Significance Testing of Word Frequencies in Corpora

This is a post-print version of the following paper: Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki and Heikki Mannila. 2016. "Significance testing of word frequencies in corpora". *Digital Scholarship in the Humanities* 31(2): 374–397. doi:10.1093/llc/fqu064, free access link: https://academic.oup.com/dsh/article/31/2/374/2462752/Significance-testing-of-word-frequencies-in?guestAccessKey=b1a8cf86-2d64-4be6-b5c9-f60402df9799

Author: Jefrey Lijffijt

Affiliation: Aalto University

Current affiliation: Ghent University

Mail: Dept. of Electronics and Information Systems, Technicum, Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium.

E-mail: jefrey.lijffijt@ugent.be

Author: Terttu Nevalainen

Affiliation: University of Helsinki

Author: Tanja Säily

Affiliation: University of Helsinki

Author: Panagiotis Papapetrou

Primary affiliation for this manuscript: Aalto University

Current affiliation: Stockholm University

Author: Kai Puolamäki

Primary affiliation for this manuscript: Aalto University

Current affiliation: Finnish Institute of Occupational Health

Author: Heikki Mannila

Affiliation: Aalto University

Abstract

Finding out whether a word occurs significantly more often in one text or corpus than in another is an important question in analysing corpora. As noted by Kilgarriff (2005), the use of the χ^2 and log-likelihood ratio tests is problematic in this context, as they are based on the assumption that all samples are statistically independent of each other. However, words within a text are not independent. As pointed out in Kilgarriff (2001) and Paquot & Bestgen (2009), it is possible to represent the data differently and employ other tests, such that we assume independence at the level of texts rather than individual words. This allows us to account for the distribution of words within a corpus. In this article we compare the significance estimates of various statistical tests in a controlled resampling experiment and in a practical setting, studying differences between texts produced by male and female fiction writers in the British National Corpus. We find that the choice of the test, and hence data representation, matters. We conclude that significance testing can be used to find consequential differences between corpora, but that assuming independence between all words may lead to overestimating the significance of the observed differences, especially for poorly dispersed words. We recommend the use of the t-test, Wilcoxon rank-sum test, or bootstrap test for comparing word frequencies across corpora.

1. Introduction

Comparison of word frequencies is among the core methods in corpus linguistics and is frequently employed as a tool for different tasks, including generating hypotheses and identifying a basis for further analysis. In this study, we focus on the assessment of the statistical significance of differences in word frequencies between corpora. Our goal is to answer questions such as 'Is word *X* more frequent in male conversation than in female conversation?' or 'Has word *X* become more frequent over time?'.

Statistical significance testing is based on computing a p-value, which indicates the probability of observing a test statistic that is equal to or greater than the test statistic of the observed data, based on the assumption that the data follow the null hypothesis. If a p-value is small (i.e. below a given threshold α), then we reject the null hypothesis. In the case of comparing the frequencies of a given word in two corpora the test statistic is the difference between these frequencies and, put simply, the null hypothesis is that the frequencies are equal.

However, to employ a test, the data have to be represented in a certain format, and by choosing a representation we make additional assumptions. For example, to employ the χ^2 test, we represent the data in a 2x2 table, as illustrated in Table 1. We refer to this representation as the bag-of-words model. This representation does not include any information on the distribution of the word *X* in the corpora. When using this representation and the χ^2 test, we implicitly assume that all words in a corpus are statistically independent samples. The reliance on this assumption when computing the statistical significance of differences in word frequencies has been challenged previously; see, for example, Evert (2005) and Kilgarriff (2005).

Table 1 The 2x2 table that is used when employing the χ^2 test

	Corpus S	Corpus T
Word *X*	*A*	*B*
Not word *X*	*C*	*D*

Hypothesis testing as a research framework in corpus linguistics has been debated but remains, in our view, a valuable tool for linguists. A general account on how to employ hypothesis testing or keyword analysis for comparing corpora can be found in Rayson (2008). We observe that the discussion regarding the usefulness of hypothesis testing in the field of linguistics has often been conflated with discussions pertaining to the assumptions made when employing a certain representation and statistical test. Kilgarriff (2005) asserts that the 'null hypothesis will never be true' for word frequencies. As a response, Gries (2005) argues that the problems posed by Kilgarriff can be alleviated by looking at (measures of) effect sizes and confidence intervals, and by using methods from exploratory data analysis. Our main point is different from that of Gries (2005). While we endorse Kilgarriff's conclusion that the assumption that all words are statistically independent is inappropriate, the lack of validity of one assumption does not imply that there are no comparable representations and tests based on credible assumptions.

