval;
            END;
            node.SetCumulative(Cumulative);
            node.SetDeclareArgTypes(DeclareArgTypes);
            node.SetDeclareProperties(DeclareProperties);
            node.SetDrawSample(DrawSample);
            node.SetLogFullLikelihood(LogFullLikelihood);
            node.SetLogPropLikelihood(LogPropLikelihood);
            node.SetLogPrior(LogPrior);
            node.SetNaturalBounds(NaturalBounds);
            node.Initialize;
            RETURN node;
        END New;

        PROCEDURE Install*;
        BEGIN
            WBDevUnivariate.Install(fact);
        END Install;

        PROCEDURE Init;
        VAR
            f: Factory;
        BEGIN
(*5*)           log2Pi := Math.Ln(2 * Math.Pi());
            NEW(f); fact := f;
        END Init;

    BEGIN
        Init;
(*1*)   END WBDevTrPoissonMF.
```

Appendix B

Data Sets

In this annex, most of the data sets used in this thesis are presented. The word frequency counts data sets in Chapter 2 where obtained from Baayen (2001), and here we summarize the information appearing there about them.

Word length and function words counts that are used in all of the other main chapters, were obtained from the original texts from an ebook edition. Some of these raw text files were obtained from the *Project Gutenberg*, *http://www.gutenberg.org/*, that is a website that facilitates the distribution of eBooks. None of the text used are protected by copyright law because their copyrights have expired.

The first step to obtain data, was to split the whole text file in individual text files for each unit of study; be it play act, chapter, sentence, or papers. To help to process this list of text files in a semi-automatized way, a basic tool was developed in Visual Basic for Windows. It performs the following tasks:

1. Remove punctuation, numbers, and other signs, to convert each text into a clean list of words.

2. Search, check and remove proper names, allowing one to do so interactively based on the list of words that appear capitalized (fully or partially) in the original text.

Thus the original text becomes a text file that includes only words that have passed the filter and they are prepared in a way such that they can be treated directly with a R script, to obtain the contingency tables of count data we need in a simple way:

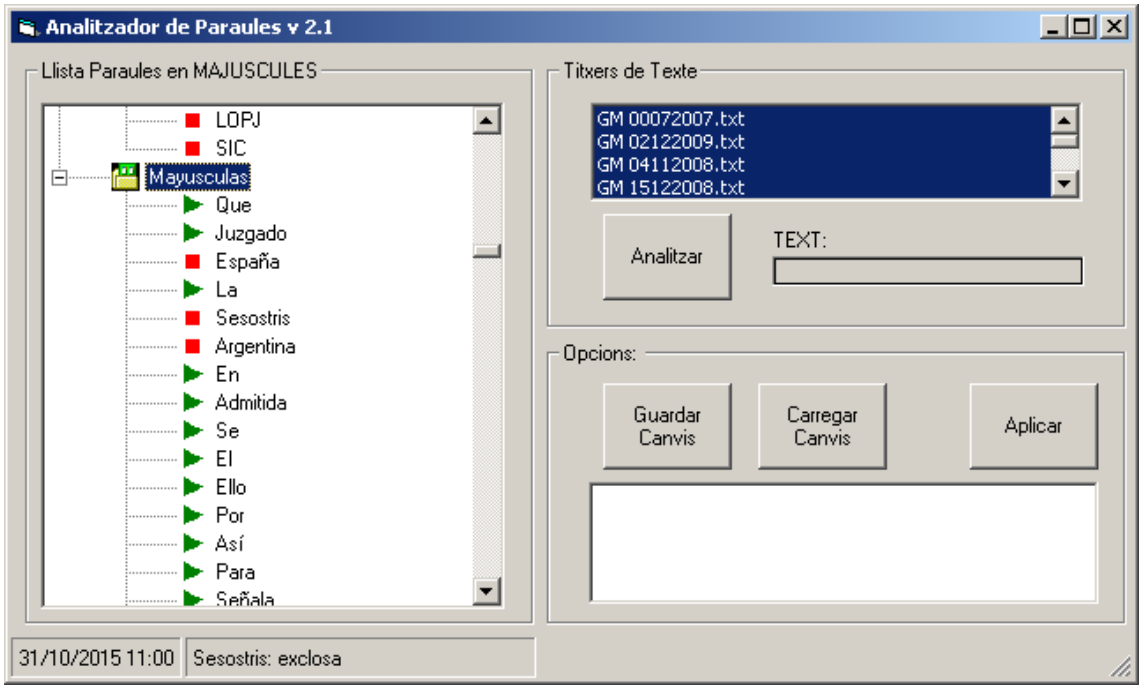


Figure B.1: Snapshot of the *Analitzador de Paraules v2.1* , tool developed to filter texts. In the left side, the tree of capitalized words shows a green arrow head (red square) when the word is included (excluded) from the text

* To obtain a row of word length frequencies from 1 to 9, plus a category 10 or more, from a text file (textfile.txt):

```
wordlist <- tolower(scan("textfile.txt",what='character'))
row<-table(cut(nchar(wordlist), breaks=c(0:9,100), labels=F))
```

* To obtain a row of function words frequencies, where fwords is a list of the function words, from a text file (textfile.txt):

```
wordlist <- tolower(scan("textfile.txt",what='character'))
row<-table(Sh01.1)[fwords]
```

  In cases where there is a 0 count on any of the function words, it will result in an error. To avoid that, the whole list of function words is added to the text and then the resulting counts are decreased in one unit.

B.1 Frequency of word frequency counts

B.1.1 Turkish text on archeology

Text in Turkish on archeology. Compared with English, Turkish is a language with a much richer morphologic system that allows one to create thousands of complex words from the same simple root.

r	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$v_{r:n}$	2326	477	178	107	53	33	22	26	7	7	12	8	4	3
r	15	16	17	18	20	21	22	23	24	28	32	34	36	38
$v_{r:n}$	2	7	4	2	1	4	1	1	2	2	2	1	1	1
r	43	44	51	56	68	69	193	222						
$v_{r:n}$	1	1	1	1	1	1	1	1						

Table B.1: Word frequency count data set for all the words in Turkish Archeology

B.1.2 Macaulay's Essay on Bacon

Data set based on the frequency count of the frequency of use names in an essay on Bacon from historian *Thomas Babington Macaulay*. This data set has been previously analysed by Yule, Good, Sichel. This author, in his books *History of England* (5 volumes, 1848-1861) and especially in his *Critical and Historical Essays* (1843), expressed high satisfaction of the English middle classes with the growing political power and prosperity they were enjoying. The sharpness and balance of Macaulay style, reflecting familiarity with the practice of parliamentary debate, contrasted with the sensitivity and beauty of the prose as contemporary authors *John Henry Newman*.

Text	Critical and Historical Essays (2 vol.)
Author	Thomas Babington Macaulay
Country	United Kigdom
Language	English
Literary genre	Essay
Editor	Alexander James Grieve
Publication date	1843

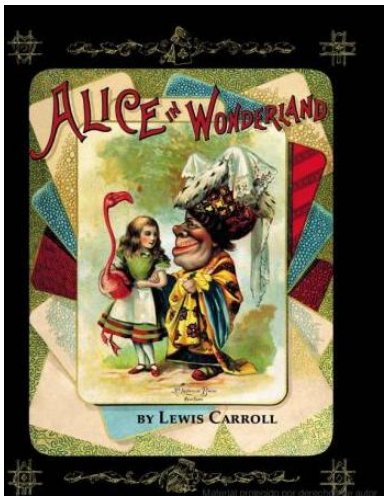
r	1	2	3	4	5	6	7	8	9	10	11	12	13
$v_{r:n}$	990	367	173	112	72	47	41	31	34	17	24	19	10
r	14	15	16	17	18	19	20	21	22	23	24	25	26
$v_{r:n}$	10	13	7	6	6	6	6	3	3	3	3	3	4
r	27	28	29	30	31	32	33	34	35	36	37	38	39
$v_{r:n}$	3	3	3	3	2	1	1	1	1	1	1	2	4
r	40	41	45	48	57	58	65	76	81	89	255		
$v_{r:n}$	1	1	2	1	1	1	1	1	1	1	1		

Table B.2: Word frequency count data set for all the words in Essay on Bacon

B.1.3 Alice's Adventures in Wonderland

Alice's Adventures in Wonderland (1865) is a work of fiction written by ECharles Lutwidge Dogson under the pseudonym of Lewis Carroll. It explains the story of a girl named Alice falling into a fantastic realm inhabited by peculiar anthropomorphic creatures. The story is full of references to Dogson friends (and their enemies), and the lessons that British schoolchildren were expected to memorize. It is considered one of the most characteristic books in the genre of the absurd.

The book is commonly referred to by short title *Alice in text Wonderland*. This alternate title was popularized by the numerous films and television adaptations of the story produced over time.



Author	Lewis Carroll
Illustrator	John Tenniel
Country	United Kingdom
Language	English
Literary Genre	Fiction Story
Editor	Macmillan
Publication date	1865
Aprox. N Pages.	224 pp
Continued by	Through the Looking-Glass

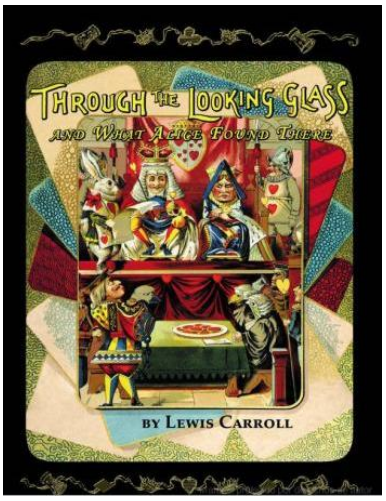
r	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$v_{r:n}$	1176	402	233	154	99	57	65	52	32	36	23	20	34	20
r	15	16	17	18	19	20	21	22	23	24	25	26	27	28
$v_{r:n}$	12	9	9	10	8	5	6	3	3	6	9	4	6	3
r	29	30	31	32	33	34	35	37	38	39	40	41	42	43
$v_{r:n}$	6	6	3	4	4	3	4	1	4	4	4	2	2	2
r	44	45	46	47	48	49	50	51	52	53	54	55	56	57
$v_{r:n}$	1	4	1	1	1	4	2	4	3	1	3	3	1	2
r	58	59	60	61	62	63	67	68	73	74	75	77	79	80
$v_{r:n}$	2	1	2	3	1	1	2	4	1	1	1	2	1	1
r	81	82	83	85	87	88	90	93	94	96	98	102	108	113
$v_{r:n}$	1	2	2	1	1	2	1	1	1	2	1	2	1	1
r	114	121	128	131	133	136	144	145	148	151	153	170	177	179
$v_{r:n}$	1	1	1	1	1	1	1	1	1	1	1	1	1	1
r	182	194	211	247	263	280	356	364	365	386	410	460	510	528
$v_{r:n}$	1	1	1	1	1	1	1	1	1	1	1	1	1	1
r	540	629	726	866	1631									
$v_{r:n}$	1	1	1	1	1									

Table B.3: Word frequency count data set for all the words in Alice's adventures in wonderland

B.1.4 Through the Looking-Glass

Through the Looking-Glass (1871) is a children's literature work written by Lewis Carroll (pseudonym of Charles Lutwidge Dodgson), generally classified within the genre of the absurd. It is the sequel to *Alice's Adventures in text Wonderland* (1865).

Although it refers to the events described in the first book, the theme and setting of *Through the Looking-Glass* makes it a sort of mirror image of *Wonderland*. The first book begins outdoors in temperate month of May in the the anniversary of *Alice* (May 4), frequently changes size as story develops, and it draws an imaginary world based on playing cards. The second book begins inside on a snowy winter night exactly six months later, on November 4, frequently changes time and spatial directions as a story develops, and it draws an imaginary world from Chess.



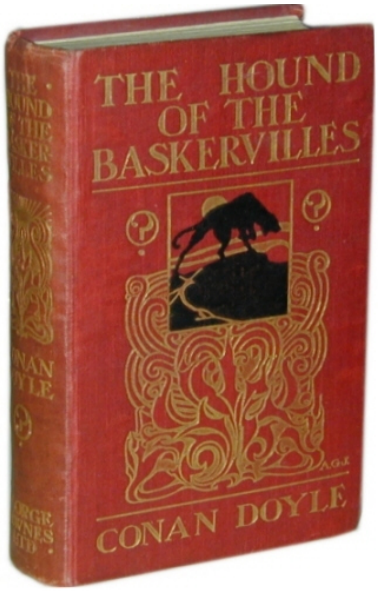
Author	Lewis Carroll
Illustrator	John Tenniel
Country	Unite Kingdom
Language	English
Literary Genre	Fiction Story
Editor	Macmillan
Publication Date	1871
Aprox. N Pages	224 pp
Preceded by	Alice's Adventures in Wonderland

r	1	2	3	4	5	6	7	8	9	10	11	12	13
$v_{r:n}$	1491	460	259	148	113	78	61	47	28	26	26	30	22
r	14	15	16	17	18	19	20	21	22	23	24	25	26
$v_{r:n}$	19	12	21	12	11	16	9	7	9	2	3	1	5
r	27	28	29	30	31	32	33	34	35	36	37	38	39
$v_{r:n}$	3	7	5	2	5	3	2	5	5	2	5	3	2
r	40	41	42	45	46	48	49	50	51	52	53	54	55
$v_{r:n}$	1	2	2	1	3	4	2	2	3	4	2	4	2
r	56	57	58	59	60	61	62	63	64	65	66	67	69
$v_{r:n}$	1	2	1	1	1	2	2	3	3	1	1	3	1
r	70	72	73	74	75	78	79	80	84	86	87	89	90
$v_{r:n}$	4	1	1	2	1	1	2	1	2	2	1	1	1
r	93	94	101	104	112	113	115	116	119	121	123	132	135
$v_{r:n}$	1	1	1	1	1	1	1	1	1	2	2	1	1
r	139	140	145	147	150	151	177	180	193	195	209	211	229
$v_{r:n}$	1	1	1	1	1	1	2	1	1	1	1	1	1
r	247	268	300	309	354	399	425	470	502	505	517	545	705
$v_{r:n}$	1	1	1	1	1	1	1	2	1	1	1	1	1
r	739	836	1555										
$v_{r:n}$	1	1	1										

Table B.4: Word frequency count data set for all the words in Through the looking-glass and what Alice found there

B.1.5 The Hound of the Baskervilles

Hound of the Baskervilles is a novel halfway between mystery and terror written by Sir Arthur Conan Doyle, originally published as a series in the Strand Magazine from August 1901 to April 1902, and located mainly in the region of Dartmoor. It is a relevant fact that Conan Doyle was Plymouth doctor at the time of writing the book. In the novel, the detective Sherlock Holmes and his assistant Dr. Watson are called to investigate an alleged curse that falls on the Baskervilles house that could explain the death of its last owner.



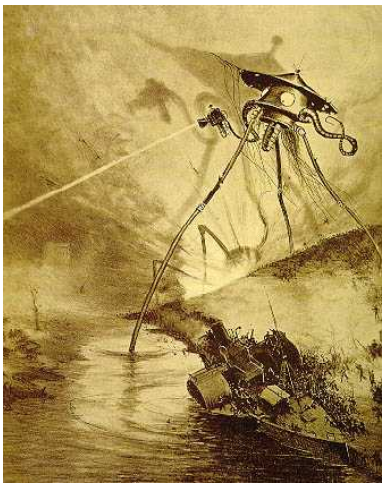
Author	Arthur Conan Doyle
Country	United Kingdom
Language	English
Series	Sherlock Holmes
Literary Genre	Thriller
Editor	George Newnes
Publication Date	1901 to 1902
Aprox. N Pages	243 pp

r	1	2	3	4	5	6	7	8	9	10	11	12	13
$v_{r:n}$	2836	889	449	280	208	137	116	92	86	52	48	40	33
r	14	15	16	17	18	19	20	21	22	23	24	25	26
$v_{r:n}$	25	34	22	20	15	13	13	17	14	9	12	7	16
r	27	28	29	30	31	32	33	34	35	36	37	38	39
$v_{r:n}$	5	8	7	7	7	4	6	8	2	7	5	3	5
r	40	41	42	43	44	45	46	47	48	49	50	52	54
$v_{r:n}$	4	3	3	8	3	3	4	1	1	2	1	1	4
r	55	57	58	60	61	62	63	64	65	66	67	68	69
$v_{r:n}$	5	3	1	2	1	5	3	2	3	2	3	2	2
r	70	71	72	73	74	77	80	82	84	85	86	87	88
$v_{r:n}$	1	1	1	1	2	2	1	1	2	1	2	1	3
r	89	90	92	94	97	99	102	104	105	107	110	111	112
$v_{r:n}$	1	2	1	1	2	2	1	1	1	1	2	1	1
r	113	114	118	123	128	137	140	141	143	146	149	151	155
$v_{r:n}$	3	1	2	1	1	2	1	1	1	1	1	1	1
r	165	167	171	175	178	182	185	190	192	199	200	201	202
$v_{r:n}$	1	1	1	1	2	1	1	1	1	1	1	2	1
r	205	207	209	211	215	222	233	240	242	244	264	286	298
$v_{r:n}$	1	1	1	1	1	1	1	1	1	1	1	1	1
r	314	315	326	329	337	350	364	374	400	405	416	419	441
$v_{r:n}$	1	1	1	1	1	1	1	1	1	1	1	1	1
r	453	479	506	541	624	689	803	827	911	914	980	1132	1305
$v_{r:n}$	1	1	1	1	1	1	1	1	1	1	1	1	1
r	1407	1465	1592	1627	3327								
$v_{r:n}$	1	1	1	1	1								

Table B.5: Word frequency count data set for all the words in HOUND OF THE BASKERVILLES

B.1.6 War of the Worlds

The War of the Worlds (1898), written by H.G. Wells is an early science fiction novel which describes an invasion of England by aliens from Mars. It is one of the first and best known descriptions of an alien invasion of Earth, and has had influence on many others. It has generated many films and TV series based on it.



Author	Herbert George Wells
Country	United Kingdom
Language	English
Literary Genre	Science fiction novel
Editor	William Heinemann
Publication Date	1898
Aprox. N Pages	303 pp

r	1	2	3	4	5	6	7	8	9	10	11	12	13
$v_{r:n}$	3613	1138	567	340	250	177	135	93	72	67	44	46	44
r	14	15	16	17	18	19	20	21	22	23	24	25	26
$v_{r:n}$	42	38	31	24	26	16	18	14	13	11	7	8	10
r	27	28	29	30	31	32	33	34	35	37	38	39	40
$v_{r:n}$	8	6	9	9	4	8	2	6	9	6	6	7	4
r	41	42	43	44	45	46	47	48	49	50	52	53	55
$v_{r:n}$	6	3	6	3	4	2	3	6	6	5	1	4	3
r	57	58	59	60	61	63	65	66	67	68	69	70	71
$v_{r:n}$	4	2	2	2	3	3	2	4	3	3	2	1	4
r	72	73	74	75	76	78	79	82	85	87	88	90	91
$v_{r:n}$	1	1	1	2	1	2	1	1	3	1	1	1	1
r	94	96	99	100	101	102	103	108	112	114	116	117	120
$v_{r:n}$	1	1	3	1	5	2	1	1	1	1	2	1	2
r	124	128	129	140	142	146	150	154	158	164	166	167	171
$v_{r:n}$	1	1	2	1	1	1	1	3	1	1	2	1	2
r	174	177	181	184	185	191	198	199	207	213	218	231	243
$v_{r:n}$	1	1	1	1	1	1	1	1	1	1	1	1	1
r	247	248	250	254	266	292	320	327	343	378	379	420	441
$v_{r:n}$	1	1	1	1	1	1	1	1	1	1	1	1	1
r	446	447	469	579	647	766	850	991	1172	1257	1605	2297	2487
$v_{r:n}$	1	1	1	1	1	1	1	1	1	1	1	1	1
r	4775												
$v_{r:n}$	1												

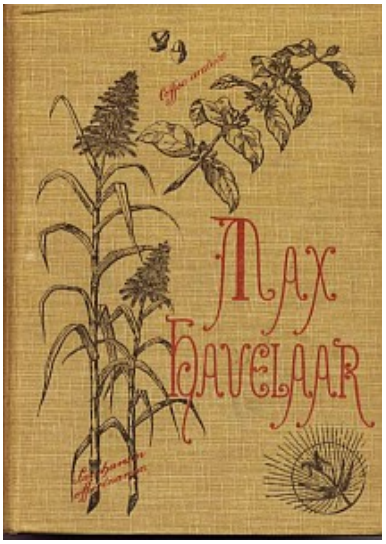
Table B.6: Word frequency count data set for all the words in War of the Worlds

B.1.7 Max Havelaar

Max Havelaar: Or the Coffee Auctions of the Dutch Trading Company (1860) written by Multatuli (pseudonym of Eduard Douwes Dekker) is a novel that played a key role in shaping and modifying policy the colonial Dutch East Indies in the nineteenth century and the beginning twentieth century In the novel, Max Havelaar tries to fight a corrupt system of government of the island of Java, which was a Dutch colony at the time.