As pointed out in Kilgarriff (2001) and Paquot & Bestgen (2009), it is possible to represent the data differently and employ other tests, such as the t-test, or the Wilcoxon rank-sum test, such that we assume independence at the level of texts rather than individual words. An alternative approach to the 2x2 table presented above is to count the number of occurrences of a word per text, and then compare a list of

(normalized) counts from one corpus against a list of counts from another corpus. An illustration of this representation is given in Table 2. This approach has the advantage that we can account for the distribution of the word within the corpus.

Table 2 The frequency lists that are used when employing the t-test. The lists do not have to be of equal length, as the corpora may contain an unequal number of texts.

Corpus S	Text S_1	Text S_2	…	Text S_N
Normalized frequency of word *X*	S_1	S_2	...	$S_{\|S\|}$

Corpus T	Text T_1	Text T_2	…	Text T_M
Normalized frequency of word *X*	T_1	T_2	...	$T_{\|T\|}$

We emphasize that the utility of hypothesis testing critically depends on the credibility of the assumptions that underlie the statistics. We share Kilgarriff's (2005) concern that application of the χ^2 test leads to finding spurious results, and we agree with Kilgarriff (2001) and Paquot and Bestgen (2009) that there are more appropriate alternatives, which, however, have not been implemented in current corpus linguistic tools. We re-examine the alternatives and provide new insights by analysing the differences between six statistical tests in a controlled resampling setting, as well as in a practical setting.

The question which method is most appropriate for assessing the significance of word frequencies or other statistics is not new. Dunning (1993) and Rayson and Garside (2000) suggest that a log-likelihood ratio test is preferable to a χ^2 test because the latter test is inaccurate when the expected values are small (< 5). Rayson *et al*. (2004) propose using the χ^2 test with a modified version of Cochrane's rule. Kilgarriff (2001) concludes

that the Wilcoxon rank-sum test[1] is more appropriate than the χ^2 test for identifying differences between two corpora, but his study is limited to a qualitative analysis of the top 25 words identified by the two methods. Kilgarriff (2005) criticizes the hypothesis testing approach because the χ^2 test finds numerous significant results, even in random data.

Hinneburg *et al.* (2007) study methods based on bootstrapping and Bayesian statistics for comparing small samples. Paquot and Bestgen (2009) present a study of the similarities and differences between the t-test, the log-likelihood ratio test, and the Wilcoxon rank-sum test; however, their study is also limited to qualitative analysis of the differences. They recommend using multiple tests, or the t-test, if only one method is to be applied. Lijffijt *et al*. (2011) illustrate that the bootstrap and inter-arrival time tests provide more conservative p-values than those that are provided by bag-of-words-based models (i.e. tests based on the assumption that all words are statistically independent), which includes the χ^2 and log-likelihood ratio tests. Lijffijt *et al*. (2012) conduct a detailed study of lexical stability over time in the *Corpus of Early English Correspondence*, using both the log-likelihood ratio and bootstrap tests, and conclude that the log-likelihood ratio test marks spurious differences as significant.[2] Relevant, but not discussed further here, is the need for balanced corpora when comparing word frequencies (Oakes and Farrow, 2007).

We find that some statistical tests that are commonly used in corpus linguistics, such as the χ^2 and log-likelihood ratio tests (Dunning, 1993; Rayson and Garside, 2000), are *anti-conservative*, that is, their p-values are excessively low, when we assume that a corpus is a collection of statistically independent texts. We perform experiments based on a subcorpus of the British National Corpus (BNC, 2007) that

contains all texts from the prose fiction genre. We quantify the potential bias of the tests based on the uniformity of p-values when we randomly divide the set of texts into two groups. This method is further explained in Section 3. Moreover, we show that the errors in the estimates differ according to each word and the dispersion of the words in the corpus. To define the *dispersion* of a word, we consider a measure of dispersion, DP_{norm}, which was introduced in Gries (2008) and refined in Lijffijt and Gries (2012).

Because the bias that we observe does not solely depend on word frequency, we cannot simply use higher cut-off values in the χ^2 or log-likelihood ratio tests to correct the bias. Notably, the rank of words, in terms of their significance, changes. Finally, we perform a keyword analysis of the differences between male and female authors, as annotated by Lee (2001), using two methods. We find that the differences between the methods are substantial and thus necessitate the use of a representation and statistical test such that the distribution of the frequency over texts is properly taken into account (the t-test, Wilcoxon rank-sum test, or the bootstrap test).

2. Why the Bag-of-Words Model is Inappropriate

The χ^2 and log-likelihood ratio tests are based on the *bag-of-words model* (illustrated in Table 1), in which all words in a corpus are assumed to be statistically independent. From the perspective of any word, the corpus is modelled as a Bernoulli process, i.e. a sequence of biased coin flips, which results in word frequencies that follow a binomial distribution (Dunning, 1993). The bag-of-words model implicitly assumes both a mean frequency and a certain variance of the frequency over texts and thus an expected dispersion. Figure 1 shows the observed frequency distribution of the word *I* in the British National Corpus and the expected frequency distribution in the bag-of-words

model. The observed distribution and the distribution that is predicted by the bag-of-words model clearly differ.

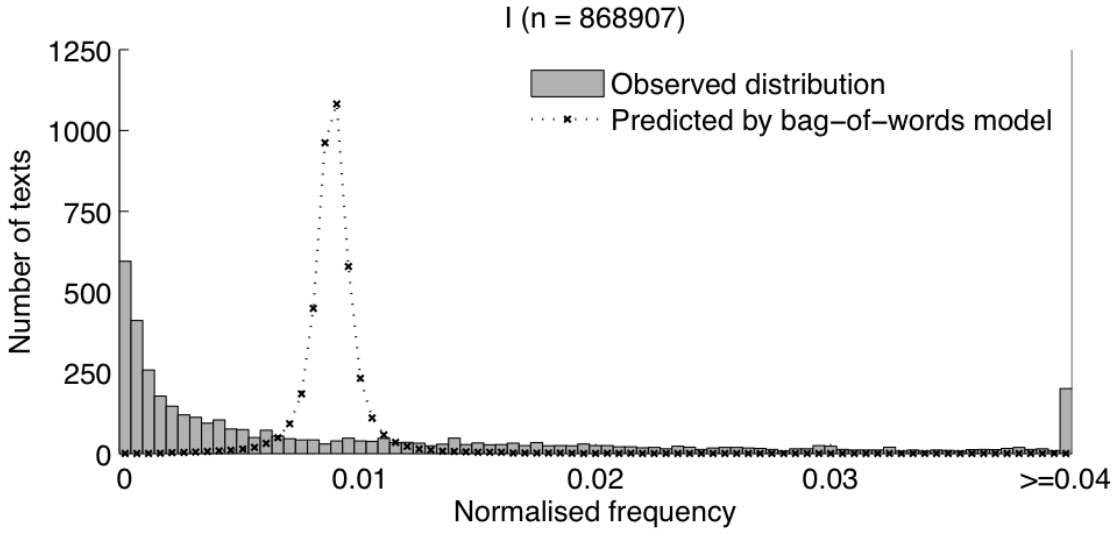

Fig. 1 The frequency distribution of *I* in the British National Corpus. The grey bars show a histogram of the observed distribution, and the black dotted line shows the expected distribution in the bag-of-words model, on which the χ^2 and log-likelihood ratio tests are based. Compared with the prediction, the observed distribution has much greater variance and thus demonstrates that the bag-of-words model is not an appropriate choice when comparing corpora, even for highly frequent words.