Despite its laconic and concise writing style, it raised the consciousness of Europeans living in Europe that the wealth they enjoyed was the result of suffering in other parts of the world. This awareness eventually led to the new political ethics through which the Dutch colonial government tried to repay its debt to the colonies by providing education to its inhabitants.

Max Havelaar was partly responsible for the end of Dutch colonialism in Indonesia in 1945, which helped to later decolonize Africa and other parts of the world.



Original Tittle	Max Havelaar, of de koffie-veilingen der Nederlandse Handel-Maatschappij
Author	Eduard Douwes Dekker
Country	Netherlands
Language	Dutch
Literary Genre	Fiction Novel
Publication Date	1860

r	1	2	3	4	5	6	7	8	9	10	11	12	13
$v_{r:n}$	6004	1731	819	491	368	258	168	137	123	108	80	52	60
r	14	15	16	17	18	19	20	21	22	23	24	25	26
$v_{r:n}$	57	39	34	37	33	19	33	21	19	20	18	14	13
r	27	28	29	30	31	32	33	34	35	36	37	38	39
$v_{r:n}$	9	9	13	13	9	9	10	9	6	5	10	7	9
r	40	41	42	43	44	45	46	47	48	49	50	51	52
$v_{r:n}$	6	8	8	8	5	3	6	4	4	2	6	3	4
r	53	54	55	56	57	58	59	61	62	63	64	65	66
$v_{r:n}$	1	3	5	4	3	4	8	8	2	2	4	2	5
r	67	68	69	70	71	72	73	74	75	76	78	79	80
$v_{r:n}$	5	2	3	3	1	2	1	1	3	3	4	2	2
r	81	82	83	86	87	88	90	92	93	96	98	101	102
$v_{r:n}$	1	2	4	1	2	2	1	1	2	3	2	3	1
r	105	106	107	109	110	111	113	114	115	116	120	121	122
$v_{r:n}$	1	1	1	3	1	1	2	1	1	1	1	1	1
r	123	125	126	127	128	135	139	145	147	151	154	156	161
$v_{r:n}$	1	1	3	1	2	1	1	2	3	1	1	1	1
r	162	165	169	170	171	177	184	188	190	194	198	202	208
$v_{r:n}$	1	1	1	1	1	1	1	1	1	1	1	3	1
r	222	223	228	234	235	236	238	242	244	262	272	283	285
$v_{r:n}$	1	1	1	1	1	1	1	2	2	1	1	1	1
r	286	289	300	308	317	323	344	358	360	365	369	384	391
$v_{r:n}$	1	1	1	1	1	1	1	1	1	1	1	1	1
r	416	430	437	443	452	453	477	479	494	541	631	650	653
$v_{r:n}$	1	1	2	1	1	1	1	1	1	1	1	1	1
r	710	714	736	920	957	990	1159	1168	1267	1335	1423	1644	1686
$v_{r:n}$	1	1	1	1	1	1	1	1	1	1	1	1	1
r	1834	1894	1955	2032	2306	2782	4826						
$v_{r:n}$	1	1	1	1	1	1	1						

Table B.7: Word frequency count data set for all the words in Max Havelaar

B.2 Word length and frequent function words counts

B.2.1 Tirant lo Blanc

Tirant lo Blanc is a chivalry book written in catalan, hailed to be "the best book of its kind in the world" by Cervantes in El Quixote, and considered by many to be the first modern novel in Europe, (see, e.g., Vargas Llosa, 1991, 93). The main body of the book was written between 1460 and 1464, but it was not printed until 1490, and there has been a long lasting debate around its authorship, originating from conflicting information in its first edition.

Where in the dedicatory letter at the beginning of the book it is stated that "*So that no one else can be blamed if any faults are found in this work, I, Joanot Martorell, take sole responsibility for it, as I have carried out the task singlehandedly*," in the colophon at the end of the book it is stated that "*Because of his death, Sir Joanot Martorell could only finish writing three parts of it. The fourth part, which is the end of the book, was written by the illustrious knight Sir Martí Joan de Galba. If faults are found in that part, let them be attributed to his ignorance.*"

Author	Joanot Martorell Martí Joan de Galba (?)
Country	Kingdom of Valencia
Language	Catalan
Literary Genre	Chivalric Romance
Publication Date	1490

Chap.	1	2	3	4	5	6	7	8	9	10+
1	21	59	44	19	33	20	16	17	9	17
2	53	113	80	49	52	33	28	36	16	16
3	109	274	239	128	112	110	76	51	43	32
4	69	150	126	71	60	71	47	32	23	21
5	119	207	231	123	128	102	61	55	29	34
6	69	136	126	69	60	61	37	27	15	15
7	32	63	51	18	29	28	15	15	19	13
8	26	52	41	19	27	29	11	16	5	11
9	23	42	48	16	15	28	12	15	14	10
10	92	191	190	93	84	72	47	47	27	24
11	60	132	144	62	55	50	26	27	9	14
12	43	66	94	66	44	29	27	13	9	13
14	125	217	245	92	101	131	59	54	28	33
15	41	118	129	38	40	44	27	25	16	7
16	27	64	59	42	16	28	16	17	14	6
17	68	151	135	73	57	82	33	46	20	15
18	57	134	119	48	65	74	25	25	10	9
19	130	302	328	150	128	132	75	59	26	37
20	53	117	98	70	51	60	28	26	14	22
21	56	142	136	67	45	64	25	28	14	13
22	69	179	137	85	58	66	32	47	20	26
23	85	163	217	99	108	79	53	27	37	24
24	66	139	171	77	78	59	35	33	13	11
25	116	174	205	128	134	98	68	29	22	13
26	158	316	274	136	157	150	80	61	37	32
27	222	442	452	188	221	218	94	89	49	36
28	52	72	78	48	43	54	18	15	21	20
29	47	128	111	49	50	47	26	19	13	17
30	33	54	48	30	16	26	6	16	5	10
31	38	67	60	27	31	45	13	10	4	9
32	84	120	152	77	85	66	42	42	18	18
33	76	133	162	75	80	66	31	50	17	19
34	72	143	134	76	56	43	23	35	16	27
35	73	184	174	79	96	79	37	40	23	23
36	43	92	111	47	54	34	28	31	11	17
37	38	53	55	22	35	37	15	15	18	16
38	79	154	150	66	76	77	33	35	9	13
39	132	212	211	116	82	134	51	36	25	33
40	24	51	55	28	22	33	14	10	9	6
41	83	194	208	100	112	100	42	27	22	34
42	75	115	106	75	91	94	57	22	14	14
43	28	62	72	36	25	23	17	11	9	13
44	90	160	125	86	78	82	51	43	24	16
52	111	128	139	73	107	62	34	39	30	22
53	41	91	83	52	48	44	13	17	9	8
54	27	65	59	22	29	37	12	24	12	10
55	115	220	195	135	128	105	49	59	23	32
56	109	207	220	95	104	96	34	44	28	23
57	129	304	282	144	146	121	70	57	28	32
58	41	61	74	27	42	31	15	13	8	14
59	181	297	282	157	162	140	66	76	48	39
60	127	272	238	132	126	121	71	45	24	25
61	23	45	60	20	29	27	12	7	6	4
63	44	77	79	25	39	46	11	9	9	4
64	115	235	221	80	117	90	46	33	34	26
65	164	320	335	121	181	144	67	54	45	31
66	27	75	92	25	37	34	8	12	10	7
67	273	532	523	273	301	240	115	86	66	45
68	467	798	781	420	452	380	191	156	98	86
71b	291	490	516	268	282	258	113	103	103	63
72	128	192	191	111	111	120	34	37	14	20
73	151	226	224	103	138	134	72	58	26	13
74	152	302	295	128	134	137	84	76	38	33
75	34	87	96	33	38	36	16	17	3	6
76	87	196	162	94	66	81	36	45	20	18
78	79	143	151	80	94	68	22	31	22	19
80	134	235	263	131	141	114	47	50	27	36
81	329	652	561	256	313	270	143	146	72	53
82	84	130	127	74	59	66	30	38	12	8
83	29	63	57	26	23	32	15	17	6	10
84	133	222	179	105	133	124	59	64	23	28
85	182	354	310	122	152	188	80	75	38	47
92	160	210	161	108	115	91	46	37	20	29
96	40	71	72	26	36	31	16	17	8	10
97	79	105	102	60	49	76	33	24	18	11
98	246	633	529	321	241	262	157	115	73	52
99	161	309	317	176	160	170	91	78	62	41
100	383	768	779	354	387	374	169	168	73	76
101	240	507	418	198	208	228	110	73	46	36
102	29	65	59	30	28	29	18	10	3	7
103	39	116	108	50	48	42	21	11	13	8
104	204	466	415	197	190	217	118	73	77	41
105	176	324	330	149	175	209	95	56	36	35
106	240	438	438	227	260	205	112	65	32	21
107a	92	172	144	91	77	96	48	31	37	23
107b	58	111	96	43	56	64	33	36	31	15
108	145	284	309	138	142	173	88	49	37	23
109	272	559	533	245	297	262	156	85	72	62
110	374	873	775	338	386	371	221	136	45	38
111	108	226	208	89	103	119	58	33	30	22
112	45	70	74	32	26	47	19	14	8	5
113	156	306	302	138	170	150	74	57	36	30
114	202	440	376	172	210	207	96	80	41	34
116	154	264	261	122	124	124	70	64	49	35
117	141	311	251	139	120	157	77	49	43	30
118	73	147	123	76	83	65	33	29	18	12
119	422	813	766	373	417	440	237	196	130	86
120	25	59	55	31	19	24	9	9	4	3
121	129	238	255	121	144	138	68	66	39	28
122	23	55	50	28	27	38	15	19	13	18
123	78	157	146	66	78	82	29	37	20	17
124	213	384	390	195	175	198	108	97	77	32
125	257	461	440	198	216	246	106	112	63	52
126	136	258	272	122	111	138	54	49	35	27
127	231	477	380	209	216	212	117	119	51	32
128	68	102	105	53	64	40	30	21	12	15
129	184	397	395	160	182	197	82	106	50	51
130	46	112	95	48	44	58	30	23	16	15
131	146	211	273	142	127	134	59	44	41	31
132	215	506	392	184	235	195	112	104	57	29
133	465	872	970	587	517	471	242	189	130	93
134	238	382	387	204	182	219	114	83	69	46
136	26	59	45	28	18	30	13	10	12	5
137	159	297	352	152	173	139	84	64	46	33
138	261	563	482	271	252	297	133	142	94	44
139	49	65	75	62	51	59	22	23	9	6
140	153	259	293	178	185	153	64	66	37	34
141	422	792	808	404	407	393	208	148	140	92
142	73	164	183	111	81	101	35	31	33	9
143	475	897	789	394	466	427	302	219	131	106
144	54	84	98	59	52	42	19	24	22	19
145	245	444	449	214	229	214	104	91	96	57
146	366	800	705	341	352	318	152	175	136	80
148	146	257	290	128	129	157	70	57	39	21
149	199	428	335	209	165	202	90	63	44	18
151	78	224	163	76	83	83	44	46	22	10
153	150	309	329	141	146	163	84	55	39	30
154	398	1033	932	414	414	466	228	170	165	75
155	236	461	450	255	178	247	110	74	75	43
156	66	131	126	57	73	75	46	37	32	12
157	406	782	753	383	403	455	232	153	109	66
159	199	499	459	205	218	263	110	84	67	27
161	502	1125	971	453	477	475	239	213	153	69
162	178	358	307	186	161	155	96	64	53	39
163	329	746	667	344	345	372	156	119	99	69
164	319	634	665	380	387	351	180	116	60	59
165	22	44	41	21	25	38	16	25	15	7
166	179	380	342	189	167	165	68	81	50	32
167	125	220	200	99	108	118	51	54	34	31
169	59	124	108	49	47	63	23	32	16	15
170	23	61	61	32	29	24	7	11	9	8
171	37	92	71	41	32	29	22	15	7	10
172	115	271	299	124	116	129	63	51	43	19
173	109	196	204	90	107	110	41	40	23	36
174	29	86	52	22	39	39	17	14	5	5
175	39	74	79	40	32	29	9	10	9	6
176	55	115	103	54	45	53	24	17	13	15
177	58	121	114	52	36	60	25	25	16	21
178	109	233	190	115	94	104	42	41	29	21
179	57	156	121	59	63	36	33	27	13	14
180	61	148	139	68	68	52	34	39	14	14
181	68	80	148	58	61	53	41	41	18	12
182	81	204	205	86	75	83	48	66	23	21
183	76	170	182	64	61	90	41	39	20	22
184	30	100	105	30	29	26	21	18	12	9
185	47	69	50	21	26	32	31	16	25	9
189	754	1462	1246	642	739	717	315	307	200	139
190	36	71	63	35	47	34	35	14	20	11
191	39	67	53	37	29	28	18	11	15	4
192	37	99	111	63	47	53	18	27	19	12
194	29	61	50	31	26	23	19	21	9	6
195	43	89	93	33	44	49	17	7	16	18
201	69	129	103	60	71	52	42	27	18	11
202	62	138	119	61	77	54	28	21	19	15
207	30	55	68	26	31	25	13	11	9	7
208	54	170	143	64	72	63	30	34	18	22
209	53	162	134	61	65	67	37	34	12	13
210	78	239	191	90	102	106	59	56	31	31
211	39	111	82	37	42	42	23	20	26	10
212	29	67	59	34	30	19	20	18	12	10
214	78	198	178	63	72	85	34	34	20	19
215	92	216	220	80	95	102	39	33	29	20
216	80	211	208	110	99	75	38	50	23	13
217	24	107	82	38	40	44	26	24	6	13
218	172	370	296	178	144	149	84	73	60	41
219	30	68	66	29	30	33	18	18	19	15
220	139	249	217	93	98	97	54	28	40	38
221	110	238	187	107	109	112	69	66	50	47
222	229	346	371	196	223	158	80	80	80	46
223	28	71	73	24	34	44	16	13	9	6
224	97	205	181	88	74	111	40	47	30	18
225	72	223	194	95	78	85	31	31	10	19
226	86	239	229	105	94	106	43	59	15	20
227	71	176	168	81	90	86	30	46	21	13
228	123	342	310	125	122	132	71	71	32	17
229	147	307	264	135	145	154	62	67	30	24
230	66	208	143	49	77	55	23	33	16	15
231	125	351	293	135	124	158	45	65	35	31
232	48	84	100	36	39	53	21	18	12	6
233	248	495	429	195	247	201	103	86	38	37
234	192	447	393	175	205	181	109	86	39	36
235	19	57	54	20	22	25	11	15	4	9
236	209	493	446	206	209	216	83	83	54	34
237	30	88	86	42	39	43	22	16	15	18
238	109	225	197	105	78	112	39	41	31	15
239	144	305	301	149	128	167	66	47	46	30
240	33	132	122	31	57	50	23	30	8	12
241	42	110	70	33	51	46	25	17	9	7
242	24	106	76	38	38	32	14	24	5	10
244	47	136	129	57	57	54	39	21	15	16
245	16	60	39	14	16	25	17	15	1	8
248	43	103	70	42	46	42	23	16	12	15
249	54	146	120	50	63	56	29	21	12	20
250	35	116	57	32	42	57	21	25	14	12
251	56	148	145	59	63	54	46	29	11	15
252	33	101	91	44	32	68	23	26	9	8
253	41	87	81	50	39	58	14	15	7	7
254	100	157	159	82	78	75	42	53	20	20
255	33	109	101	29	43	49	28	18	12	9
256	48	80	79	44	37	26	23	20	11	6
257	38	88	63	47	50	51	15	17	8	14
258	129	263	263	113	142	123	53	45	35	36
259	24	68	54	42	23	25	16	10	5	10
260	139	303	238	130	131	121	76	59	29	36
262	352	844	701	328	361	353	205	128	79	75
263	141	286	256	141	129	126	61	42	15	29
264	211	418	375	196	176	196	92	70	38	53
265	79	153	132	68	62	67	51	21	15	14
266	41	133	119	42	44	45	27	17	27	12
267	21	67	63	23	34	25	18	19	12	3
268	93	270	226	97	118	108	69	35	24	24
269	179	413	369	179	185	168	82	83	41	40
270	18	67	55	17	26	20	23	21	3	6
271	58	131	124	65	56	35	24	43	13	10
272	37	69	57	23	39	28	20	24	11	8
273	30	81	82	35	42	30	23	16	12	12
274	20	78	58	29	30	39	19	22	4	6
275	74	118	117	58	54	61	39	31	21	14
276	33	107	81	62	47	45	21	25	14	12
277	105	202	164	108	91	91	53	48	23	20
278	57	105	76	38	53	62	36	32	11	17
279	65	128	106	47	52	43	22	17	12	17
280	42	110	86	39	41	39	22	17	16	12
281	165	305	272	161	141	157	62	73	34	41
282	117	252	236	102	122	122	56	62	26	25
283	187	387	351	169	197	165	84	67	26	55
284	25	62	47	29	35	35	23	20	8	18
285	31	99	66	39	35	26	20	13	4	18
286	141	264	284	148	162	122	72	51	38	26
288	177	357	339	168	168	164	90	70	44	36
289	23	43	45	24	25	16	11	9	16	3
290	83	222	183	112	87	82	43	27	35	23
291	220	472	457	220	198	214	100	94	76	43
292	205	399	384	200	175	214	84	81	50	38
293	91	153	143	77	67	84	38	40	26	20
294	32	99	72	43	44	37	32	30	13	9
295	74	242	201	112	122	77	75	45	41	35
296	82	169	153	76	78	65	51	42	13	24
297	34	109	107	41	44	47	25	17	11	15
299	232	583	520	300	288	233	136	93	45	56
300	24	89	85	43	31	32	22	12	6	7
301	173	374	367	180	165	159	80	91	44	30
302	48	95	74	35	50	32	35	16	12	12
303	34	114	112	36	39	32	40	16	9	7
304	207	361	373	205	207	224	94	76	43	36
305	17	57	56	29	22	14	16	10	6	11
306	29	51	46	9	27	37	22	18	11	5
307	133	293	282	140	127	133	78	43	38	22
308	29	119	93	40	45	28	19	19	14	9
309	235	499	495	235	230	229	131	106	63	55
310	265	616	569	294	286	231	158	111	59	51
311	28	84	67	40	26	29	20	10	10	7
312	105	233	203	90	85	126	38	52	30	13
313	59	102	107	60	39	51	24	15	9	12
314	32	74	50	26	31	28	16	12	8	7
315	254	419	422	240	226	204	112	65	48	51
316	24	91	90	39	43	24	18	23	11	9
317	82	227	211	98	90	79	42	51	24	28
318	42	86	91	45	28	38	19	23	10	8
319	145	365	380	160	143	160	88	68	52	27
320	22	49	52	21	16	22	10	14	3	8
321	146	315	340	135	159	144	81	60	41	33
322	40	107	104	41	45	41	30	18	10	12
323	46	136	133	47	60	55	35	31	11	15
324	57	196	135	88	60	59	39	34	17	20
325	48	146	106	57	51	58	37	25	23	19
326	24	67	59	29	34	19	12	23	14	8
327	51	156	138	59	56	51	31	30	26	17
328	94	128	188	66	82	76	53	44	25	19
329	67	139	130	53	66	56	36	35	20	11
330	121	228	232	110	110	106	72	56	37	20
331	46	118	87	61	67	38	23	19	9	10
332	17	72	68	16	34	37	13	11	8	3
333	61	153	155	78	79	66	30	32	26	13
334	182	320	335	187	180	154	98	76	54	33
335	77	135	158	90	82	64	35	27	24	9
336	12	63	51	17	21	21	5	12	3	11
337	111	220	228	106	117	119	46	28	30	28
338	24	54	59	22	24	28	22	18	8	10
339	100	207	225	99	94	111	71	32	15	16
340	241	353	476	235	237	225	113	85	79	39
341	22	61	61	25	17	29	10	14	10	3
343	147	241	308	142	157	140	81	58	35	29
344	109	203	185	108	107	94	61	43	28	16
345	84	163	186	86	72	80	38	29	28	19
346	48	107	107	51	33	41	29	34	9	20
347	65	145	167	76	74	62	41	30	26	24