Another example is presented in Table 3, which depicts p-values for the hypothesis that the name *Matilda* is used at an equal frequency by male and female authors in the prose fiction subcorpus of the British National Corpus. This subcorpus is presented in Section 4. The frequency for male authors is 56.7 per million words (absolute frequency 408), and the frequency for female authors is 20.2 per million words (absolute frequency 169). With more than 500 occurrences in the fiction subcorpus, we may easily trust the results of the χ^2 and log-likelihood ratio tests, which show that male authors use this name more often than female authors. However, the other tests (the t-test, Wilcoxon rank-sum test, inter-arrival time test, and bootstrap test) indicate that the observed frequency difference is not unlikely to occur at random. The

reason that the methods disagree is that the word is used in only 5 of 409 total texts (1 text written by a male author and 4 texts written by female authors), with an uneven frequency distribution: one text contains 408 instances, followed by, in the other texts, 155 instances, 11 instances, 2 instances, and 1 instance, respectively. This uneven distribution should lead to an uncertain estimate of the mean frequency. In other words, the variance of the frequency of *Matilda* is very high. The χ^2 and log-likelihood ratio tests do not account for the uneven distribution, as these tests use only the total number of words in a corpus, and as a result they underestimate the uncertainty.

Table 3 p-values for the hypothesis that male and female authors use the name *Matilda* at an equal frequency, based on the prose fiction subcorpus of the British National Corpus

χ^2 test[3]	Log-likelihood ratio test	Welch's t-test	Wilcoxon rank-sum test	Inter-arrival time test	Bootstrap test
< 0.0001	< 0.0001	0.4393	0.1866	0.5826	0.7768

The remainder of this paper is structured as follows. In Section 3, we present the significance testing methods, the uniformity test, and the dispersion measure. In Section 4, we describe the data that are used. In Section 5, we compare the methods in a series of experiments based on random divisions of the corpus, and in Section 6 we describe the differences between male and female authors that were identified using various methods. Section 7 briefly concludes the paper.

3. Methods

In this section, we briefly discuss the mathematical models and assumptions that underlie each of the six methods discussed in the introduction. A summary of the essential differences is given in Section 3.8. The statistical test employed in the controlled random sampling experiment (Section 5) is presented in Section 3.9, and the measure of dispersion that we use is presented in Section 3.10. Readers less interested in the specifics of the statistical tests may proceed directly to 3.8 and then to Section 4.

3.1 Notation

We use q to denote the word that we intend to compare in two corpora, and S and T to denote the two corpora. Corpus S contains $|S|$ texts and $size(S)$ words. We use subscripts to indicate individual texts: S_1, S_2, …, $S_{|S|}$. We express the relative frequency of word q in corpus S as $freq(q,S)$. Each of the following six methods computes a p-value for the hypothesis of a word having an equal frequency in the two corpora, $freq(q,S) = freq(q,T)$, against the alternative hypothesis that the frequencies are not equal: $freq(q,S) > freq(q,T)$ or $freq(q,S) < freq(q,T)$. Thus, conforming to the tradition in corpus linguistics, all methods provide two-tailed p-values.

3.2 Method 1: Pearson's χ^2 Test

Pearson's χ^2 test, which is also known as the χ^2 test for independence or simply as the *χ^2 test*, is based on the assumption that a text or corpus can be modelled as a sequence of independent Bernoulli trials. Each Bernoulli trial is a random event with a binary outcome; thus, the entire sequence is similar to a sequence of biased coin flips. Under the assumption of independent Bernoulli trials, the probability distribution for the word frequency is given by the probability mass function of the binomial distribution. Let n

be the size of the corpus and p the relative frequency of a word. The probability of observing this word exactly k times is given by

$$\Pr(K = k) = p^2(1 - p)^{n-k} \binom{n}{k}. \quad (1)$$

This distribution is approximately normal with mean np and variance $np(1-p)$ when $np(1-p) > 5$ (Dunning, 1993). The fact that this distribution is well approximated by a normal distribution is used in the χ^2 test. The test is conducted as follows. Let $O_1 = freq(q,S) \cdot size(S)$ and $O_2 = freq(q,T) \cdot size(T)$, which are the observed frequencies of q in S and T, respectively. Let p be the relative frequency over the combined corpora, i.e. $p = (O_1+O_2)/(size(S)+size(T))$. We define the expected frequency in S and T as $E_1 = p \cdot size(S)$ and $E_2 = p \cdot size(T)$, respectively. The test statistic X^2 using Yates' correction is given by

$$X^2 = \frac{(|O_1 - E_1| - 0.5)^2}{E_1} + \frac{(|O_2 - E_2| - 0.5)^2}{E_2}. \quad (2)$$

The test statistic asymptotically follows a χ^2 distribution with one degree of freedom. The p-value can be obtained by comparing the test statistic to a table of χ^2 distributions. The χ^2 test is available in most statistical software programs and implemented in tools such as WordSmith Tools (Scott, 2012) and BNCweb (Hoffmann *et al.*, 2008).