348	26	57	36	22	28	18	20	15	8	5
349	234	446	446	246	253	254	126	104	84	76
350	121	232	240	104	122	120	54	49	38	37

Table B.8: Word length counts for all the words in Tirant lo Blanc (1/2)

Chap.	1	2	3	4	5	6	7	8	9	10+
351	42	96	96	30	49	33	27	33	10	12
352	23	43	40	22	27	24	17	8	6	7
353	71	156	156	86	60	58	42	55	31	23
354	77	198	188	92	73	71	39	46	44	24
355	145	361	313	154	140	163	102	81	47	50
356	29	106	98	35	41	40	26	31	12	10
357	87	231	217	102	96	105	50	42	31	41
358	21	67	58	38	19	26	18	16	9	7
359	38	67	69	27	32	27	21	19	5	6
360	37	91	89	45	34	45	26	24	9	16
362	44	87	100	43	35	58	24	23	15	17
363	21	53	46	40	31	20	11	6	10	6
364	24	84	80	34	38	30	11	26	10	12
365	18	69	54	27	29	22	12	17	5	8
366	68	134	105	69	52	66	31	20	14	20
367	68	134	105	69	52	66	31	20	14	20
368	37	82	71	42	23	27	19	11	6	13
369	35	119	88	42	44	51	36	25	8	15
370	53	116	92	52	48	62	31	22	16	13
371	71	197	156	78	80	59	26	45	17	21
372	91	184	172	86	93	96	49	58	27	36
373	42	79	80	40	46	38	19	18	14	12
374	147	296	241	120	138	130	142	78	73	76
375	42	77	53	37	29	31	24	26	16	22
376	86	140	108	74	64	77	57	48	52	46
377	34	74	45	21	33	23	27	13	16	17
378	46	118	92	71	48	34	45	30	18	21
379	12	51	41	15	34	26	15	18	3	15
380	31	97	66	44	51	38	19	27	20	21
381	27	66	43	24	21	30	14	22	12	13
382	41	72	61	30	32	44	12	20	8	10
383	42	60	52	30	30	30	16	17	16	16
384	57	101	110	58	43	50	32	24	19	18
385	30	53	58	37	16	27	21	16	13	16
386	64	117	113	55	53	79	39	34	20	14
387	163	321	350	170	180	165	103	80	63	44
388	56	76	53	39	32	29	19	17	16	20
390	43	99	109	40	38	52	15	29	17	12
391	54	71	60	27	30	30	13	22	9	17
392	42	51	59	25	24	39	9	10	14	14
393	55	70	58	19	28	36	10	18	14	19
394	106	190	198	121	117	131	71	50	43	32
395	33	46	43	38	26	29	14	9	9	10
397	41	76	76	30	36	47	9	17	10	14
399	26	42	42	15	29	21	6	5	8	9
400	32	88	57	27	30	38	28	24	17	21
401	109	175	202	95	96	92	43	28	33	34
402	29	62	46	31	40	36	26	27	17	13
403	102	258	225	117	101	98	69	54	45	39
404	63	91	92	55	58	49	17	16	19	11
405	33	65	53	33	19	26	11	4	4	13
406	71	128	111	58	55	51	30	28	20	21
407	36	79	72	52	37	38	15	16	8	17
408	151	291	252	133	139	125	59	40	32	32
409	137	241	281	134	127	114	53	47	52	37
410	229	382	373	219	194	150	86	73	34	37
411	29	69	81	39	38	38	26	16	14	9
412	29	68	45	28	28	27	15	17	7	7
413	49	68	49	34	54	39	13	16	15	20
414	137	256	243	133	106	135	63	39	42	34
415	117	206	165	75	107	101	40	45	31	35
416	85	164	134	67	94	58	39	31	19	23
417	29	42	26	20	24	26	26	17	10	13
418	147	241	276	160	171	147	48	48	34	34
420	141	192	204	78	89	115	36	44	39	22
422	50	111	129	52	53	66	34	22	22	24
423	62	76	93	51	48	60	22	18	8	20
424	85	180	177	64	96	103	29	33	19	23
425	43	88	101	69	69	59	26	26	20	31
426	39	55	47	35	39	31	18	23	23	24
427	57	130	127	74	73	63	33	23	31	30
428	31	71	91	22	33	43	25	15	17	9
429	29	61	77	25	27	24	18	24	8	14
430	129	205	152	90	108	77	67	53	40	30
431	76	166	159	76	68	72	44	52	33	35
432	92	196	202	69	86	110	57	53	17	31
433	27	63	54	33	34	31	14	16	8	0
434	120	291	235	102	117	113	62	54	32	28
435	25	101	70	35	40	45	18	16	14	10
436	30	66	53	26	25	48	24	23	11	6
437	17	61	55	20	28	31	16	18	9	13
438	47	97	92	42	48	52	30	25	7	21
439	43	108	58	34	37	52	21	28	11	12
440	54	74	60	29	32	40	23	12	16	5
442	86	163	155	61	75	89	41	44	28	25
443	26	47	21	9	20	36	17	20	13	18
444	80	114	133	60	79	61	46	36	24	26
445	96	225	168	75	75	112	53	58	37	23
446	115	233	240	142	117	121	68	52	47	57
447	58	74	64	45	33	43	19	21	17	18
448	147	213	226	129	127	133	72	62	49	28
449	56	100	112	48	46	42	31	22	27	13
450	178	276	208	119	159	135	96	68	52	39
451	57	126	89	46	41	77	41	45	18	16
452	176	254	239	107	121	170	85	67	70	54
454	37	77	53	22	35	22	27	21	12	16
456	169	303	275	155	153	163	95	68	55	51
457	31	45	45	16	27	28	13	12	5	9
458	34	78	55	23	39	47	21	11	12	10
459	135	302	220	138	158	146	109	68	57	29
460	30	72	66	27	36	55	27	16	16	15
462	28	72	68	39	40	32	24	11	15	14
463	211	335	219	148	189	149	120	91	63	54
464	46	59	53	30	24	36	18	17	21	17
465	58	95	71	40	45	66	32	23	8	15
466	73	111	93	40	68	56	30	33	24	23
467	56	128	114	49	48	73	36	22	28	28
468	45	66	62	51	40	51	26	32	19	19
471	96	196	183	80	93	96	42	36	33	26
472	100	182	171	92	99	103	52	49	41	24
473	53	126	106	72	75	68	29	29	30	15
474	99	260	252	126	157	118	84	37	50	41
475	34	98	76	36	38	39	19	23	15	6
476	128	198	199	93	89	79	53	52	31	29
477	140	233	179	130	105	88	42	38	33	24
478	165	298	283	131	127	152	62	72	43	38
479	107	183	175	77	89	84	51	35	34	34
480	78	123	150	57	54	65	42	25	34	13
481	159	282	262	137	124	122	63	71	56	46
482	50	47	61	18	32	47	23	32	14	11
483	158	220	207	80	120	93	65	54	62	50
484	59	67	68	37	26	32	15	14	17	6
485	96	174	106	57	77	86	42	54	24	25
486	45	88	91	46	40	28	13	30	11	10
487	48	49	62	53	41	36	21	9	16	13

Table B.9: Word length counts for all the words in Tirant lo Blanc (2/2)

Chap.	e	de	la	que	lo	en	a	per	no	l
1	12	15	9	8	10	6	1	4	1	7
2	26	28	19	9	10	12	11	8	3	2
3	66	46	48	53	26	20	22	20	19	9
4	33	29	34	13	9	21	13	11	5	7
5	63	46	42	34	33	17	16	21	8	12
6	35	15	27	23	27	16	13	11	7	10
7	20	20	10	16	3	6	4	5	5	5
8	13	9	13	6	1	9	6	6	4	5
9	12	9	9	7	6	4	4	7	3	4
10	44	27	29	21	16	14	19	18	11	15
11	23	18	23	25	19	6	18	8	6	11
12	20	8	8	13	13	5	5	5	4	4
14	64	37	27	36	35	17	20	20	6	22
15	21	26	13	23	15	9	10	9	16	4
16	12	6	9	19	10	4	8	9	6	4
17	42	19	25	22	22	11	9	9	13	4
18	37	25	18	12	19	15	8	12	5	1
19	71	40	45	43	48	24	18	19	18	19
20	29	25	17	15	14	15	13	6	7	1
21	35	21	16	23	9	16	10	16	8	3
22	44	28	45	19	10	14	14	12	7	5
23	47	24	27	34	22	14	14	15	12	2
24	33	23	19	27	15	15	10	12	11	7
25	60	20	30	32	35	15	18	11	9	14
26	86	48	61	37	44	32	24	17	7	14
27	114	72	79	79	67	31	35	26	32	25
28	26	15	10	14	10	8	8	9	4	6
29	24	23	21	18	20	13	7	14	2	7
30	10	12	4	7	5	4	6	0	6	14
31	12	15	6	13	4	9	8	5	2	12
32	54	28	12	26	18	17	10	13	8	9
33	37	28	16	19	19	10	13	16	12	18
34	32	30	22	17	22	11	13	12	9	16
35	34	26	21	18	42	19	10	18	5	19
36	15	19	2	17	18	5	7	11	4	11
37	22	14	4	6	8	9	5	3	5	5
38	45	20	34	23	28	11	8	18	8	12
39	61	37	21	37	25	15	22	20	9	26
40	12	10	6	7	5	5	8	4	1	0
41	45	36	23	36	33	18	10	16	18	6
42	43	31	13	19	13	11	7	6	6	2
43	14	10	15	6	11	5	3	5	1	4
44	58	41	32	14	14	12	8	12	3	4
52	60	26	11	19	17	8	13	8	5	6
53	24	17	11	14	15	9	4	6	5	1
54	14	20	14	5	5	4	5	5	2	0
55	62	49	21	34	19	30	11	15	10	11
56	60	46	20	30	30	8	24	15	20	8
57	69	38	47	49	34	17	22	29	26	9
58	17	15	5	4	13	3	7	6	3	5
59	92	35	42	38	59	35	33	11	12	34
60	62	57	44	43	28	22	27	22	15	12
61	15	4	6	14	8	2	5	6	1	0
63	21	9	9	17	11	7	5	13	5	7
64	60	46	31	28	16	12	22	21	15	12
65	77	60	42	46	34	31	34	27	30	20
66	11	10	2	13	12	6	8	9	7	4
67	138	86	69	81	69	43	55	38	26	32
68	213	140	100	108	118	67	86	55	42	58
71	149	96	78	85	75	46	49	37	33	39
72	51	31	29	40	28	13	13	14	15	42
73	74	38	18	45	29	29	25	13	16	27
74	75	65	42	48	34	28	33	30	16	15
75	21	13	7	20	9	6	6	6	7	3
76	37	34	22	15	15	19	14	9	10	9
78	43	21	23	17	16	10	9	13	4	9
80	77	41	36	36	47	15	17	27	8	12
81	163	96	96	83	77	55	70	49	46	46
82	40	18	19	24	18	8	14	9	7	14
83	19	11	9	8	1	6	4	9	3	1
84	64	45	30	27	33	19	28	15	6	18
85	87	80	61	48	57	35	20	13	12	15
92	63	31	34	32	18	18	33	18	10	27
96	23	7	12	12	15	7	8	7	2	4
97	31	25	12	15	9	5	18	11	8	11
98	140	103	97	114	90	54	35	42	36	32
99	95	63	42	56	43	34	19	19	17	16
100	209	149	141	136	59	82	68	58	43	26
101	133	79	85	65	64	41	48	41	27	18
102	13	11	12	11	4	4	4	3	2	4
103	18	13	19	14	10	7	8	7	15	1
104	123	93	83	86	60	51	36	36	21	18
105	105	64	58	64	51	23	30	27	11	14
106	127	72	80	80	59	41	35	36	31	31
107	50	41	22	23	21	19	17	11	5	14
107	33	18	22	13	7	10	15	8	10	3
108	84	44	37	59	41	35	32	30	12	7
109	146	102	103	84	46	42	50	41	33	24
110	178	115	121	125	148	73	78	56	45	54
111	50	38	45	40	18	16	24	21	6	8
112	25	22	7	8	14	7	5	3	2	7
113	78	49	65	34	33	48	25	20	16	16
114	113	98	69	58	49	45	45	24	20	15
116	75	58	40	42	25	27	23	24	17	27
117	68	44	81	39	31	25	25	34	17	27
118	33	37	20	28	10	12	18	9	16	6
119	196	177	127	146	78	68	76	58	42	65
120	10	2	4	10	6	5	4	3	9	1
121	59	37	40	45	20	30	31	29	19	22
122	16	16	12	9	5	6	1	4	3	2
123	30	36	25	29	5	14	16	8	10	13
124	105	71	91	67	34	39	42	47	24	34
125	119	83	75	86	30	34	45	34	28	47
126	66	54	43	49	16	20	26	27	20	23
127	106	65	87	70	36	38	45	24	33	23
128	35	20	17	22	9	10	13	9	8	6
129	79	69	75	70	29	28	44	41	28	22
130	21	16	20	20	8	11	8	10	6	11
131	71	44	40	39	22	22	26	30	11	21
132	108	104	80	52	54	35	35	25	27	29
133	233	136	132	146	117	81	95	69	71	49
134	124	68	48	62	49	26	45	28	35	29
136	12	15	7	7	5	6	5	5	2	2
137	75	57	47	53	24	16	39	32	18	13
138	137	97	91	72	47	51	49	51	42	37
139	27	13	13	8	13	7	12	9	0	5
140	79	57	35	35	59	17	28	22	7	17
141	220	135	130	133	110	62	66	66	57	67
142	41	24	15	32	26	14	18	9	16	6
143	256	142	114	124	66	89	81	67	96	49
144	35	8	6	16	18	6	5	9	1	7
145	119	62	65	63	54	40	46	34	25	35
146	164	115	111	146	53	77	61	71	78	63
148	73	37	37	35	44	25	27	24	19	13
149	103	70	49	58	65	28	37	26	23	16
151	39	36	23	24	26	22	13	16	16	7
153	74	49	41	31	32	25	35	35	22	5
154	201	164	124	139	116	86	80	78	103	42
155	96	72	81	77	40	45	55	40	30	46
156	39	20	32	14	7	14	14	12	10	3
157	210	147	152	105	108	58	74	56	41	57
159	112	84	74	78	85	32	32	36	36	22
161	248	165	182	187	101	71	106	100	85	71
162	92	37	60	55	24	30	29	22	27	18
163	182	109	95	131	82	50	49	55	67	34
164	181	119	114	88	71	62	50	32	39	31
165	7	11	9	9	2	3	2	7	6	2
166	100	67	57	56	38	30	30	22	26	22
167	56	48	36	34	21	27	27	14	13	22
169	31	13	29	17	6	5	15	8	11	6
170	6	12	5	13	4	0	7	10	2	4
171	15	14	8	11	5	4	7	10	9	6
172	66	43	45	42	22	31	19	33	28	11
173	49	32	37	40	18	19	29	18	13	17
174	15	12	19	13	6	5	7	3	4	2
175	16	10	14	11	9	4	8	8	9	5
176	24	17	24	16	5	6	9	5	7	10
177	24	18	20	18	20	6	10	13	5	11
178	58	32	42	28	21	15	23	23	10	13
179	23	22	25	18	9	16	16	8	11	7
180	35	18	22	24	8	11	12	14	15	6
181	37	15	10	23	13	5	5	15	5	10
182	43	29	29	31	25	20	10	26	20	8
183	39	27	20	29	24	14	11	19	14	12
184	17	10	8	13	22	7	4	8	7	3
185	22	16	17	7	2	7	6	3	3	10
189	354	232	245	187	168	133	104	101	95	135
190	15	16	12	4	7	11	3	4	3	6
191	16	6	10	6	8	6	8	2	3	8
192	20	18	8	17	11	11	3	6	12	5
194	19	9	18	17	3	8	4	2	2	5
195	27	16	10	12	8	7	6	15	7	3
201	34	23	27	9	8	9	17	9	7	7
202	32	24	28	19	18	10	10	10	5	12
207	15	9	12	13	5	5	4	4	0	5
208	27	32	30	24	13	15	10	16	13	10
209	19	20	20	29	12	10	15	5	17	4
210	31	37	41	35	17	23	21	17	18	12
211	26	13	16	18	6	14	5	7	5	4
212	16	7	13	11	1	6	9	12	4	0
214	32	27	26	33	11	18	20	12	15	9
215	48	34	17	32	16	18	15	19	23	7
216	46	34	33	46	19	12	14	18	14	5
217	10	14	17	23	7	10	4	6	13	5
218	83	57	53	45	33	29	30	17	31	20
219	18	15	12	7	4	9	3	11	6	3
220	59	30	43	60	10	19	28	14	26	29
221	67	34	63	54	24	20	7	8	14	8
222	109	64	45	42	54	27	35	32	20	31
223	13	7	4	11	9	9	5	6	8	3
224	41	35	33	28	15	28	21	10	25	16
225	33	21	25	43	23	20	21	14	22	2
226	42	29	39	43	19	14	21	22	20	7
227	28	29	29	28	12	9	13	22	22	11
228	62	46	53	51	15	37	26	25	32	12
229	71	52	41	36	25	31	26	27	30	22
230	32	22	29	22	7	13	11	10	18	6
231	62	42	60	64	30	35	24	21	26	14
232	24	12	6	24	4	10	7	7	3	7
233	132	58	96	81	45	35	36	32	48	29
234	84	53	72	84	45	42	40	26	33	29
235	9	8	8	7	3	4	4	5	7	0
236	101	57	62	84	54	41	43	39	47	31
237	16	13	20	16	7	9	5	8	5	4
238	49	36	25	34	30	21	15	9	21	24
239	69	57	51	53	31	27	18	24	18	17
240	16	21	15	25	11	12	9	11	8	1
241	24	15	18	16	7	9	5	3	10	5
242	10	11	16	20	6	8	6	7	7	3
244	24	24	25	28	5	9	9	8	12	3
245	10	8	8	9	6	6	1	1	7	1
248	17	11	24	14	4	7	9	4	11	6
249	21	21	23	22	11	11	11	10	15	10
250	15	20	14	14	12	8	3	2	6	3
251	24	14	20	24	8	9	9	9	16	5
252	13	21	15	12	7	12	13	6	8	4
253	19	11	12	14	10	10	11	6	5	2
254	47	25	33	20	10	16	17	15	11	22
255	16	14	13	18	9	7	4	11	7	6
256	21	7	6	17	4	6	10	7	6	4
257	13	8	10	15	8	9	8	9	8	7
258	65	40	47	45	14	23	17	26	22	26
259	13	9	7	10	3	6	3	7	5	5

Table B.10: Most frequent function word counts in Tirant lo Blanc (1/2)

Chap.	e	de	la	que	lo	en	a	per	no	l
260	63	38	55	52	30	32	16	11	19	23
262	168	93	130	127	69	104	63	46	76	59
263	65	36	36	48	30	30	24	17	16	19
264	94	68	66	79	24	37	38	35	35	32
265	37	31	30	22	15	19	16	15	5	11
266	30	19	15	15	8	17	3	13	16	3
267	15	10	2	9	5	8	3	8	11	0
268	55	41	34	37	21	11	15	20	26	7
269	84	41	60	63	36	40	34	44	24	25
270	10	10	13	8	2	9	1	5	5	2
271	33	16	22	19	8	12	5	15	10	4
272	20	12	12	10	7	9	10	7	4	5
273	15	14	16	18	7	5	6	6	9	1
274	7	19	16	14	2	4	6	8	4	3
275	35	20	11	20	8	15	14	18	11	17
276	24	15	19	14	8	14	2	7	4	0
277	52	28	34	31	21	25	16	13	18	18
278	30	25	17	11	5	16	9	6	4	9
279	32	14	15	20	10	9	15	6	13	5
280	14	12	10	12	9	13	10	8	13	4
281	72	54	53	54	19	30	27	20	21	21
282	48	34	35	45	26	29	29	24	24	17
283	88	56	77	64	31	50	33	28	19	26
284	10	12	10	7	2	5	5	4	5	0
285	18	13	16	7	7	15	3	3	6	1
286	63	34	50	41	19	20	36	21	28	11
288	87	57	59	58	25	29	33	25	26	18
289	14	9	10	8	3	3	3	0	6	1
290	40	24	34	31	14	16	24	7	33	8
291	120	61	81	62	44	41	44	33	37	23
292	102	57	83	89	41	41	37	27	21	31
293	37	25	34	31	12	13	16	14	13	18
294	17	15	12	6	8	13	8	6	10	1
295	34	38	45	33	18	16	17	25	21	8
296	45	29	38	32	13	19	17	7	10	6
297	22	22	13	14	4	7	7	9	9	3
299	108	77	120	89	36	59	40	47	29	36
300	9	12	11	16	5	9	6	6	9	1
301	96	53	51	47	53	35	29	28	17	16
302	21	14	6	13	16	14	9	7	4	5
303	17	20	9	22	4	12	9	8	13	0
304	117	52	47	56	55	28	35	29	29	16
305	6	9	9	9	0	2	5	7	3	1
306	16	10	13	6	8	4	9	7	1	0
307	74	41	38	47	28	26	25	26	20	11
308	16	22	8	12	8	10	7	13	15	1
309	130	98	58	89	56	47	36	35	41	18
310	163	93	95	76	89	50	30	48	32	18
311	15	11	9	19	8	6	1	7	7	2
312	58	31	33	42	30	16	12	11	19	12
313	26	12	13	19	11	7	10	15	5	7
314	19	17	12	8	1	3	8	1	5	1
315	142	64	91	63	37	34	38	40	20	27
316	13	9	17	20	8	6	4	9	3	1
317	46	38	32	26	24	28	14	19	19	8
318	15	13	12	17	13	7	9	8	12	6
319	75	54	50	68	40	30	26	27	32	11
320	12	5	7	9	7	4	7	6	8	2
321	80	71	41	58	48	30	21	29	21	13
322	21	17	19	20	7	6	5	5	8	1
323	21	25	14	26	12	12	8	18	16	3
324	29	26	24	34	12	14	10	11	18	1
325	26	19	23	21	6	15	10	13	16	2
326	13	12	11	10	7	8	4	1	3	1
327	20	26	19	33	11	14	12	14	8	6
328	52	25	8	25	24	16	14	20	2	6
329	42	24	19	12	16	16	11	10	3	4
330	65	45	32	33	36	23	19	19	9	8
331	25	17	8	20	7	9	6	3	19	3
332	10	5	8	17	4	12	3	4	11	1
333	40	30	23	26	19	14	12	18	12	2
334	111	46	51	56	32	20	22	31	22	12
335	35	18	24	27	17	11	17	12	10	8
336	7	6	11	8	5	2	3	3	6	0
337	62	30	28	24	40	23	19	12	18	6
338	13	11	7	13	2	7	4	5	3	1
339	66	28	35	37	20	23	10	15	16	7
340	142	66	59	73	54	36	39	41	19	13
341	10	16	5	12	2	7	7	7	6	1
343	88	45	31	45	30	24	26	15	21	6
344	61	29	29	28	39	16	20	11	12	10
345	47	22	25	21	25	18	17	15	14	4
346	36	13	14	13	14	12	4	9	5	3
347	37	23	21	17	18	9	17	14	14	4
348	17	9	10	5	3	4	5	3	5	1
349	135	71	78	71	71	37	45	36	18	13
350	61	39	52	47	20	20	26	16	7	9
351	23	18	18	21	10	7	9	6	9	2
352	14	9	3	6	3	5	6	3	7	1
353	33	29	21	29	9	12	14	14	14	7
354	51	37	32	24	13	16	14	21	21	4
355	71	61	55	39	26	38	26	20	31	15
356	13	13	13	20	8	8	5	11	11	0
357	48	41	45	31	18	24	15	20	17	12
358	14	8	10	10	9	5	4	4	1	1
359	22	8	9	16	11	7	3	4	8	3
360	20	9	17	17	13	4	5	6	4	4
362	23	16	25	12	6	12	7	7	6	6
363	12	15	9	8	4	4	1	3	4	7
364	11	17	8	10	12	8	7	8	6	0
365	7	10	14	14	6	4	3	4	4	1
366	40	24	28	17	9	21	9	11	7	8
367	40	24	28	17	9	21	9	11	7	8
368	20	6	12	19	9	8	6	5	7	2
369	7	27	10	11	7	9	7	11	15	6
370	29	28	26	18	9	6	11	13	8	5
371	36	27	31	35	11	17	14	15	10	3
372	47	34	39	36	17	19	16	13	10	13
373	23	16	18	19	1	6	7	6	8	5
374	86	49	57	39	12	26	28	17	34	10
375	33	9	15	7	4	4	8	1	8	0
376	55	24	25	11	12	13	12	12	15	10
377	25	14	13	2	6	5	1	7	9	4
378	27	14	14	15	9	11	8	11	16	5
379	8	8	8	4	4	7	0	3	9	4
380	10	16	17	7	9	12	2	8	10	5
381	6	12	10	9	6	2	8	6	4	1
382	19	14	8	9	9	2	9	8	5	3
383	25	21	10	9	3	8	3	5	1	7
384	39	22	18	10	15	16	9	13	2	2
385	14	16	14	6	9	3	4	6	1	6
386	38	29	32	17	17	5	7	9	6	3
387	92	70	57	63	45	30	25	27	23	14
388	24	14	15	6	10	6	10	7	1	19
390	28	19	17	17	14	9	7	11	3	3
391	25	21	18	15	2	5	8	1	3	16
392	20	15	12	7	2	4	9	1	3	7
393	29	18	19	9	8	5	8	5	1	14
394	62	38	51	19	22	14	19	29	15	8
395	24	9	11	7	4	3	6	2	0	3
397	25	9	14	21	9	12	4	6	7	6
399	17	8	10	6	3	10	2	5	2	5
400	16	12	15	6	11	7	8	6	3	1
401	75	57	23	37	24	16	19	16	2	4
402	15	13	11	4	11	3	5	8	2	2
403	54	53	47	36	19	30	16	27	19	16
404	43	24	12	14	16	6	5	7	2	5
405	18	11	10	6	16	9	7	3	0	7
406	42	30	16	11	21	17	15	9	3	6
407	18	21	11	16	17	8	8	10	3	7
408	96	104	56	33	35	17	23	8	7	16
409	87	61	44	33	27	29	17	33	6	6
410	131	55	54	52	53	37	32	28	35	27
411	15	11	12	21	4	12	9	9	3	1
412	17	10	14	13	6	9	4	6	4	2
413	36	15	17	7	7	4	7	7	3	3
414	81	63	40	35	36	26	21	15	10	26
415	60	51	45	29	16	9	18	20	7	24
416	44	24	48	38	14	13	14	14	7	13
417	10	10	13	3	1	3	7	6	4	5
418	92	51	50	62	24	25	16	16	20	20
420	79	39	49	44	24	12	24	23	9	27
422	31	18	15	33	21	8	9	12	19	3
423	35	19	14	24	6	5	10	3	1	4
424	53	26	34	44	14	15	15	18	10	6
425	27	18	15	15	20	12	7	10	3	1
426	17	14	10	4	3	6	9	4	1	2
427	30	23	32	26	16	17	12	12	11	5
428	16	15	12	14	15	2	12	10	8	1
429	21	11	7	11	9	7	4	11	4	0
430	67	34	62	16	16	20	24	11	3	31
431	38	31	37	33	10	23	13	10	8	12
432	46	32	28	33	12	17	20	25	18	13
433	13	7	12	8	6	5	6	4	9	4
434	51	47	69	48	20	20	39	14	15	12
435	9	14	22	16	9	9	6	5	9	3
436	12	11	12	9	1	2	3	5	15	5
437	4	9	15	13	3	4	8	3	7	1
438	17	13	24	19	8	7	7	6	8	6
439	18	14	24	15	5	9	10	7	5	6
440	22	13	14	14	8	5	9	5	6	14
442	42	30	35	29	18	18	18	23	10	14
443	11	9	11	2	7	3	6	1	2	1
444	41	22	16	32	18	9	15	16	8	11
445	39	33	61	31	31	13	24	18	10	13
446	79	48	36	55	40	23	23	15	11	6
447	26	15	15	13	10	6	10	3	4	7
448	82	48	42	28	25	32	20	15	6	32
449	36	24	9	20	12	12	10	8	4	5
450	86	68	55	19	39	22	31	17	3	43
451	30	27	23	17	10	7	11	6	12	9
452	85	50	49	46	35	23	29	26	11	41
454	16	18	14	6	11	9	7	4	3	9
456	109	58	73	39	54	8	21	21	10	10
457	16	10	12	5	3	3	7	0	3	4
458	20	14	17	5	11	3	4	4	2	3
459	82	60	64	37	46	26	15	20	5	14
460	7	14	17	11	5	5	6	5	8	3
462	21	17	9	9	15	11	4	8	2	2
463	102	80	118	39	10	27	31	28	7	35
464	23	13	25	6	3	5	9	13	2	9
465	41	26	20	8	16	6	1	14	4	9
466	40	33	23	10	13	11	11	15	6	14
467	32	22	15	18	21	14	9	14	7	10
468	16	11	15	6	6	11	9	6	2	3
471	59	52	32	24	32	13	13	17	3	8
472	33	37	32	17	22	12	15	13	9	16
473	29	21	28	16	16	5	5	5	10	5
474	67	43	48	44	26	23	14	13	18	6
475	16	18	11	11	7	8	6	4	9	4
476	70	42	35	27	14	25	35	12	11	10
477	74	42	39	26	16	23	27	16	11	18
478	126	60	66	23	14	35	14	20	11	5
479	57	50	41	34	16	17	19	17	5	18
480	45	32	15	33	19	9	9	19	3	12
481	82	54	42	40	43	26	32	37	12	28
482	31	8	11	14	1	3	9	7	5	7
483	85	59	39	36	24	12	23	16	14	25
484	31	19	13	12	10	7	15	3	2	7
485	59	66	28	14	12	21	7	8	2	15
486	28	29	14	10	14	13	4	14	1	8
487	29	13	8	10	8	4	4	4	2	10

Table B.11: Most frequent function word counts in Tirant lo Blanc (2/2)

B.2.2 Don Quijote de la Mancha

Don Quijote de la Mancha, fully titled in spanish; *El ingenioso hidalgo don Quijote de la Mancha*, is a Spanish novel by Miguel de Cervantes Saavedra. Published in two volumes, in 1605 and 1615, Don Quixote is considered one of the most influential works of literature from the Spanish Golden Age and the entire Spanish literary canon. As a founding work of modern Western literature and one of the earliest canonical novels, it regularly appears high on lists of the greatest works of fiction ever published, such as the Bokklubben World Library collection that cites Don Quixote as authors' choice for the "best literary work ever written". It follows the adventures of a nameless hidalgo who reads so many chivalric romances that he loses his sanity and decides to set out to revive chivalry, undo wrongs, and bring justice to the world, under the name Don Quixote.


EL INGENIOSO
HIDALGO DON QVI-
XOTE DE LA MANCHA,
Compuesto por Miguel de Cervantes Saauedra.
DIRIGIDO AL DVQVE DE BEIAR,
Marques de Gibraleon, Conde de Benalcaçar, y Bañares, Vizconde de la Puebla de Alcozer, Señor de las villas de Capilla, Curiel, y Burguillos.
Año, 1605.
CON PRIVILEGIO,
EN MADRID Por Iuan de la Cuesta.
Vendese en casa de Francisco de Robles, librero del Rey nro señor.

Author	Miguel de Cervantes
Country	Spain
Language	Spanish
Literary Genre	Burlesque
Editor	Francisco de Robles
Publication Date	1605 part I 1615 part II

PART I											PART II										
Chap.	1	2	3	4	5	6	7	8	9	10+	Chap.	1	2	3	4	5	6	7	8	9	10+
1	155	428	304	181	229	199	151	74	70	87	12	179	537	440	195	252	208	190	131	99	98
2	166	489	385	157	277	232	193	133	93	81	13	178	526	394	253	262	228	153	107	74	86
3	178	511	392	215	301	199	217	134	81	89	14	322	924	769	340	419	424	346	222	166	129
4	186	547	428	260	288	249	194	126	81	79	15	61	172	124	55	80	75	64	48	32	24
5	130	372	270	157	214	164	139	61	39	54	16	256	893	623	288	428	351	298	177	132	146
6	175	575	457	294	333	229	185	120	77	78	17	327	881	681	360	459	373	291	200	132	138
7	133	432	360	173	231	173	138	91	77	59	18	236	742	536	253	375	264	259	144	103	104
8	194	710	537	287	327	315	242	189	98	94	19	181	690	495	253	317	267	218	147	106	112
9	158	451	341	166	238	167	151	122	69	82	20	238	745	531	303	376	316	280	193	103	101
10	128	455	340	204	247	188	147	96	77	49	21	196	584	378	196	296	201	210	136	94	96
11	163	458	407	188	258	189	173	125	80	95	22	246	622	443	292	353	248	220	133	114	125
12	168	563	423	250	285	210	177	84	71	92	23	279	881	645	388	476	387	305	166	135	174
13	268	793	639	304	401	323	291	195	132	165	24	185	565	386	217	313	244	162	121	88	92
14	187	737	517	240	337	273	236	155	113	116	25	284	858	644	372	487	310	266	142	128	110
15	214	610	501	249	373	281	213	139	106	84	26	254	683	507	297	419	263	213	169	106	137
16	238	614	486	262	335	255	257	143	102	103	27	193	558	442	249	313	218	205	125	108	92
17	237	750	570	318	391	308	236	145	110	108	28	139	476	341	217	200	206	168	101	46	46
18	268	853	716	390	438	413	332	188	152	128	29	217	497	417	226	292	229	173	118	95	94
19	207	620	468	295	331	286	203	130	106	113	30	149	451	297	163	218	171	148	104	73	51
20	376	1146	927	531	674	524	378	250	161	127	31	243	687	513	318	320	292	251	174	106	91
21	301	1049	790	484	520	418	348	179	129	124	32	413	1196	928	446	634	490	370	260	193	215
22	315	911	720	423	431	400	333	190	169	101	33	212	651	512	290	326	262	183	113	81	67
23	308	891	699	372	523	398	323	213	129	126	34	212	654	476	254	343	274	204	129	108	110
24	244	853	587	271	405	305	276	228	114	128	35	235	630	430	261	292	288	176	131	86	121
25	462	1485	1109	689	731	690	434	297	196	184	36	159	540	326	192	271	202	163	110	77	81
26	228	743	512	282	380	271	223	138	96	75	37	57	174	131	80	93	62	47	52	34	17
27	444	1571	1071	618	809	528	460	394	217	226	38	175	528	380	206	258	207	172	137	86	115
28	415	1417	1041	565	684	501	379	372	172	234	39	69	189	153	79	122	101	69	63	43	40
29	363	1076	812	475	543	398	366	255	155	136	40	137	454	350	184	244	190	155	113	79	57
30	256	893	591	422	457	330	307	180	89	115	41	290	850	661	402	477	399	274	175	115	118
31	263	826	621	398	435	372	256	161	103	82	42	133	463	346	170	215	199	130	95	68	74
32	166	581	431	269	310	255	194	108	73	101	43	163	479	377	212	215	234	149	99	53	83
33	585	1804	1345	702	952	654	675	317	234	269	44	249	816	586	291	359	271	284	181	113	120
34	541	1752	1219	651	917	748	635	319	215	266	45	233	652	365	300	316	267	160	122	100	89
35	262	827	552	357	417	332	279	125	85	97	46	115	331	243	120	162	144	121	94	50	60
36	317	854	673	353	434	334	304	258	114	138	47	255	800	464	339	388	319	210	145	97	132
37	300	960	683	411	480	382	311	222	112	140	48	267	726	527	335	379	316	254	136	105	108
38	125	367	290	143	218	150	129	83	45	65	49	358	955	666	440	510	426	261	186	133	166
39	252	735	505	283	336	325	261	170	95	116	50	284	777	590	361	419	346	214	173	84	86
40	354	1061	825	498	588	431	367	226	115	164	51	231	746	551	303	334	321	206	143	110	122
41	629	1797	1329	712	966	739	593	447	236	289	52	241	612	430	275	333	247	203	131	75	82
42	217	657	476	269	335	196	253	169	85	94	53	182	494	337	211	264	202	129	148	59	82
43	259	911	645	379	429	325	267	188	110	125	54	261	621	466	276	376	331	192	160	83	109
44	262	809	579	344	448	253	255	164	88	106	55	216	603	464	273	332	258	217	118	72	102
45	233	705	523	323	345	201	261	133	105	109	56	128	460	296	157	229	180	165	87	61	66
46	232	757	527	302	348	265	236	181	123	128	57	70	292	210	92	142	112	92	98	43	45
47	310	861	630	319	464	356	271	187	168	186	58	299	907	802	344	487	408	344	250	159	182
48	204	626	552	250	316	255	212	140	120	150	59	197	644	501	271	325	310	234	167	92	90
49	196	656	444	236	340	264	206	151	106	109	60	402	1000	786	375	566	387	426	251	194	187
50	193	524	436	226	315	200	202	163	97	86	61	84	207	182	69	104	108	80	63	59	44
51	167	471	357	184	298	149	187	108	54	89	62	326	1017	813	414	477	452	382	215	156	158
52	285	862	692	333	447	391	303	232	140	147	63	261	816	635	312	422	367	285	213	138	117
PART II											64	129	413	262	146	185	148	139	90	72	50
1	296	946	680	390	481	373	353	223	132	174	65	143	477	364	186	219	175	158	119	62	71
2	124	397	337	148	180	191	139	78	60	57	66	140	449	298	196	217	180	142	90	66	62
3	163	651	539	264	275	273	197	188	105	112	67	124	414	311	161	176	156	126	112	82	82
4	149	480	301	200	230	181	130	91	58	66	68	157	400	323	153	216	198	134	90	65	71
5	197	560	406	228	281	240	186	137	67	45	69	154	420	315	167	224	213	117	104	65	79
6	164	495	384	170	228	174	158	127	82	83	70	180	606	479	260	276	270	221	145	92	91
7	203	620	465	243	275	265	187	146	84	86	71	153	486	351	210	246	245	178	87	62	60
8	206	600	487	236	353	268	215	156	113	103	72	123	413	342	159	207	205	142	78	63	50
9	105	309	252	138	182	137	106	79	50	28	73	149	386	314	150	188	182	134	91	71	66
10	273	709	545	338	394	357	273	177	101	103	74	173	567	402	269	261	235	224	129	76	112
11	190	564	405	207	260	241	206	118	104	82											

Table B.12: Word length counts for all the words in El Quijote

PART I																			
Chapter	que	y	de	la	a	en	el	no	los	se	con	por	las	lo	le	su	don	del	me	como
1	88	105	120	33	44	38	40	20	21	30	27	17	8	16	23	35	5	18	4	13
2	113	103	92	64	58	54	45	33	17	30	29	20	29	21	26	33	12	15	6	18
3	125	118	113	69	59	46	33	34	30	44	33	28	36	18	30	30	21	13	5	18
4	131	126	100	64	57	38	59	41	22	24	30	35	17	27	31	24	19	15	12	16
5	90	71	78	34	56	23	46	24	20	20	14	10	12	12	31	20	13	17	4	4
6	124	135	134	53	38	41	93	36	41	40	28	29	21	11	25	20	4	29	7	14
7	108	83	83	32	44	38	39	29	24	27	21	30	7	27	24	18	19	10	10	3
8	165	119	143	65	65	74	65	55	25	34	31	34	29	29	40	41	30	25	15	18
9	98	109	102	62	44	42	50	25	23	26	26	23	18	17	20	18	12	13	12	9
10	118	76	91	43	48	44	29	41	17	21	14	15	16	17	16	19	17	11	14	11
11	122	106	91	49	56	40	40	25	36	25	26	21	28	20	18	18	8	18	18	11
12	125	111	131	67	52	52	47	41	42	40	24	25	8	28	11	39	9	20	9	16
13	187	169	191	91	96	83	69	47	66	47	39	35	24	35	23	34	22	26	15	21
14	153	101	145	88	68	61	66	49	46	33	45	42	28	16	14	38	5	21	28	16
15	150	117	133	56	87	57	36	45	28	37	29	18	31	24	25	25	23	15	21	15
16	141	135	124	109	100	58	62	44	31	44	29	14	29	17	26	30	23	26	11	13
17	185	155	152	71	77	64	65	54	34	48	44	31	27	32	44	39	22	21	17	20
18	207	177	189	83	79	85	83	61	78	46	32	32	50	20	27	36	32	37	22	26
19	134	119	120	77	73	58	62	44	25	28	34	22	25	26	29	29	26	26	15	20
20	317	219	230	104	135	92	100	108	50	64	46	51	45	59	36	37	40	27	26	38
21	239	174	196	132	106	84	105	85	36	63	37	60	41	36	34	41	24	27	26	27
22	221	193	184	107	109	84	79	84	54	45	45	53	38	35	37	27	40	20	25	21
23	236	197	185	81	95	69	75	74	42	47	44	62	22	56	55	29	29	21	22	17
24	204	138	169	93	97	65	65	52	32	40	50	37	22	45	37	44	23	18	41	26
25	399	294	296	145	142	146	96	107	42	57	42	74	63	65	55	39	34	37	51	42
26	170	142	144	78	73	69	62	56	26	47	14	30	23	31	63	35	17	25	10	25
27	383	277	307	195	149	164	138	90	55	75	75	58	39	59	55	60	29	28	79	34
28	347	247	281	135	149	142	100	109	83	67	68	68	34	51	37	51	27	19	76	33
29	279	210	199	123	136	88	103	69	43	71	52	49	35	49	51	48	42	18	27	26
30	198	165	137	98	80	93	65	84	24	39	45	44	14	36	35	22	29	19	36	20
31	225	147	164	76	102	68	33	71	27	42	27	40	13	30	37	31	29	19	47	16
32	154	119	123	61	41	37	65	45	26	29	31	18	12	30	22	18	14	14	19	17
33	502	366	349	202	191	163	126	144	82	89	99	103	59	69	76	90	0	37	46	62
34	461	312	334	213	210	149	108	153	47	86	94	66	52	79	100	103	0	26	32	38
35	187	161	161	108	92	75	82	63	27	43	35	37	16	32	49	49	9	25	16	15
36	218	188	139	110	120	87	79	70	47	42	57	39	26	37	30	26	27	14	18	28
37	231	194	165	136	92	105	92	65	51	59	43	43	39	46	33	43	32	28	18	30
38	106	78	86	40	38	39	29	19	25	29	16	17	28	16	24	19	5	11	5	3
39	177	153	166	93	93	88	90	25	43	38	29	20	18	30	21	27	14	16	16	19
40	278	238	201	133	108	103	107	51	54	57	60	61	25	64	26	23	0	30	36	27
41	454	356	346	236	246	172	141	113	100	104	90	74	54	60	59	64	0	32	50	44
42	162	124	140	84	86	66	73	23	34	36	39	19	32	26	30	51	12	20	11	16
43	209	148	183	110	90	72	61	76	27	51	44	36	28	34	42	40	17	25	26	21
44	199	146	143	96	108	74	80	59	39	43	30	30	19	43	44	42	38	16	16	23
45	149	131	145	76	92	58	82	60	42	52	28	29	13	27	31	30	53	17	13	21
46	158	164	164	95	63	65	60	57	28	42	27	38	29	37	35	32	34	22	14	22
47	190	200	207	94	91	84	79	59	62	37	44	36	31	25	30	29	26	24	18	27
48	195	143	139	72	51	59	46	42	50	40	45	23	52	37	17	13	4	17	12	27
49	154	128	176	73	58	63	52	38	38	29	22	20	21	29	19	18	18	12	13	30
50	131	132	121	55	54	32	53	34	36	25	26	24	24	19	17	17	8	21	16	24
51	118	101	126	76	59	41	37	26	28	33	22	14	19	9	12	37	1	13	1	7
52	205	171	209	103	109	87	111	44	68	52	40	25	39	25	34	43	43	27	11	13