3.3 Method 2: Log-Likelihood Ratio Test

The χ^2 test is based on two approximations: the normal distribution approximates the binomial distribution, and the test statistic asymptotically follows a χ^2 distribution. Because of this double approximation, the χ^2 test is inapplicable when the word frequency is small (< 5). For this reason, Dunning (1993) introduces a test which is not

based on the normality approximation but on the likelihood ratio. This test is called the *log-likelihood ratio test* and is also known as the G^2 test.

The likelihood function $H(p;n,k)$ is the same as $Pr(K = k)$ in Equation (1); the only difference is that we explicitly mention the parameter p. The likelihood ratio is the ratio of the probability when we have two parameters, p_1 and p_2 (one for each corpus), divided by the probability when we have only one parameter, p (for both corpora). The precise mathematical formulation is given by $p_1 = freq(q,S)$, $n_1 = size(S)$, $k_1 = freq(q,S) \cdot size(S)$, $p_2 = freq(q,T)$, $n_2 = size(T)$, $k_2 = freq(q,T) \cdot size(T)$, and $p = (k_1+k_2)/(n_1+n_2)$. The likelihood ratio is defined as

$$\lambda = \frac{H(p;n_1,k_1) \times H(p;n_2,k_2)}{H(p_1;n_1,k_1) \times H(p_2;n_2,k_2)}. \qquad (3)$$

We set the parameters p_1, p_2, and p to the values that maximize the likelihood function. The full derivation can be found in Dunning (1993).

The log-likelihood ratio test is based on the fact that the quantity $-2 \log \lambda$ asymptotically follows a χ^2 distribution with degrees of freedom that are equal to the difference in the number of parameters between the ratios (i.e. one in this instance). The quantity $-2 \log \lambda$ is used as the test statistic. Dunning (1993) claims that this test statistic approaches its asymptotical distribution much faster than the test statistic in the χ^2 test and is thus preferable, especially when the expected frequency is low. Again, the final p-value is computed by comparing the test statistic to a table of χ^2 distributions. The log-likelihood ratio test is available in many statistical software programs and implemented in tools such as WMatrix (Rayson, 2008), WordSmith Tools (Scott, 2012), and BNCweb (Hoffmann *et al.*, 2008).

Similar to the χ^2 test, this method is based on the bag-of-words model, the representation illustrated in Table 1, and thus on the assumption that each word can be modelled as an independent Bernoulli trial. As a result, the test ignores all structure in the corpus and even in texts and sentences. We refer to any method that is based on this assumption as a *bag-of-words test*.

There exist other bag-of-words tests that are not based on approximations of the probability mass function given in (1) but are directly based on the summation of values in Equation (1). Such tests provide more accurate probabilities, especially for small frequencies, under the bag-of-words assumption. Examples include Fisher's exact test and the binomial test. We expect these methods to perform similarly to the χ^2 and log-likelihood ratio tests for low word frequencies, and as the frequency increases, the results will converge because all of these tests are based on the bag-of-words assumption and Equation (1). For brevity, we do not consider other bag-of-words tests in this paper.

3.4 Method 3: Welch's t-Test

A t-test is a significance test in which the test statistic follows a Student's t-distribution. We intend to compare two groups of samples and make a minimum number of assumptions. We use *Welch's t-test*, which is based on the assumption that the mean frequency follows a Gaussian distribution. Welch's t-test is more general than Student's t-test because the former test does not assume equal variance in the two populations. Welch's t-test provides a p-value for the hypothesis that the means of the two distributions are equal.