PART II																				
1	212	195	197	93	80	98	119	74	53	44	44	49	24	24	34	42	33	24	23	25
2	98	74	89	40	46	33	35	23	27	18	14	17	18	21	10	6	20	11	19	3
3	150	88	146	88	65	60	50	61	52	46	22	21	35	21	21	11	26	21	12	21
4	116	86	103	31	52	43	36	39	23	24	15	11	14	21	14	17	11	10	16	9
5	131	119	106	60	67	42	35	62	23	19	44	28	11	26	12	17	6	9	20	17
6	119	111	98	49	40	54	38	48	48	20	22	23	17	21	18	16	9	7	6	13
7	146	134	104	58	58	34	46	62	18	42	28	27	26	37	19	37	22	12	21	16
8	149	121	129	77	63	77	53	32	46	37	24	29	28	22	15	21	21	22	12	13
9	72	58	57	41	41	34	30	22	11	11	13	15	4	10	11	12	15	10	10	8
10	157	162	142	88	97	59	48	58	26	27	23	29	38	27	25	29	26	27	19	36
11	108	113	132	74	68	46	53	25	49	35	37	12	30	12	17	32	24	10	12	8

Table B.13: Most frequent function word counts in El Quijote (1/2)

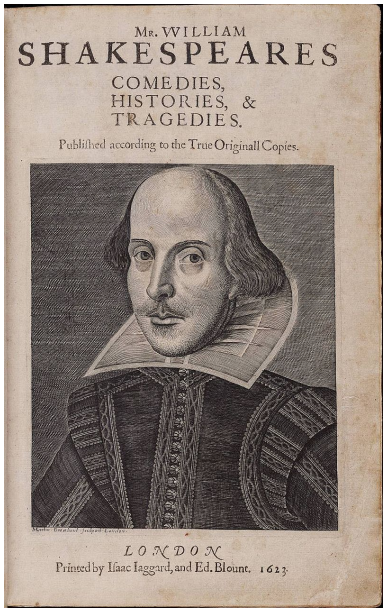
PART II																				
Chapter	que	y	de	la	a	en	el	no	los	se	con	por	las	lo	le	su	don	del	me	como
12	122	112	114	70	58	44	56	29	48	36	15	20	33	24	15	18	22	30	3	13
13	107	106	109	39	65	39	65	38	20	22	26	23	14	24	15	7	0	31	18	11
14	226	205	198	103	103	82	98	63	64	35	35	38	39	41	41	44	40	52	23	16
15	33	37	33	9	21	9	27	13	5	16	8	17	3	11	8	13	12	2	2	4
16	160	163	207	111	78	92	81	67	65	41	40	35	48	28	30	29	25	23	12	20
17	213	197	155	105	117	83	97	60	65	49	41	33	37	43	30	35	49	31	22	18
18	150	137	150	91	85	55	75	37	44	42	30	26	39	36	27	28	65	20	17	12
19	147	121	157	85	51	60	79	53	40	40	39	34	27	14	19	19	15	22	11	24
20	136	175	195	85	57	66	80	41	42	36	30	28	47	26	20	8	20	22	9	17
21	116	135	135	86	60	56	52	41	37	37	26	25	27	11	19	20	7	14	10	18
22	136	153	123	81	86	60	53	46	19	37	26	19	41	21	34	12	26	16	11	14
23	185	194	165	87	69	83	78	68	45	27	49	35	51	36	38	27	25	23	44	23
24	132	112	113	82	63	44	64	49	25	24	25	30	24	20	21	12	16	13	12	16
25	198	183	133	62	86	86	130	59	39	46	29	44	35	40	42	17	23	42	17	14
26	138	177	114	80	74	73	64	55	33	48	42	25	39	22	21	36	38	20	24	10
27	143	116	112	66	71	58	44	37	34	33	22	35	26	30	32	26	19	19	6	26
28	118	85	94	44	50	51	30	28	14	26	21	17	18	17	12	6	15	14	23	9
29	101	120	108	54	70	55	47	32	38	26	23	32	26	18	16	12	21	16	8	12
30	91	84	82	57	60	45	35	27	22	26	19	9	11	11	15	22	20	17	13	8
31	171	125	137	96	100	65	54	48	34	39	37	31	20	36	30	15	34	22	18	19
32	261	261	233	173	131	111	100	97	80	46	67	69	58	38	47	31	35	31	34	40
33	163	143	130	87	59	40	51	53	31	26	23	36	27	26	23	17	8	21	23	19
34	127	137	137	78	70	65	81	40	49	41	28	26	26	18	25	14	22	17	6	20
35	117	142	152	67	78	44	53	43	45	28	24	32	27	18	9	15	9	16	25	12
36	104	99	111	78	52	40	53	32	24	23	25	18	15	17	17	16	6	10	23	9
37	44	31	29	22	22	19	17	13	10	13	6	9	11	6	2	4	3	0	6	8
38	101	110	119	92	58	38	50	35	34	28	14	27	27	16	7	18	13	19	21	17
39	36	54	45	44	13	19	10	8	15	13	19	6	11	6	5	7	5	2	3	9
40	100	76	81	61	50	45	45	30	25	23	24	35	31	9	16	10	5	11	14	17
41	202	185	152	100	90	81	77	79	53	44	43	53	42	27	21	23	22	34	27	24
42	107	80	101	70	46	36	36	37	20	20	26	14	25	15	12	7	3	24	4	13
43	139	107	74	46	50	38	50	50	31	23	18	19	9	13	8	10	11	10	18	15
44	174	153	199	113	84	68	62	62	30	40	38	31	33	17	31	31	26	28	19	18
45	118	157	97	86	64	61	80	33	35	45	21	21	26	22	36	20	3	28	27	12
46	58	75	58	58	39	27	34	22	21	19	29	16	16	12	22	16	23	9	2	5
47	158	174	166	82	69	67	82	67	30	37	21	30	26	30	26	16	5	14	26	18
48	134	174	152	112	85	72	47	51	21	35	41	29	29	20	28	25	25	19	17	22
49	218	235	192	93	102	71	93	85	62	50	53	30	21	37	27	31	0	19	32	16
50	177	187	157	88	90	54	83	51	25	32	42	31	36	34	21	26	14	16	23	22
51	189	151	133	91	71	70	64	48	47	33	37	27	33	37	21	20	7	24	24	19
52	140	135	119	88	98	51	57	33	32	27	29	26	27	23	25	17	15	15	19	15
53	98	122	96	47	51	46	48	25	25	34	25	13	18	13	28	17	1	9	25	12
54	159	157	111	47	96	71	53	45	32	42	31	29	26	27	17	28	5	15	16	23
55	139	140	135	46	69	51	53	45	26	32	21	41	12	29	25	15	16	14	18	15
56	88	88	97	61	35	40	52	29	28	24	15	25	14	13	26	26	19	13	6	4
57	54	39	53	32	30	31	19	20	20	13	12	8	26	9	9	7	10	8	11	11
58	234	194	207	108	96	96	78	60	60	47	58	42	52	37	20	26	43	30	18	20
59	153	127	133	70	62	57	54	41	28	37	26	25	25	30	22	34	56	25	13	11
60	219	220	214	96	167	109	61	80	98	46	50	45	36	37	63	48	40	15	23	10
61	38	52	54	33	29	18	37	14	23	12	13	10	18	6	6	7	17	5	0	2
62	215	204	227	125	114	94	108	71	56	58	48	54	47	43	43	32	87	17	21	26
63	194	162	144	111	88	97	103	40	52	45	46	38	30	26	34	31	20	19	15	16
64	63	66	88	73	56	34	43	25	2	19	22	13	13	13	15	12	35	11	6	14
65	102	89	75	48	52	56	53	41	15	27	38	18	11	13	23	20	41	11	13	8
66	84	79	86	45	54	28	44	32	6	24	21	15	25	28	9	12	15	12	8	10
67	82	82	95	35	33	46	39	34	47	12	10	21	22	9	8	9	13	8	12	17
68	88	86	82	47	63	37	43	22	49	17	17	10	10	11	14	6	19	14	13	9
69	75	90	87	48	61	49	42	29	30	27	22	19	29	16	13	8	13	13	9	8
70	118	107	126	61	65	61	55	49	37	21	34	34	32	17	24	25	31	20	28	14
71	122	89	76	50	57	42	44	45	18	26	24	22	13	19	19	16	19	13	23	10
72	91	77	90	44	44	38	56	25	21	12	13	20	14	13	15	21	57	13	16	6
73	90	98	76	49	45	43	34	28	22	21	12	22	13	14	11	26	18	20	6	9
74	111	111	127	54	55	45	59	28	31	29	25	21	27	14	29	30	24	23	12	19

Table B.14: Most frequent function word counts in El Quijote (2/2)

B.2.3 William Shakespeare Plays

The texts used are the ones from, *Mr. William Shakespeares Comedies, Histories and Tragedies* is the 1623 published collection of William Shakespeare's plays named *First Folio*

Printed in folio format and containing 36 plays, it was prepared by Shakespeare's colleagues John Heminges and Henry Condell. Although eighteen of Shakespeare's plays had been published in quarto prior to 1623, the First Folio is arguably the only reliable text for about twenty of the plays, and a valuable source text for many of those plays previously published. The Folio includes all of the plays generally accepted to be Shakespeare's, with the exception of *Pericles*, *Prince of Tyre*, *The Two Noble Kinsmen*, and the two lost plays, *Cardenio* and *Love's Labour's Won.*



Author	William Shakespeare
Country	England
Language	Early Modern English
Literary Genre	English Renaissance theatre
Editor	Edward Blount and William and Isaac Jaggard
Publication Date	1623

num	Play	num	Play	num	Play
1	A Midsommer Nights Dreame	13	The Life of King Henry the Eight	25	The Tragedie of Julius Caesar
2	Alls well, that ends well	14	The Life of Timon of Athens	26	The Tragedie of King Lear
3	As you like it	15	The Merchant of Venice	27	The Tragedie of Macbeth
4	Loues Labours lost	16	The merry Wiues of Windsor	28	The Tragedie of Othello, the Moore of Venice
5	Measvre, for Measure	17	The Second Part of Henry the Fourth	29	The Tragedie of Richard the Third
6	Much adoe about Nothing	18	The second Part of Henry the Sixt	30	The Tragedie of Romeo and Juliet
7	The Comedie of Errors	19	The Taming of the Shrew	31	The Tragedie of Titus Andronicus
8	The First Part of Henry the Fourth	20	The Tempest	32	The Tragedy of Coriolanus
9	The first Part of Henry the Sixt	21	The third Part of Henry the Sixt	33	The two Gentlemen of Verona
10	The life and death of King John	22	The Tragedie of Anthonie, and Cleopatra	34	The Winters Tale
11	The life and death of King Richard the Second	23	The Tragedie of Cymbeline	35	Twelfe Night, or what you will
12	The Life of Hennry de Fift	24	The Tragedie of Hamlet		

Table B.15: List of Willian Shakespeare plays in First Folio

Play.Act	1	2	3	4	5	6	7	8	9	10+	Play.Act.	1	2	3	4	5	6	7	8	9	10+
1.I	215	704	972	1026	667	413	292	165	83	106	19.I	211	1290	1690	1452	1007	719	497	295	228	221
1.II	207	658	841	855	509	365	234	156	64	77	19.II	237	1075	1317	1052	715	512	339	213	173	153
1.III	170	491	611	646	386	252	209	109	70	38	19.III	211	905	1270	1088	795	485	315	199	160	131
1.IV	88	301	384	463	317	195	160	86	51	51	19.IV	106	473	578	461	323	208	158	108	55	51
1.V	134	407	557	540	403	235	166	112	54	59	19.V	183	727	927	891	534	442	255	159	104	99
2.I	155	497	650	623	351	246	149	119	69	64	20.I	180	786	967	992	722	474	344	236	144	110
2.II	290	850	1143	1057	579	417	257	132	100	114	20.II	150	626	768	766	579	345	242	182	104	98
2.III	162	606	785	814	454	344	157	99	105	68	20.III	129	638	777	728	485	432	271	164	103	105
2.IV	229	591	837	736	472	279	196	102	92	54	20.IV	287	1340	1536	1553	1088	790	510	397	232	213
2.V	104	305	434	452	248	171	97	80	56	28	20.V	37	165	140	157	87	82	55	48	31	21
3.I	304	772	983	846	527	376	222	99	84	99	21.I	160	870	1146	1071	729	468	323	256	166	127
3.II	308	787	1148	1085	577	381	257	124	99	126	21.II	167	661	868	839	570	409	250	140	107	72
3.III	376	913	1091	1069	594	483	310	143	109	111	21.III	236	1144	1369	1380	924	603	381	331	177	155
3.IV	183	731	1021	850	532	354	242	136	83	101	21.IV	292	1044	1368	1357	876	597	350	232	151	120
3.V	140	418	572	485	334	236	156	69	31	72	21.V	116	495	588	553	446	242	144	119	46	59
4.I	127	577	742	622	395	285	184	116	58	75	22.I	173	772	1023	973	791	474	283	208	102	115
4.II	262	972	1297	1153	697	498	335	135	100	79	22.II	194	884	1206	1203	895	635	347	238	117	112
4.III	186	783	981	879	526	371	259	142	89	139	22.III	203	747	1051	991	641	499	261	208	82	69
4.IV	191	729	969	837	511	344	279	164	75	110	22.IV	135	723	1019	958	628	439	292	293	115	84
4.V	179	758	915	915	516	363	279	154	91	87	22.V	158	616	934	847	645	393	249	205	93	75
5.I	78	364	477	432	254	164	155	82	51	43	23.I	365	1481	1793	1744	1104	780	472	351	172	191
5.II	133	496	573	602	318	236	139	65	44	53	23.II	156	646	733	648	458	322	186	139	57	71
5.III	143	468	623	614	345	281	118	59	33	36	23.III	242	1152	1440	1418	828	560	411	278	148	174
5.IV	220	657	920	753	451	327	201	89	59	60	23.IV	272	1113	1256	1419	921	667	373	278	145	176
5.V	168	637	842	675	432	298	209	105	61	72	23.V	128	661	740	727	573	369	253	188	97	76
6.I	223	599	742	684	398	234	185	107	76	56	24.I	206	881	1134	1037	772	497	306	195	153	121
6.II	344	976	1338	1124	695	404	285	170	103	97	24.II	253	946	1292	1143	735	475	352	222	163	139
6.III	232	754	1107	851	516	353	247	141	108	72	24.III	245	939	1124	1149	730	452	322	200	137	103
6.IV	176	539	719	674	375	246	192	111	51	59	24.IV	118	459	651	531	402	222	171	130	69	56
6.V	244	745	1074	1074	576	368	284	163	91	68	24.V	230	803	1138	1027	689	440	245	185	110	111
7.I	203	627	795	639	527	349	173	127	82	85	25.I	245	1000	1357	1309	980	623	397	258	116	119
7.II	76	356	442	389	292	168	105	73	60	35	25.II	184	909	1330	1080	735	565	352	219	115	140
7.III	114	230	290	315	167	145	78	35	24	52	25.III	207	847	1166	1067	740	445	303	179	86	118
7.IV	322	957	1333	1142	819	533	321	220	123	119	25.IV	207	847	1166	1067	740	445	303	179	86	118
7.V	370	1361	1795	1714	1182	786	438	296	192	193	25.V	190	853	1131	1096	720	457	298	194	104	111

Table B.16: Word length counts for all the words in Willian Shakespeare plays (1/2)

Play.Act	1	2	3	4	5	6	7	8	9	10+
8.I	128	493	526	542	381	285	160	87	87	56
8.II	148	506	773	663	502	318	189	110	63	54
8.III	254	812	1145	1161	716	482	259	171	99	91
8.IV	108	350	485	407	285	210	112	62	44	43
8.V	128	474	653	674	448	309	180	90	64	61
9.I	245	738	903	746	514	364	224	166	90	80
9.II	283	930	1214	1055	707	489	303	229	122	87
9.III	275	946	1197	1100	696	495	296	199	96	87
9.IV	191	657	891	790	514	295	203	154	79	59
9.V	136	449	603	494	355	221	141	89	33	24
10.I	219	822	1095	1000	576	381	274	178	105	70
10.II	216	713	968	909	557	366	259	142	77	77
10.III	312	1000	1318	1106	714	464	303	152	110	115
10.IV	189	491	796	686	380	258	156	97	58	61
10.V	231	651	855	696	401	295	213	125	61	50
11.I	353	1176	1514	1246	795	629	363	229	167	118
11.II	198	573	811	677	410	293	157	115	95	43
11.III	134	452	620	506	392	226	153	85	67	50
11.IV	289	944	1371	1105	736	481	300	154	134	87
11.V	117	419	626	503	384	250	154	82	68	34
12.I	207	744	974	865	614	427	234	154	110	124
12.II	319	977	1280	1237	732	517	299	216	127	107
12.III	203	676	856	771	544	318	210	158	83	73
12.IV	281	919	1197	1000	645	439	254	209	98	143
12.V	196	599	859	791	470	293	176	124	61	56
13.I	268	823	1029	975	587	391	232	130	85	86
13.II	262	862	1033	948	616	371	232	154	80	88
13.III	280	988	1322	1073	672	506	262	175	108	91
13.IV	96	315	420	385	235	124	91	46	26	35
13.V	166	563	719	664	399	270	168	105	69	55
14.I	171	769	871	854	477	355	264	163	114	119
14.II	162	700	864	814	509	362	190	124	99	84
14.III	149	560	731	638	445	287	179	123	87	73
14.IV	436	1447	1950	1769	1100	701	529	307	181	184
14.V	178	788	1086	951	616	442	282	181	129	81
15.I	226	1143	1404	1470	1038	698	422	222	158	190
15.II	29	105	117	134	96	56	44	12	14	11
15.III	193	725	832	913	640	441	229	169	100	91
15.IV	215	831	934	962	635	448	266	124	100	91
15.V	125	757	868	837	611	366	269	127	105	113
16.I	174	885	952	895	723	449	318	222	140	114
16.II	153	920	1109	1008	754	441	308	215	129	128
16.III	121	750	935	957	672	490	299	157	96	130
16.IV	113	462	478	594	340	252	162	91	57	52
16.V	182	745	900	1006	713	436	242	123	64	100
17.I	199	855	1035	997	645	429	270	163	130	146
17.II	437	1129	1489	1512	965	601	406	198	140	125
17.III	272	883	1140	976	708	461	301	171	104	160
17.IV	113	532	697	527	420	267	179	103	77	82
17.V	211	785	944	866	592	367	232	128	100	85
18.I	216	842	1106	1003	633	497	254	167	110	140
18.II	321	996	1396	1260	806	509	379	202	129	155
18.III	183	473	823	727	401	299	170	101	53	98
18.IV	219	1120	1438	1403	968	659	393	271	192	204
18.V	200	625	859	800	520	375	252	118	64	83

Play.Act.	1	2	3	4	5	6	7	8	9	10+
26.I	115	655	820	784	545	389	262	183	108	114
26.II	156	632	908	926	641	419	225	189	100	77
26.III	115	504	692	734	463	322	183	97	52	51
26.IV	159	688	1074	919	575	426	275	146	87	75
26.V	194	686	1073	990	577	454	272	200	141	71
27.I	290	910	1222	1199	863	509	287	185	98	89
27.II	288	918	1151	1239	802	451	251	167	104	76
27.III	309	1055	1350	1431	956	575	338	217	112	96
27.IV	145	576	675	737	462	298	169	95	49	50
27.V	176	604	690	777	519	315	176	112	71	68
28.I	207	711	816	903	579	376	247	165	79	92
28.II	97	358	396	453	298	187	125	65	45	41
28.III	160	732	969	870	532	393	236	145	96	86
28.IV	183	634	1039	1146	746	504	309	202	97	111
28.V	91	426	608	593	415	264	169	102	50	73
29.I	208	794	1060	899	617	434	246	142	76	82
29.II	187	771	912	832	611	475	190	141	82	77
29.III	203	769	1041	907	697	538	281	148	100	61
29.IV	149	576	656	684	420	311	198	112	41	55
29.V	116	483	578	650	393	286	164	121	56	31
30.I	92	600	797	754	569	390	264	162	78	96
30.II	103	387	599	521	427	285	164	102	46	67
30.III	129	580	845	741	550	364	246	116	75	60
30.IV	166	560	802	759	534	398	232	171	67	77
30.V	107	480	606	628	486	292	176	126	50	49
31.I	201	1224	1361	1269	814	642	333	245	134	135
31.II	243	1019	1244	1140	685	489	318	203	124	145
31.III	272	1243	1521	1411	909	635	416	287	137	142
31.IV	142	778	995	948	570	385	283	137	78	91
31.V	223	884	1141	1083	648	431	295	181	117	107
32.I	305	1163	1410	1434	866	614	415	275	153	145
32.II	217	744	1042	1037	682	479	302	175	110	110
32.III	164	598	902	894	653	348	202	150	85	101
32.IV	227	737	1015	949	671	450	260	156	81	81
32.V	124	506	690	686	449	266	180	117	52	54
33.I	262	1038	1159	1028	723	490	347	214	156	130
33.II	243	959	1269	1077	696	500	336	239	143	153
33.III	325	1098	1281	1243	710	536	307	191	116	118
33.IV	260	918	1122	1035	628	436	264	183	89	76
33.V	235	693	850	841	534	380	207	134	54	58
34.I	146	671	887	801	610	421	252	178	95	80
34.II	232	974	1221	1251	816	586	391	205	113	87
34.III	188	999	1278	1264	799	585	442	241	139	114
34.IV	174	790	945	1012	790	444	291	170	103	73
34.V	150	656	701	712	541	358	206	141	84	61
35.I	317	1037	1225	1187	787	520	347	235	134	132
35.II	178	631	829	749	523	324	219	144	69	74
35.III	242	1117	1263	1209	858	600	403	222	138	142
35.IV	167	752	995	890	629	422	283	147	98	84
35.V	284	1127	1525	1332	956	682	447	274	194	110

Table B.17: Word length counts for all the words in WILLIAN SHAKESPEARE PLAYS (2/2)

Play.Act	the	and	i	to	of	a	you	my	that	in	is	not	it	for	with	me	your	his	this	be
1.I	148	121	110	88	84	91	37	83	55	49	27	48	33	38	38	49	12	29	43	28
1.II	102	102	99	71	79	96	55	52	52	38	48	40	45	34	27	40	25	32	38	30
1.III	69	96	105	56	49	54	44	60	27	28	27	30	31	22	23	33	24	15	35	21
1.IV	47	68	51	41	30	29	29	43	20	15	22	11	13	15	33	12	17	6	31	24
1.V	78	97	83	45	48	43	45	54	35	31	27	25	12	20	18	21	20	16	38	24
2.I	85	73	95	79	44	58	60	40	36	39	29	33	39	42	29	36	37	26	10	24
2.II	102	110	175	135	62	103	121	104	73	51	68	64	52	60	48	68	51	41	30	27
2.III	74	87	96	86	54	64	34	66	70	40	56	41	35	52	31	47	35	12	23	44
2.IV	77	85	168	89	38	59	80	62	68	30	39	41	30	47	28	61	35	20	27	32
2.V	40	47	63	48	28	35	29	30	35	15	31	23	21	25	12	25	11	1	20	18
3.I	111	113	204	73	74	87	106	64	47	57	106	48	65	34	25	24	47	34	23	36
3.II	118	112	201	105	94	99	136	79	42	51	71	48	41	45	53	41	48	32	33	37
3.III	114	121	236	107	92	121	155	87	47	82	56	44	46	49	38	67	63	35	40	50
3.IV	138	105	106	85	84	74	85	62	43	51	69	39	31	30	46	30	38	28	22	44
3.V	83	88	71	49	51	66	55	27	26	36	24	25	19	12	17	23	20	9	12	24
4.I	115	90	77	88	82	50	63	34	40	56	28	40	41	32	40	29	30	21	17	23
4.II	153	118	152	153	88	105	148	57	96	74	57	70	92	53	42	41	93	36	37	56
4.III	153	109	83	139	87	101	88	40	56	66	65	34	65	41	42	22	46	38	47	44
4.IV	142	118	108	124	67	83	114	39	44	48	58	49	54	46	35	32	59	32	43	36
4.V	85	105	111	90	65	66	109	91	65	50	56	43	46	54	42	43	74	30	51	37
5.I	89	61	48	70	52	30	31	51	30	22	11	27	10	27	13	27	18	12	11	9
5.II	57	72	94	58	29	37	46	47	35	45	30	25	36	34	32	43	23	25	19	20
5.III	72	70	82	57	37	55	63	48	39	51	25	27	32	29	21	37	41	2	17	20
5.IV	119	108	136	91	41	83	104	47	61	55	50	44	42	48	33	79	25	29	27	19
5.V	100	132	120	74	64	46	56	83	48	49	26	45	38	31	40	67	19	35	43	16
6.I	101	85	120	68	60	97	83	57	47	57	59	33	52	32	34	22	30	27	27	26
6.II	171	148	199	117	98	132	124	79	69	81	74	61	69	55	45	52	41	44	23	41
6.III	141	127	117	104	72	105	104	37	56	53	67	54	65	51	45	23	37	27	24	46
6.IV	66	79	94	61	60	62	71	45	65	43	50	54	40	25	26	30	21	10	42	33
6.V	95	161	151	104	71	87	107	80	48	66	52	47	40	59	45	59	65	24	36	26
7.I	136	104	119	104	83	79	39	38	55	55	67	28	46	46	39	19	17	17	30	28
7.II	53	39	41	57	45	34	48	29	33	35	28	22	20	17	22	16	25	42	4	17
7.III	42	46	40	25	31	61	22	22	15	23	24	4	16	16	10	8	18	4	16	9
7.IV	266	133	124	121	113	158	75	63	81	90	84	63	70	74	29	31	30	36	30	32
7.V	329	235	179	156	165	166	157	91	99	121	98	96	87	84	63	58	83	58	75	50
8.I	94	86	69	81	48	50	54	49	34	42	35	12	24	23	27	22	35	12	11	15
8.II	139	131	93	46	57	49	50	36	33	50	26	47	24	33	47	45	24	18	24	22
8.III	115	166	154	107	65	78	118	49	64	62	46	65	29	38	55	76	39	31	47	34
8.IV	85	73	70	47	49	34	19	37	13	33	28	20	22	18	16	18	15	9	22	16
8.V	128	106	57	56	50	51	32	33	40	52	55	27	33	31	30	13	15	23	45	17
9.I	135	118	139	95	93	101	100	62	41	64	61	43	37	41	43	60	31	31	31	36
9.II	187	145	166	134	94	106	86	114	63	65	60	61	31	52	52	71	42	43	35	51
9.III	207	165	162	102	121	105	107	120	72	84	62	47	65	55	45	55	42	31	26	44
9.IV	176	105	116	95	85	66	82	49	46	43	49	41	61	43	25	40	41	35	39	30
9.V	111	67	78	40	51	58	63	36	33	37	33	29	46	31	26	26	22	15	17	23
10.I	135	110	154	102	82	58	114	84	75	57	60	59	36	49	54	52	55	45	33	35
10.II	132	167	129	81	82	77	67	61	70	65	47	51	38	39	36	29	35	32	37	31
10.III	180	150	153	86	136	143	119	52	91	97	100	81	63	48	53	51	36	38	25	45
10.IV	88	104	100	67	61	81	89	30	51	42	30	24	39	31	20	36	32	29	32	19
10.V	120	123	133	116	57	85	117	36	52	51	49	31	33	39	30	27	37	20	35	29
11.I	146	235	203	165	70	145	144	89	68	91	67	67	62	84	53	75	68	42	43	63
11.II	62	113	117	72	43	80	100	66	44	45	41	40	16	38	32	52	39	11	19	35
11.III	78	77	76	75	37	55	60	46	29	30	26	40	24	23	32	20	21	15	15	35
11.IV	155	216	181	134	73	108	138	89	60	54	79	55	79	55	55	61	40	16	43	50
11.V	58	88	73	67	23	44	65	49	32	20	48	29	15	26	16	27	27	17	17	15
12.I	123	122	121	90	85	82	75	75	67	78	50	41	60	38	24	26	55	52	24	46
12.II	146	172	191	150	100	113	118	129	69	65	83	73	62	52	33	47	58	36	31	37
12.III	134	96	123	79	87	76	84	55	68	55	56	42	45	32	35	23	34	48	32	25
12.IV	184	111	150	125	131	126	103	70	67	63	56	61	59	46	39	31	47	75	40	45
12.V	83	105	121	71	56	67	100	64	52	39	34	34	64	32	24	43	39	22	47	12

Table B.18: Most frequent function word counts in Willian Shakespeare plays (1/3)

Play.Act	the	and	i	to	of	a	you	my	that	in	is	not	it	for	with	me	your	his	this	be
13.I	113	109	142	92	94	110	112	83	70	81	64	55	48	42	38	32	53	21	26	41
13.II	118	107	136	93	112	106	67	95	65	55	56	56	60	43	39	54	31	14	40	42
13.III	158	142	171	109	123	102	157	81	60	58	75	63	49	61	57	58	57	44	42	50
13.IV	41	49	73	37	28	21	31	27	29	24	22	23	11	15	15	25	8	7	28	17
13.V	86	99	90	61	55	62	77	70	50	40	26	32	31	26	19	51	32	22	36	27
14.I	89	102	94	110	59	73	80	106	59	42	47	53	39	27	41	46	38	32	35	49
14.II	108	93	97	96	64	63	83	63	52	32	36	56	51	40	31	40	49	29	39	41
14.III	135	91	90	91	69	52	31	49	29	25	34	33	45	41	23	26	23	12	27	30
14.IV	260	227	251	209	180	168	166	134	110	74	81	104	84	79	62	68	101	52	74	77
14.V	162	112	112	105	99	60	85	66	68	46	45	59	56	35	37	38	66	46	28	40
15.I	217	220	115	169	206	101	62	97	120	114	59	52	45	64	56	28	79	62	109	41
15.II	10	23	18	21	13	10	4	7	10	1	9	9	7	6	10	9	0	3	5	11
15.III	115	134	87	95	103	77	35	66	73	56	42	52	30	25	39	36	20	29	49	42
15.IV	159	143	112	105	118	92	80	62	50	54	55	44	52	34	58	56	45	34	52	40
15.V	185	168	67	114	113	55	43	61	52	48	35	32	37	25	41	35	40	29	65	36
16.I	136	164	96	144	121	73	27	129	50	81	37	33	27	41	43	33	24	62	35	36
16.II	152	164	101	137	113	50	44	104	62	73	84	59	47	55	61	37	28	68	35	38
16.III	165	162	62	108	127	59	37	96	55	48	50	33	26	39	60	24	42	47	32	37
16.IV	74	76	74	69	69	37	33	64	50	33	21	21	27	16	24	24	25	20	32	17
16.V	146	134	117	93	74	58	19	93	63	51	56	51	40	44	47	48	10	39	37	34
17.I	181	194	100	103	137	91	46	58	53	65	25	41	43	56	49	27	30	35	43	50
17.II	217	200	216	120	156	200	101	100	70	101	72	83	57	64	53	60	31	37	48	59
17.III	171	204	146	94	132	118	82	83	51	101	42	45	39	32	36	52	28	30	33	29
17.IV	130	113	55	74	93	58	31	40	30	43	32	36	20	22	26	25	15	34	20	18
17.V	155	139	110	84	121	91	46	62	48	58	41	49	38	35	38	35	34	36	39	20
18.I	218	161	110	107	129	102	64	70	67	74	65	52	49	40	59	28	59	58	35	22
18.II	205	171	173	136	104	135	145	80	52	74	69	67	57	57	56	70	51	40	27	44
18.III	123	115	85	58	75	90	70	35	33	34	48	18	38	25	32	27	21	18	18	18
18.IV	259	260	114	160	193	90	95	139	87	85	73	50	89	42	77	46	64	60	55	36
18.V	116	121	122	83	78	68	96	58	46	45	36	35	28	30	38	42	47	22	20	44
19.I	313	303	87	176	242	112	101	71	113	144	70	51	66	75	88	23	68	88	51	44
19.II	252	226	103	115	146	120	81	67	56	71	106	50	54	64	54	26	54	56	16	37
19.III	238	183	92	103	128	96	41	64	59	84	56	48	51	50	50	33	16	71	31	44
19.IV	96	103	66	44	75	37	43	26	30	52	62	17	50	22	15	16	40	47	26	15
19.V	117	160	109	94	99	71	99	62	50	70	55	34	36	41	34	54	58	22	27	21
20.I	193	129	112	116	107	60	32	59	35	72	47	39	23	31	60	43	21	46	41	48
20.II	127	132	92	80	97	56	32	84	51	55	38	35	16	42	40	27	30	46	44	31
20.III	119	137	65	75	94	64	35	46	55	59	28	34	25	36	36	21	32	38	33	38
20.IV	236	288	154	189	222	118	75	133	100	99	73	63	46	64	89	57	71	67	66	100
20.V	16	32	17	21	28	20	6	14	10	12	18	7	2	11	15	4	7	7	3	13
21.I	255	206	100	101	165	53	35	95	50	82	36	31	28	58	50	43	43	55	45	64
21.II	152	158	99	100	90	61	32	93	34	54	29	36	30	32	37	31	45	37	45	36
21.III	217	197	143	145	106	90	54	127	100	84	82	55	48	71	74	60	40	90	39	63
21.IV	235	262	153	144	119	130	55	86	73	81	67	56	51	75	67	53	42	47	42	66
21.V	81	91	64	95	56	47	19	40	29	36	31	26	21	25	30	19	9	26	13	19
22.I	172	196	125	106	104	47	37	88	61	50	41	50	40	38	52	51	24	56	36	49
22.II	198	193	103	126	106	76	22	88	82	68	44	30	39	72	66	34	33	75	49	40
22.III	118	178	127	145	61	74	52	104	58	43	44	41	30	58	54	39	41	45	38	47
22.IV	149	183	97	115	83	36	56	88	65	67	34	40	21	55	48	41	37	53	35	40
22.V	163	170	89	95	51	61	21	71	60	46	33	37	16	55	47	24	14	48	34	26
23.I	234	251	224	219	170	108	149	158	181	109	64	86	80	99	76	106	64	76	59	64
23.II	102	120	96	84	68	52	52	87	32	40	42	31	28	35	31	43	33	30	27	28
23.III	248	194	156	192	143	75	117	145	78	90	48	63	52	62	63	54	101	85	55	31
23.IV	207	195	167	174	150	83	44	151	99	84	73	52	34	53	64	73	62	27	32	70
23.V	158	133	69	101	104	53	23	83	41	76	41	24	16	32	33	40	21	30	23	27
24.I	192	141	89	118	111	108	75	63	68	67	48	44	50	40	31	29	55	67	64	30
24.II	193	172	144	139	133	100	105	94	78	64	39	58	53	57	44	62	82	44	69	52
24.III	189	146	140	129	114	94	101	120	71	64	40	31	42	45	34	63	72	82	53	39
24.IV	129	84	55	59	77	49	33	24	44	40	27	20	12	15	31	37	14	39	12	14
24.V	149	158	130	111	105	97	100	82	55	53	35	38	28	49	43	45	61	34	55	43

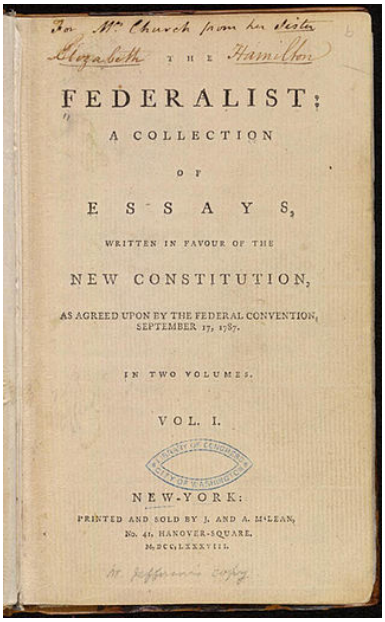
Table B.19: Most frequent function word counts in Willian Shakespeare plays (2/3)

Play.Act	the	and	i	to	of	a	you	my	that	in	is	not	it	for	with	me	your	his	this	be
25.I	228	152	137	142	95	103	147	56	77	80	54	63	57	49	74	47	69	42	32	39
25.II	204	160	100	152	113	84	150	42	71	76	32	56	46	60	46	18	84	64	19	41
25.III	157	158	115	113	98	84	108	55	51	59	50	51	40	40	48	38	40	44	34	34
25.IV	157	158	115	113	98	84	108	55	51	59	50	51	40	40	48	38	40	44	34	34
25.V	167	130	97	139	92	84	107	69	52	53	39	56	39	45	78	30	76	74	48	31
26.I	119	199	67	102	83	42	50	87	66	90	21	32	15	38	55	31	35	39	37	35
26.II	137	149	82	100	54	72	40	53	53	42	42	47	36	36	48	41	34	21	59	35
26.III	75	94	62	69	38	44	15	68	48	41	25	30	26	41	46	35	8	17	22	21
26.IV	182	155	87	101	72	64	66	63	49	57	39	36	44	50	57	36	24	47	45	29
26.V	126	210	118	105	69	72	55	66	61	48	33	25	31	53	50	48	29	35	42	30
27.I	187	145	153	108	126	124	85	85	70	78	75	52	46	43	64	43	26	48	46	39
27.II	145	150	132	133	74	129	45	77	91	68	87	59	56	54	44	51	16	33	35	44
27.III	148	172	133	133	84	128	82	97	100	77	93	83	64	61	60	71	40	24	42	62
27.IV	82	92	81	89	34	37	61	51	36	42	41	35	34	36	38	46	18	12	40	35
27.V	95	131	98	85	62	53	16	64	46	51	44	29	30	28	40	53	9	20	54	31
28.I	88	86	116	117	70	82	77	73	48	45	37	44	40	28	34	41	30	36	32	30
28.II	36	61	58	50	30	37	50	48	22	24	21	24	22	7	22	27	25	17	13	13
28.III	87	118	99	106	67	55	65	100	50	53	44	47	45	45	31	33	41	75	26	31
28.IV	164	150	95	84	85	76	40	32	71	53	38	50	40	36	41	34	25	36	33	35
28.V	74	92	45	59	67	44	43	29	33	49	27	29	31	26	36	12	21	33	17	17
29.I	144	152	130	105	83	73	113	30	68	59	56	48	56	39	39	45	29	38	39	39
29.II	143	142	113	125	82	57	68	45	69	47	60	68	46	40	28	42	36	22	23	27
29.III	144	152	126	99	94	54	102	33	69	52	52	56	37	46	41	43	32	55	37	39
29.IV	63	97	96	60	49	43	77	54	38	36	39	46	32	31	27	41	40	25	26	27
29.V	83	84	65	57	44	31	30	47	41	31	43	38	27	20	14	16	12	17	41	15
30.I	153	134	50	95	83	39	43	38	60	50	38	30	34	17	31	32	34	45	20	32
30.II	114	88	57	58	49	42	28	21	30	35	28	34	36	14	23	19	15	13	21	11
30.III	142	116	71	97	70	51	61	47	52	49	41	30	31	36	40	18	33	40	18	31
30.IV	131	117	91	64	75	63	52	57	48	39	37	40	34	30	22	24	30	30	28	36
30.V	105	90	62	69	59	44	19	40	37	26	36	31	26	12	37	20	14	18	17	27
31.I	208	206	105	179	157	90	107	125	73	105	55	79	130	44	60	49	55	55	63	38
31.II	197	178	132	129	117	107	119	119	70	78	60	51	69	49	49	44	39	64	45	30
31.III	252	212	135	165	164	122	160	120	89	82	73	69	78	57	68	57	72	58	68	52
31.IV	125	136	67	99	71	72	86	62	57	58	56	46	56	46	39	26	57	52	37	34
31.V	209	128	107	111	98	106	50	73	72	59	70	54	80	37	36	52	30	56	62	37
32.I	163	178	186	141	149	107	143	158	95	73	43	66	74	47	52	52	70	47	58	56
32.II	142	118	133	102	84	71	104	94	59	47	34	68	33	28	48	50	48	37	45	29
32.III	161	118	71	90	67	77	62	68	59	54	41	30	21	19	32	32	30	41	44	19
32.IV	160	117	139	80	67	70	69	79	63	56	43	56	38	26	32	59	48	29	35	28
32.V	101	104	76	57	57	35	47	52	37	31	38	31	23	21	23	24	30	19	40	11
33.I	170	182	170	141	133	90	88	116	63	76	65	48	57	56	69	51	60	31	45	48
33.II	204	189	140	154	108	102	77	55	79	76	66	54	56	59	39	45	41	42	58	40
33.III	119	157	220	142	89	94	119	121	96	70	55	101	80	55	44	72	55	33	46	67
33.IV	112	147	172	89	58	83	108	81	65	56	58	72	79	36	39	65	48	36	34	35
33.V	88	92	155	56	52	58	84	75	67	41	44	43	47	28	24	45	21	23	40	23
34.I	144	99	85	95	77	58	62	41	44	55	46	45	38	19	44	29	22	40	26	30
34.II	179	128	149	136	89	80	110	67	62	58	66	72	64	51	52	59	45	35	32	46
34.III	199	172	111	122	120	76	85	86	76	69	57	61	47	52	46	46	38	85	32	41
34.IV	134	136	95	96	75	68	41	87	53	29	56	43	42	36	45	69	21	28	44	28
34.V	107	87	94	78	82	48	59	51	54	48	45	37	32	27	20	40	31	30	33	34
35.I	175	144	188	141	114	115	120	102	79	66	67	59	60	50	51	38	90	56	46	57
35.II	113	97	117	73	83	57	75	34	49	28	45	51	61	27	28	28	54	22	33	32
35.III	175	156	141	142	121	100	69	106	99	82	79	57	63	55	41	54	30	33	41	66
35.IV	142	112	106	92	67	59	42	61	47	49	40	62	42	37	35	34	16	44	37	38
35.V	220	183	156	160	130	122	103	103	90	76	53	55	56	65	57	63	50	57	62	54

Table B.20: Most frequent function word counts in Willian Shakespeare plays (3/3)

B.2.4 Federalist Papers

The Federalist Papers, first known as *The Federalist*, is a collection of 85 articles and essays written under the pseudonym by Alexander Hamilton, James Madison, and John Jay. 75 of them were published serially in The Independent Journal and The New York Packet and they were later compiled and published together with eight additional ones in two volumes in 1788 by J. and A. McLean. The authors of The Federalist Papers foremost wished to influence the vote in favor of ratifying the United States Constitution.