The test statistic is the normalized difference between the means of the word frequencies. Let x_1 be the mean of the frequency of q over texts in S, and let s_1 be the

standard deviation. Likewise, let x_2 be the mean of the frequency of q over texts in T, and let s_2 be the standard deviation. The test statistic t is given by

$$t = \frac{x_1 - x_2}{\sqrt{\frac{s_1^2}{|S|} + \frac{s_2^2}{|T|}}}. \quad (4)$$

The test statistic follows a *Student's t-distribution* with degrees of freedom that depend on the variance of the populations. An exact solution to this problem is unknown, but Welch's t-test is based on the Welch-Satterthwaite equation, which provides an approximate solution (Welch, 1947). Implementations of this test are available in statistical software programs, including *R* and *Microsoft Excel*.

NB. It is often claimed that Student's and Welch's t-test are only applicable if the data follow a normal distribution. This is not true; the assumption is that the test statistic follows a normal distribution. In this case, the test statistic is the difference between the two means. This statistic does not in general follow a normal distribution. However, the Central Limit Theorem (CLT) states that, under very general conditions, the mean of a set of independent random variables approaches normality very fast when the number of samples increases. Since the frequency of a word per text is bounded, the conditions for the CLT are met, and the means x_1 and x_1, as well as their difference are approximately normal when the number of texts is sufficiently large. For small corpora, it is a priori not clear if the test is an appropriate choice.

3.5 Method 4: Wilcoxon Rank-Sum Test

The *Wilcoxon rank-sum test*, which is also known as the Mann-Whitney U-test, is a statistical test that does not make any assumption regarding the shape of the distribution for the quantity of interest. It is based on the fact that if the distributions of q for two

corpora are equal, then it is possible to induce a probability distribution over the rank orders (Wilcoxon, 1945; Mann and Whitney, 1947).

The test is performed as follows. We order all samples based on the frequency of word *q*, regardless of the corpus in which these samples are located. This approach gives us a ranked series, an example of which is shown in Table 4.

Table 4 Example of a ranked series

Rank	1	2	3	4	5	6	7	8	9	10
Corpus	S	T	T	T	S	S	S	T	T	S

The test statistic *U* is then defined as the sum of the ranks of texts of the smaller corpus. In this situation, because both corpora have a size of 5, we can select either S or T. We find that $U_S = 1+5+6+7+10 = 29$ and $U_T = ((n^2+n)/2) - 28 = 26$.

We obtain a p-value for small *n* by comparing the test statistic with a statistical table, and if $n > 20$, then the distribution of the test statistic is well approximated by a Gaussian distribution using known parameters. Implementations of this test are available in statistical software programs, such as *R*.

Multiple texts may have equal frequencies for a word. Particularly for infrequent words, numerous texts in a corpus may have a frequency of zero. The Wilcoxon rank-sum test accounts for texts with equal frequencies by assigning to each text the average rank over all equal-frequency texts. For example, if there are five texts with a frequency of zero, then each text is assigned a rank of 3.

3.6 Method 5: Inter-Arrival Time Test

A novel significance test that is specifically designed for frequency counts in sequences is the *inter-arrival time test*, which was introduced by Lijffijt *et al*. (2011). This test is based on the spatial distribution of a word in a corpus, as modelled by the distribution of inter-arrival times between words. The assumption is that the inter-arrival time distribution of a word captures the behavioural pattern of the word in a corpus. Savický and Hlaváčová (2002) use the inter-arrival time distribution to define a corrected frequency that captures whether words that are frequent in a corpus are ``common'' or not, and Altmann *et al*. (2009) reports that the inter-arrival time distribution of a word, as summarized in a burstiness parameter, is a good predictor of word class.

The significance test is performed as follows. The *inter-arrival times* are obtained by counting the number of words between each consecutive occurrence of word *q*, plus one. The texts in the corpus are ordered randomly and the corpus is treated as though it were placed on a ring: the end of the corpus is attached to the beginning. We begin counting at the first occurrence and continue until we again reach the first occurrence. For example, assume that we have a corpus with ten words and two occurrences of word *q* (Table 5).

Table 5 Example of a small corpus

Index	1	2	3	4	5	6	7	8	9	10
Word	*x*	*x*	*q*	*x*	*x*	*x*	*q*	*x*	*x*	*x*

The inter-arrival times for this corpus are *3+1 = 4* and *5+1 = 6*; thus, the *empirical inter-arrival time distribution* is *{4, 6}*. By definition, the number of inter-

arrival times is equal to the number of occurrences in the corpus, and the sum of the inter-arrival times equals the size of the corpus.