Author	A.Hamilton, J. Madison, and J. Jay
Country	United States
Language	English
Literary Genre	Essay
Editor	J. and A. McLean
Publication Date	1788

Paper	1	2	3	4	5	6	7	8	9	10+	Paper	1	2	3	4	5	6	7	8	9	10+
1	39	355	310	222	159	99	106	85	61	147	44	47	589	583	360	351	224	182	156	148	250
2	33	311	307	287	148	132	127	103	71	142	45	23	427	452	272	192	177	134	136	100	202
3	16	304	257	213	130	139	113	87	69	109	46	46	523	538	355	265	216	175	121	106	265
4	16	322	341	236	180	116	143	88	61	123	47	41	573	605	313	234	152	186	142	182	307
5	10	238	233	220	160	103	114	90	74	94	48	44	390	351	235	180	122	113	91	120	208
6	51	378	377	264	184	144	151	121	89	179	49	35	372	334	164	177	117	105	89	90	160
7	47	485	390	282	262	181	142	116	110	235	50	19	226	200	145	104	81	103	68	46	107
8	47	422	350	263	195	167	134	136	102	170	51	48	431	349	219	169	164	145	107	88	191
9	53	433	348	263	198	130	134	119	102	194	52	40	400	337	256	211	136	121	78	91	171
10	81	619	584	376	287	230	243	199	147	228	53	53	451	427	300	178	148	174	127	117	186
11	70	529	454	315	273	175	195	139	112	226	54	39	464	369	288	192	151	128	125	76	163
12	48	461	369	288	225	184	163	134	89	180	55	59	429	375	275	238	169	138	115	79	157
13	26	202	162	146	100	77	64	56	37	87	56	48	299	305	230	159	115	98	68	96	146
14	39	455	416	286	242	175	132	124	93	174	57	41	499	458	285	236	164	125	120	86	189
15	73	707	552	399	316	214	220	169	159	264	58	56	432	391	286	206	160	134	125	89	201
16	38	482	364	257	214	153	136	99	106	183	59	52	443	351	259	218	112	107	112	97	157
17	32	300	305	187	163	123	111	110	77	154	60	59	559	397	272	224	146	139	129	96	212
18	38	356	475	258	164	196	178	123	112	185	61	40	352	277	230	150	116	80	85	63	122
19	27	368	442	240	175	161	182	141	104	178	62	82	524	444	305	216	209	145	136	93	227
20	29	271	310	168	156	102	128	101	84	161	63	69	647	591	383	281	263	214	169	161	255
21	66	458	352	238	197	155	134	105	89	197	64	32	488	476	385	192	163	151	143	112	163
22	97	776	627	445	357	266	233	200	163	308	65	45	466	381	257	223	138	133	124	86	161
23	28	393	375	212	173	148	121	101	90	154	66	54	516	428	328	189	182	110	100	110	203
24	50	433	326	216	178	118	137	109	94	156	67	29	357	336	191	143	148	101	92	103	135
25	45	465	348	256	206	165	153	115	89	139	68	31	348	312	196	132	125	93	67	73	119
26	55	575	437	307	222	206	151	114	114	194	69	64	616	585	334	244	205	180	174	136	201
27	36	314	274	208	104	103	103	72	64	139	70	75	659	614	395	288	211	212	150	178	277
28	33	341	320	188	165	117	108	88	76	152	71	46	392	343	217	146	128	113	85	95	136
29	54	549	428	281	199	147	164	122	102	183	72	58	461	408	257	188	136	140	95	114	177
30	39	426	389	246	205	137	144	123	93	160	73	65	517	455	346	223	166	131	121	126	193
31	26	404	322	231	184	129	100	103	82	147	74	30	234	187	115	109	60	61	69	57	73
32	46	320	273	173	189	94	108	81	64	134	75	57	437	362	244	195	168	133	102	104	136
33	49	357	338	241	179	116	119	74	82	129	76	55	536	437	326	242	170	141	94	101	194
34	47	496	414	288	268	151	157	115	103	170	77	73	457	355	260	190	147	103	81	113	190
35	50	493	414	308	243	166	146	141	112	175	78	64	690	588	408	274	210	167	155	162	293
36	55	646	506	383	272	197	165	159	121	223	79	25	220	211	141	98	75	53	57	50	89
37	57	549	517	386	241	205	171	142	144	297	80	43	532	478	266	295	215	165	128	120	208
38	97	700	647	478	293	278	244	159	142	277	81	95	907	730	479	380	280	228	189	189	304
39	53	617	537	291	200	209	158	159	127	253	82	32	334	295	174	185	133	99	80	78	122
40	57	651	609	396	275	229	176	151	132	333	83	157	1353	993	804	628	425	347	263	248	487
41	86	726	712	447	370	264	303	202	174	256	84	85	879	761	537	404	314	270	189	148	325
42	52	586	550	321	272	218	223	178	120	254	85	99	600	534	311	234	205	167	134	125	248
43	95	749	666	387	315	270	249	193	182	316											

Table B.21: Word length counts for all the words in FEDERALIST PAPERS

Paper	on	would	upon	there	by	to	and	the	these	in	at	latter	several	I	if	might	any	kind	had	between	those	an	he	this	very	against	no	were	into	same
1	9	2	6	2	14	70	40	126	3	26	8	1	1	14	4	2	6	0	1	0	9	11	0	14	0	1	3	1	2	1
2	8	5	1	0	10	52	83	105	4	34	10	1	0	4	3	0	1	2	4	0	2	1	0	14	7	1	1	7	6	6
3	6	2	0	1	18	55	60	91	3	25	1	1	2	3	7	0	5	1	0	0	6	3	2	6	0	3	2	1	0	2
4	11	17	0	3	14	50	90	84	6	24	2	0	2	0	13	2	5	0	3	2	4	3	1	1	0	6	1	1	10	0
5	5	37	0	0	10	44	72	64	6	28	4	0	0	1	3	6	3	1	0	3	9	4	0	6	1	5	2	6	7	5
6	2	6	4	8	10	56	72	154	6	55	6	1	0	2	5	1	0	1	4	7	7	11	4	11	1	4	0	5	7	5
7	12	51	11	9	28	80	49	201	10	40	11	1	2	0	12	4	7	1	7	8	8	15	0	22	0	1	3	9	2	3
8	9	27	3	2	11	78	52	155	3	39	10	1	1	1	7	1	1	0	3	2	6	13	0	16	2	5	3	0	5	3
9	9	8	4	3	13	70	45	168	5	37	10	1	2	6	7	2	4	4	5	3	6	13	9	15	3	4	2	11	5	2
10	18	6	0	6	39	99	121	259	8	63	8	5	0	3	6	0	4	0	1	3	4	14	4	11	0	5	5	0	9	12
11	5	50	6	8	20	82	70	186	7	66	11	1	2	1	6	7	4	3	1	4	7	15	0	24	1	1	5	2	4	3
12	12	22	7	9	15	80	62	174	8	54	13	1	1	1	6	2	7	2	3	5	3	11	1	17	2	3	2	0	7	2
13	3	14	2	9	5	42	17	72	0	14	1	0	2	0	6	0	3	0	0	1	0	3	0	5	1	1	4	0	7	3
14	17	5	0	0	18	71	60	200	2	40	7	2	1	2	6	1	3	0	3	4	13	9	0	13	0	4	10	3	2	3
15	10	13	10	18	32	116	74	251	6	73	24	0	0	4	7	0	3	6	2	2	10	18	0	24	1	4	6	8	4	6
16	4	36	6	4	13	88	42	191	2	39	11	1	1	1	14	2	11	5	5	3	10	13	0	18	2	6	6	7	3	4
17	2	12	6	4	9	57	52	160	0	29	7	2	1	2	2	3	2	1	3	4	9	9	1	11	0	2	0	4	3	3
18	15	6	1	3	33	53	79	235	4	41	4	3	2	1	0	1	2	0	23	2	5	5	3	16	3	4	2	16	5	6
19	17	4	0	1	22	57	82	203	4	41	5	0	0	0	4	1	6	0	11	4	0	9	6	12	1	6	6	5	4	0
20	7	1	1	0	19	41	54	135	7	39	8	0	3	1	1	0	0	0	3	1	1	5	8	8	2	0	2	0	3	3
21	6	12	6	8	22	54	49	182	3	48	4	0	1	2	11	2	6	8	1	4	9	10	0	17	0	6	12	2	0	2
22	8	20	13	14	30	143	80	288	3	86	12	0	5	1	9	10	7	5	3	5	13	19	2	30	1	1	6	9	4	6
23	2	4	7	4	11	96	56	185	6	26	3	0	0	4	6	0	9	0	1	2	4	11	0	14	1	1	2	1	2	3
24	11	25	7	6	14	84	54	133	9	50	7	1	1	4	13	1	6	1	5	4	1	9	18	18	0	6	3	0	1	0
25	11	21	2	2	22	89	46	173	0	40	11	1	0	1	5	8	4	4	4	2	6	7	0	20	1	4	2	1	1	4
26	7	16	6	8	21	94	49	199	3	64	12	2	0	2	16	1	10	1	7	5	5	22	2	15	3	6	4	7	8	1
27	4	3	4	8	14	59	33	144	5	28	4	1	2	4	5	2	8	1	0	2	4	5	0	8	2	0	3	0	3	2
28	1	12	3	7	7	65	35	164	3	40	7	0	0	0	10	3	2	1	3	0	4	13	0	7	0	8	6	2	1	6
29	3	19	10	13	10	111	59	220	0	41	11	2	0	4	15	1	6	1	2	1	4	14	1	14	3	3	5	6	4	5
30	8	22	13	5	14	75	45	162	4	49	6	1	0	1	4	5	8	3	0	3	5	7	2	17	2	0	5	1	6	2
31	5	5	13	6	9	81	44	160	4	49	5	4	0	2	3	5	5	2	0	1	9	12	0	13	0	5	3	0	5	4
32	14	23	2	8	10	46	41	148	2	40	2	1	0	10	6	9	6	2	2	1	1	15	0	16	1	1	7	2	1	4
33	5	15	9	2	14	66	44	156	6	33	2	1	0	5	8	4	7	1	3	0	6	8	0	18	3	5	0	4	4	6
34	11	20	10	9	4	106	48	184	6	55	6	2	2	2	12	4	8	0	2	6	10	13	0	14	2	6	8	1	2	1
35	8	16	9	5	10	100	65	186	9	47	4	0	1	5	7	4	8	0	2	7	9	13	2	14	1	1	6	4	5	4
36	6	5	6	18	21	119	66	248	1	65	4	2	3	11	8	5	10	5	3	5	9	15	0	23	1	2	5	1	4	2
37	19	7	1	2	30	84	101	228	14	62	2	3	3	1	2	1	3	0	3	5	4	10	1	17	6	1	5	3	5	4
38	15	15	4	3	37	116	94	269	9	62	6	5	2	5	6	6	9	0	6	0	3	18	6	24	3	13	17	10	7	5
39	25	8	0	0	33	90	64	298	6	73	7	3	6	0	5	2	4	0	0	2	1	6	1	21	1	1	7	6	3	6
40	13	6	0	4	55	119	99	292	22	68	7	3	7	2	8	1	2	0	17	2	7	12	1	18	2	5	9	19	7	3
41	27	9	0	1	32	118	88	334	15	63	5	3	3	5	13	3	6	1	7	1	6	20	1	28	5	14	9	6	7	9
42	22	15	2	3	31	92	96	260	6	62	3	6	6	2	4	5	11	0	4	3	6	16	0	18	7	6	8	4	8	1
43	33	12	0	2	47	111	79	354	2	55	7	3	0	2	6	5	8	0	3	5	5	12	1	22	2	15	14	1	7	8

Table B.22: Function words counts in FEDERALIST PAPERS for the thirty words with the largest T_i (1/2)

Paper	on	would	upon	there	by	to	and	the	these	in	at	latter	several	I	if	might	any	kind	had	between	those	an	he	this	very	against	no	were	into	same
44	28	29	0	4	28	80	88	313	12	66	4	3	6	4	8	12	15	0	10	2	4	10	0	25	2	6	18	2	11	10
45	10	11	0	3	11	65	63	276	4	40	4	9	4	3	7	2	4	0	4	2	8	7	0	5	8	5	6	6	3	2
46	32	29	0	0	21	85	73	301	13	46	6	4	2	4	5	0	6	0	3	0	6	9	0	9	4	4	4	3	4	5
47	20	4	0	3	42	64	86	325	6	62	7	4	7	6	5	1	5	0	3	0	1	10	10	24	3	1	15	7	1	14
48	16	3	0	2	28	53	51	167	10	45	4	1	5	8	2	3	1	0	10	1	4	12	0	14	2	6	6	4	3	7
49	16	22	0	2	15	57	42	176	1	34	2	1	4	2	3	6	3	0	4	1	1	11	3	5	2	9	4	2	1	4
50	11	11	1	0	11	27	33	99	3	28	9	0	4	5	4	3	5	0	9	0	1	3	0	7	3	1	0	1	2	9
51	21	9	0	4	23	49	40	200	4	50	3	4	6	2	7	4	1	0	1	2	3	5	0	13	4	8	4	4	5	5
52	19	8	0	0	22	71	37	184	7	33	7	1	2	5	5	3	7	0	2	2	4	3	1	15	6	0	4	10	1	5
53	8	6	0	2	31	72	62	191	10	45	2	0	4	3	7	3	6	0	0	7	5	6	3	15	9	2	10	4	2	2
54	19	6	2	1	26	60	38	202	3	65	2	2	3	4	5	4	7	0	3	0	6	9	2	21	6	2	8	6	5	7
55	9	10	0	5	14	77	48	180	3	30	17	3	2	11	5	1	8	0	3	5	0	4	0	12	7	4	4	6	0	3
56	11	4	0	3	10	38	53	135	6	31	5	1	1	0	0	4	5	1	1	0	4	3	0	13	10	2	2	2	4	4
57	19	6	0	4	25	73	54	214	8	40	6	0	0	4	6	1	5	0	1	4	5	6	3	13	7	5	4	2	2	4
58	18	12	0	2	22	60	47	211	6	58	5	6	2	5	6	8	5	0	2	3	4	9	0	14	2	7	5	2	2	3
59	6	16	3	7	17	72	34	176	3	62	8	1	0	3	6	6	11	1	2	2	5	24	2	18	1	3	8	2	5	0
60	6	28	8	8	21	86	36	221	4	79	9	3	3	6	6	6	12	1	0	5	8	12	0	16	1	1	4	2	5	5
61	6	17	3	5	5	60	25	149	1	47	12	1	4	5	7	4	5	1	3	1	4	5	0	14	0	3	5	0	3	8
62	19	5	0	0	28	79	69	190	6	50	5	2	0	6	5	3	6	0	0	4	5	13	2	17	0	4	9	0	6	1
63	20	11	0	8	51	87	68	288	9	68	12	6	1	11	8	5	9	0	7	2	5	18	0	17	3	8	6	5	6	6
64	14	7	0	6	30	87	103	172	7	53	5	1	1	0	8	0	8	0	3	0	15	6	3	11	5	0	4	0	1	1
65	5	25	10	5	16	84	37	218	3	51	6	3	2	1	3	9	3	0	3	2	6	11	1	18	0	3	3	1	3	4
66	7	10	11	5	13	83	41	244	2	68	1	2	3	6	16	10	7	0	2	2	6	12	0	21	2	7	5	3	2	6
67	3	5	6	3	14	83	46	181	5	36	5	1	0	5	4	3	3	0	3	1	4	8	5	12	0	2	3	0	1	0
68	3	7	2	1	12	75	30	140	3	40	6	1	1	1	4	8	8	1	0	0	4	9	2	13	0	3	3	2	0	1
69	5	27	12	9	19	93	90	301	3	66	6	1	5	2	12	2	6	1	1	5	5	12	11	30	1	3	10	3	7	4
70	14	11	6	13	15	118	79	282	8	87	12	1	2	9	10	3	13	0	5	6	9	19	4	16	1	7	9	7	3	1
71	7	16	3	3	15	75	38	173	0	37	8	1	0	1	9	8	7	1	4	2	3	9	14	7	6	1	1	3	0	3
72	6	25	5	9	17	98	52	176	7	35	12	0	0	2	5	11	2	0	8	1	1	12	26	10	2	0	6	4	0	4
73	5	29	13	5	27	82	42	203	2	62	10	2	0	2	5	10	12	1	2	1	6	9	10	26	5	5	5	3	4	5
74	4	9	3	4	2	35	23	102	1	24	2	0	1	3	3	8	3	2	3	0	3	7	1	8	0	2	0	2	2	2
75	5	27	5	3	16	90	36	206	1	44	2	2	1	5	3	5	2	1	1	1	4	12	2	14	4	2	2	0	2	2
76	4	28	10	7	25	95	55	208	1	59	5	0	2	4	3	11	6	1	2	3	7	16	9	18	1	2	6	3	0	2
77	3	32	10	3	16	71	41	180	2	60	6	0	1	9	6	6	6	0	3	0	1	19	6	17	1	1	2	2	0	2
78	10	19	9	12	25	126	74	306	4	68	3	4	2	4	10	2	9	1	3	6	10	16	2	26	2	3	18	1	2	1
79	5	5	2	3	4	41	24	87	1	35	5	2	0	2	2	2	8	0	0	2	0	2	1	8	1	0	4	0	1	1
80	9	10	6	7	13	113	68	253	12	47	3	2	3	3	5	1	7	2	1	26	15	8	1	14	0	5	7	2	1	8
81	16	21	13	19	32	156	85	375	5	130	7	4	5	9	6	7	15	0	4	4	12	22	0	42	3	6	8	2	2	9
82	0	11	4	0	4	82	41	166	6	38	2	5	0	9	2	1	1	0	0	2	6	10	0	14	0	0	3	1	1	1
83	18	48	20	22	79	213	119	474	15	207	19	5	5	22	24	11	16	2	5	9	13	20	0	59	2	6	20	2	1	9
84	20	15	11	16	28	130	85	364	8	86	10	5	5	17	7	2	30	1	1	2	8	14	2	36	3	8	26	7	2	12
85	17	6	12	10	11	113	72	240	8	73	6	1	1	30	4	1	12	0	2	3	9	19	5	13	2	5	14	2	2	1

Table B.23: Function words counts in Federalist Papers for the thirty words with the largest T_i (2/2)

Bibliography

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

Argamon, S. (2008). Interpreting Burrow's delta: geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23, 131-147.

Baayen, H. (2001). *Word Frequency Distributions.* Dordretch: Kluwer.

Banfield, J.D. and Raftery, A.E. (1993). Model based Gaussian and non-Gaussian clustering. *Biometrics*, 49, 803-821.

Besag, J., York, J. and Mollie, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.*, 43, 1-59.

Binongo, J.N.G. (1994). Joaquin's Joaquinesquerie, Joaquinesqueri's Joaquin: A Statistical Expression of a Filipino Writer's Style. *Literary and Linguistic Computing*, 9, 267-279.

Binongo, J.N.G. (2003). Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, 16, 9-17.

Bosch, R.A. and Smith, J.A. (1998). Separating hyperplanes and the authorship of the disputed Federalist Papers. *American Mathematical Monthly*, 105, 601-608.

Brinegar, C.S. (1963). Mark Twain and the Quintus Curtius Snodgrass letters: A statistical test of authorship. *Journal of the American Statistical Association*, 58, 85-96.

Bruno, A.M. (1974). *Toward a Quantitative Methodology for Stylistic Analysis of Narrative Style*. Berkeley, University of California Press.

Burrows, J.F. (1987). Words patterns and story shapes: the statistical analysis of narrative style. *Literary and Linguistic Computing*, 2, 61-70.

Burrows, J.F. (1992). Not unless you ask nicely: the interpretative nexus between analysis and information. *Literary and Linguistic Computing*, 7, 91-109.

Burrows, J.F. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3): 267287.

Burrows, J.F. (2007). All the way through: testing for authorship in different frequency strata. *Literary and Linguistic Computing*, 22, 27-47.

Casella, G., Moreno, E. and Giron, J. (2014). Cluster analysis, model selection and prior distributions on models. *Bayesian Analysis*,9, 613-658.

Chaski, C.E. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4, 1-13.

Collins, J., Kaufer, D., Vlachos, P., Butler, B. and Ishizaki, S. (2004). Detecting collaborations in text : Comparing the authors rhetorical language choices in the Federalist Papers. *Computers and the Humanities*, 38, 15-36.

Craig, H. and Kinney, A.F. (2009). *Shakespeare, Computers, and the Mistery of Authorship*. Cambridge: Cambridge University Press.

Diederich, J., Kindermann, J., Leopold, E., and Paass, G. (2003). Authorship attribution with support vector machines. *Applied Intelligence*, 19, 109-123.

Dunn, P.K., Smyth, G.K. (2008). Evaluation of Tweedie exponential dispersion model densities by Fourier inversion. *Statistics and Computing*, 18 (1) , 73-86

Edmondson, P. and Wells, S. (2013). *Shakespeare Beyond Doubt: Evidence, Argument, Controversy.* Cambridge: Cambridge University Press.

Fernandez, C. and Green, P.J. (2002). Modelling spatially correlated data via mixtures: a Bayesian approach. *Journal of the Royal Statistical Society B,* 64, 805-826.

Font, M., Puig, X. and Ginebra, J. (2013). A Bayesian analysis of frequency count data. *Journal of Statistical Computation and Simulation,* 83, 229-246.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3, 1289-1305.

Forsyth, R. and Holmes, D. (1996). Feature-finding for text classification. *Literary and Linguistic Computing*, 11, 163-174.

Fraley, C. and Raftery, A.E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association,* 97, 611-631.

Gale, W.A., Church, K.W., and Yarowsky, D. (1993). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26, 415-439.

Gareth, J., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

Gelfand, A.E., and Dey, D.K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Serie B*, 56, 501-514.

Gelman, A., Carlin, J.C., Stern, H. and Rubin, D.B. (2004). *Bayesian Data Analysis* (2nd ed). New York: Chapman and Hall.

Ginebra, J. (2007). On the measure of the information in a statistical experiment. *Bayesian Analysis*, 2, 167-212.

Ginebra, J. and Puig, X. (2010). On the measure and the estimation of evenness and diversity. *Computational Statistics and Data Analysis*, 54, 2187-2201.

Giron, J., Ginebra, J. and Riba, A. (2005). Bayesian analysis of a multinomial Sequence and Homogeneity of Literary Style, *The American Statistician*, 59, 19-30.

Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40, 237-264.

Grant, T.D. (2007). Quantifying evidence for forensic authorship analysis. *International Journal of Speech Language and the Law*, 14, 1-25.

Greenacre, M. (1988). Clustering the rows and columns of a contingency table, *Journal of Classification*, 5, 39–51.

Greenacre, M. (2007). *Correspondence Analysis in Practice, 2nd Ed.* London, Chapman Hall.

Grieve, J. (2007). Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing*, 22, 251-270.

Hilton, M.L. and Holmes, D.I. (1993). An assessment of cumulative control charts for authorship-attribution, *Literary and Linguistic Computing*, 8, 73–80.

Holla, M.S. (1966). On a Poisson-inverse Gaussian distribution. *Metrika*, 11, 115-121.

Holmes, D.I. (1985). The analysis of literary style. A review, *Journal of the Royal Statistical Society, Ser A*, 148, 328–341.

Holmes, D.I. (1992). A stylometric analysis of Mormon scripture and related texts, *Journal of the Royal Statistical Society, Ser. A*, 155, 91–120.

Holmes, D.I. (1994). Authorship attribution. *Computers and the Humanities,* 28, 87-106.

Holmes, D.I. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing,* 13, 111-117.

Holmes, D.I. (1999). Stylometry. *Encyclopedia of Statistical Sciences; Update Vol.3*, pp. 721-727. New York: Wiley.

Holmes, D.I. and Forsyth, R.S. (1955). The federalist revised: new directions in author attribution. *Literary and Linguistic Computing*, 10, 111-127.

Hope, J. (1994). *The Authorship of Shakespeare's Plays*. Cambridge: Cambridge University Press.

Hope, J. (2010). *Shakespeare and Language: Reason, Eloquence and Artifice in the Renaissance*. London: The Arden Shakespeare.

Hoover, D.L. (2001). Statistical stylistics and authorship attribution: an empirical investigation. *Literary and Linguistic Computing*, 10, 111-127.

Hoover, D.L. (2003). Multivariate analysis and the study of style variation. *Literary and Linguistic Computing*, 18, 341-360.

Hoover, D.L. (2004). Testing Burrow's Delta. *Literary and Linguistic Computing*, 19, 453-475.

Jackson, M.P. (2003). *Defining Shakespeare: Pericles as Test Case*. Oxford: Oxford University Press.

Joachims, T.T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceed.of the 10th European conference on machine learning*, pp. 137-142.

Jockers, M.L. and Witten, D.M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, 25, 215-223.

Jockers, M.L., Witten, D.M. and Criddle, C.S. (2008). Reassessing authorship in the book of Mormon using nearest Shrunken centroid classification. *Literary and Linguistic Computing*, 23, 465-491.

Juola, P., Sofko, J. and Brennan, P. (2006). A prototype for authorship attribution studies. *Literary and Linguistic Computing*, 21, 169-178.

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data*. Wiley, New York.

Khmelev, D.V. and Tweedie, F.J. (2001). Using Markov chains for identification of writers. *Literary and Linguistic Computing*, 16, 299-307.

Koppel, M., Akiva, N. and Dagan, I. (2006). Feature instability as a criterion for selecting potential style markers. *Journal of the American Society for Information Science and Technology*, 57, 1519-1525.

Lewis, D.D. (1998). Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. *Proceed. of the 10th European Conference on Machine Learning*, pp. 4-15

Li, J., Zheng, R. and Chen, H. (2006). From fingerprint to writeprint. *Communications of the ACM*, 49, 76-82.

Lunn, D. (2003). WinBUGS Development Interface (WBDev). *ISBA Bulletin*, **10**, 1011.

Lunn, D.J., Jackson, C. (2004). WinBUGS Development Interface (WBDev)- Implementing Your Own Univariate Distributions *WinBUGS WBDev website*

Lunn, D.J., Jackson, C., Best, N., Thomas, A. and Spiegelhalter, D. (2013). *The BUGS Book. A Practical Introduction to Bayesian Analysis*. Chapman Hall, London.

Lunn, D.J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS – A Bayesian modelling framework: concepts, structure and extensibility. *Statistics and Computing*, 10, 325-337.

Luyckx, K. (2010). *Scalability Issues in Authorship Attribution*. Brussels: University Press Antwerp.

Martindale, C. and McKenzie, D. (1995). On the utility of content analysis in author attribution: The Federalist. *Computers and the Humanities*, 29, 259-270.

Martorell, J, and Galba, M.J. (1490), *Tirant lo Blanc* (in catalan), ed. M. Riquer 1983, Barcelona: Edicions 62. (Translated into English by D. Rosenthal in 1986, Baltimore, Johns Hopkins University Press, and by La Fontaine in 1993, Boston, Peter Lang).

Matthews, R. and Merriam, T. (1993). Neural computation in stylometry: A application to the works of Shakespeare and Fletcher. *Literary and Linguistic Computing*, 8, 203-209.

McCallum, A. and Nigan K. (1998). A comparison of event models for naive Bayes text classification. *AAAI-98 Workshop on Learning for Text Categorization*, Madison, Wisconsin.

Mendenhall, T.C. (1887). The characteristic curves of composition, *Science*, IX, 237–249.

Mendenhall, T.C. (1901). A mechanical solution of a literary problem, *The Popular Science Monthly*, LX, 97-105.

Merriam, T. and Matthews, R. (1994). Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe. *Literary and Linguistic Computing*, 9, 1-6.

Miranda-Garcia, A., and Calle-Martin, J. (2007). Function words in authorship attribution studies. *Literary and Linguistic Computing*, 22: 27-47.

Mollie, A. (1996). Bayesian mapping of disease. In *Markov Chain Monte Carlo in Practice* (eds W.R. Gilks, S. Richardson and D.J. Spiegelhalter), pp. 359-379. London: Chapman and Hall.

Morton, A.Q. (1978). *Literary Detection*, New York: Scribners.

Mosteller, F. and Wallace, D.L. (1964, 84). *Applied Bayesian and Classical Inference; the Case of The Federalist Papers*, 1rst and 2nd edn, Berlin: Springer-Verlag.

Murtagh, F. and Raftery, A.E. (1984). Fitting straight lines to point patterns. *Pattern Recognition*, 17, 479-483.

Oakes, M.P. (1998), *Statistics for Corpus Linguistics*, Edimburg: Edimburgh University Press.

Peng, F., Shuurmans, D. and Wang, S. (2004). Augmenting naive Bayes classifiers with statistical language models. *Information Retrieval Journal*, 7, 317-345.

Pollatschek, M. and Radday, Y.T. (1981). Vocabulary richness and concentration in Hebrew biblical literature. *Association for Literary and Linguistical Computing Bulletin*, **8**, 217-231.

Puig, X., Font, M., and Ginebra, J. (2014). Classification of literary style that takes order into consideration. *Journal of Quantitative Linguistics* 22, 177-201.

Puig, X., and Ginebra, J. (2014a). A Bayesian cluster analysis of election results. *Journal of Applied Statistics* 41, 73-94.

Puig, X. and Ginebra, J. (2014b). A Cluster analysis of vote transitions. *Computational Statistics and Data Analysis* 70, 328-344.

Puig, X., Ginebra, J., and Font, M. (2010). The Sichel model and the mixing and truncation order. *Journal of Applied Statistics* 37, 1585-1603

Puig, X., Ginebra, J., and Perez-Casany, M. (2009). Extension of the zero truncated inverse Gaussian-Poisson model. *Statistical Modelling*, 9, 151-171.

Ramsey J., and Silverman B.W. (2005). *Functional Data Analysis*. New York: Springer Verlag.

Riba, A. and Ginebra, J. (2005). Change-point estimation in a multinomial sequence and homogeneity of literary style. *Journal of Applied Statistics*, 32, 61-74.

Riba, A. and Ginebra, J. (2006). Diversity of vocabulary and homogeneity of literary style. *Journal of Applied Statistics*, 33, 729-741.

Riquer, M. (1990), *Aproximació al Tirant lo Blanc* (in Catalan), Barcelona:Quaderns Crema.

Rybicki, J., and Eder, M. (2011). Deeper Delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing*, 26, 315-321.

Shahan, J.M. and Waugh, A. (2013). *Shakespeare Beyond Doubt? Exposing and Industry in Denial*. London: Llumina Press.

Schneider, K.M. (2003). A comparison of event models for Naive Bayes anti-spam e-mail filtering. Proceed. of the tenth conference on the European chapter of the Association for Computational Linguistics, Vol. 1, pp. 307-314

Sebastiani, F. (2002), Machine learning in automated text categorization. *ACM Computing Surveys*, 34, 1-47.

Seshadri, V. (1993). *The inverse Gaussian distribution: A case study in exponential families.* Oxford: Clarendon Press.

Seshadri, V. (1998). *The Inverse Gaussian Distribution: Statistical Theory and Applications.* New York: Springer Verlag.

Sichel. H.S. (1975). On a distribution law for words frequencies. *Journal of the American Statistical Association*, **70**, 542-547.

Sichel, H.S. (1985). A bibliometric distribution that really works. *Journal of the American Society for Information Science*, **36**, 314-321.

Sichel, H.S. (1986a). Word frequency distributions and type-token characteristics. *Mathematical Scientist*, **11**, 45-72.

Sichel, H.S. (1986b). Parameter estimation for a word frequency distribution based on occupancy theory. *Communications in Statistics, Theory and Methods*, **15**, 935-949.

Smith, M.W.A. (1983). Recent experience and new developments of methods for the determination of authorship. *Association for Literary and Linguistic Computing Bulletin*, 11, 73-82.

Smith, M.W.A. (1990). A Note on the Authorship of "Pericles." *Computers and the Humanities*, 24, 295-300.

Spiegelhalter, D.J., Thomas, A., Best, N. and Lunn, D. (2003). WinBUGS version 1.4 User Manual. http://wwww.mrc-bsu.cam.ac.uk/bugs/WinBUGS/manual14.pdf.

Stamatatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society of Information Science and Technology*, 60, 538-556.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26, 471-495.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35, 193-214.

Tambouratzis, G., Markantonatou, S., Hairetakis, N., Vassiliou, M., Carayannis, G. and Tambouratzis, D. (2004). Discriminating the registers and styles in the Modern Greek language-Part 2. Exteding the feature vector to optimize author discrimination. Literary and Linguistic Computing. 19, 221-242.

Tweedie, F., Singh, S. and Holmes, D. (1996). Neural network applications in stylometry: The Federalist papers. *Computers and the Humanities*, 30, 1-10.

Uzuner, O. and Katz, B. (2005). *A comparative study of language models for book and author recognition*. Lecture Notes in Computer Science, Springer Verlag.

Vargas Llosa, A. (1991), *Carta de Batalla por Tirant lo Blanc*, (in Spanish), Barcelona: Seix Barral.

Vargas Llosa, A. (1993), Tirant lo Blanc: Las palabras como hechos, (in Spanish), in *Actes del Symposion Tirant lo Blanc*, 587–604, Barcelona: Quaderns Crema.

Vickers, C.B. (2004). *Shakespeare, Co-Author: A Historical Study of Five Collaborative Plays*. Oxford: Oxford University Press.

Wetzels, R., Lee, M.D., Wagenmakers, E.J. (2009). Bayesian Inference Using WBDev: A Tutorial for Social Scientists. http://www.winbugs-development.org.uk/files/wbdev_tutorial.pdf.

Williams, C.B. (1975). Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon. *Biometrika*, 62, 207-212.

Yu, B. (2008). An evaluation of text classification methods for literary study. *Literary and Linguistic Computing*, 23, 327-342.

Yule, G.U. (1938). On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship. *Biometrika*, 30, 363-390.

Yule, G.U. (1944). On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship. *Biometrika*, 30, 363-390.

Zheng, R., Li, J., Chen, H. and Huang, Z. (2006). A framework for authorship identification of onlie messages: Writing style features and classification techniques. *Journal of the American Society of Information Science and Technology*, 57, 378-393.

Zhao, Y., and Zobel, J. (2005). Effective and scalable authorship attribution using function words. *Information Retrieval Technology*, 3689, 174-189.

List of Tables

List of Figures