The significance test is based on the production of random corpora by repeatedly sampling inter-arrival times from the empirical inter-arrival time distribution. The first occurrence must be sampled from a different distribution (Lijffijt *et al*., 2011). After we obtain the index of the first occurrence, we sample uniformly at random an inter-arrival time from the empirical inter-arrival time distribution and insert a new occurrence of *q* at the position given by this inter-arrival time. This process is repeated until we exceed the length of the corpus.

In Lijffijt *et al*. (2011), the significance test is based on a foreground corpus *S* and a background corpus *T*. The test is performed by comparing the observed frequency of *q* in *S* to the frequency in randomized corpora with sizes equal to *S* but based on the inter-arrival time distribution of *T*. The test is one-tailed, and the alternative hypothesis is $freq(q,S) > freq(q,T)$. The test is also asymmetrical in that the p-value for $freq(q,S) > freq(q,T)$ is not necessarily the same as $freq(q,S^*) < freq(q,T^*)$ if we set $S^* = T$ and $T^* = S$ because only one corpus is randomized. We adopt a slightly different approach that does not have these issues. We create random corpora S^1 to S^N, based on the inter-arrival time distribution of *S*, and random corpora T^1 to T^N, based on the inter-arrival time distribution of *T*, with all sizes equal to the smaller corpus. The one-tailed p-value is given by the mid-P test (Berry and Armitage, 1995):

$$p = \frac{\sum_{i=1}^{N} H\left(freq(q,T^i) - freq(q,S^i)\right)}{N}, \qquad (5)$$

$$\text{where } H(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0.5 & \text{if } x = 0 \\ 0 & \text{if } x < 0 \end{cases}.$$

We can convert this to a two-tailed p-value (Dudoit *et al*., 2003) using the following equation:

$$p_{two} = 2 \times \min(p, 1 - p). \tag{6}$$

Because the p-value is an empirical estimate and the real p-value that we are approximating may be small, the use of smoothing is appropriate (North *et al*., 2002). Thus, the final p-value is computed as follows:

$$p^* = \frac{p_{two} \times N + 1}{N + 1}. \tag{7}$$

The value p^* is used as the p-value for this test in our experiments.

Obtaining the p-values takes longer compared to the other methods, as it requires sampling many pseudorandom numbers. Specifically, it takes N times the number of tokens in a corpus steps to compute the p-values for all types. For example, for the experiment presented in Section 6, this process takes several minutes.

3.7 Method 6: Bootstrap Test

Bootstrapping (Efron and Tibshirani, 1994) is a statistical method for estimating the uncertainty of some quantity in a data sample by resampling the data several times. We can employ bootstrapping to create a significance test as follows. Similar to the procedure used in the inter-arrival time test, we create a series of corpora S^1 to S^N, but we produce a random corpus by sampling $|S|$ texts from S. Likewise, we create a series

T^1 to T^N by repeatedly sampling $|S|$ texts from T. The p-value is again obtained using Equations (5) through (7).

This method makes no assumptions regarding the shape of the frequency distribution for words and is thus generally applicable. This method is almost identical to the bootstrap test used by Lijffijt *et al*. (2011), but our method differs in that we use a two-tailed p-value and resample both *S* and *T* concurrently. Implementations in *R* and *Matlab* can be found in Lijffijt (2012).

3.8 Summary of Methods

Table 6 summarizes the assumptions underlying the six methods that are described above. The χ^2 and log-likelihood ratio tests represent the data in a 2x2 table, while Welch's t-test, the Wilcoxon rank-sum test, and the bootstrap test take as input a list of frequencies per text for each word. The inter-arrival time test is based on the spatial distribution of a word in the corpora. The Wilcoxon rank-sum and bootstrap tests make the fewest assumptions regarding the frequency distribution and are thus the most generally applicable.

Table 6 Summary of the six methods that are presented in this paper and the assumptions regarding the frequency distribution for each test

Test	Assumption regarding frequency distribution
Pearson's χ^2 test	All words are statistically independent (bag-of-words model)
Log-likelihood ratio test	All words are statistically independent (bag-of-words model)
Welch's t-test	All texts are statistically independent, and the mean frequency follows a normal distribution
Wilcoxon rank-sum test	All texts are statistically independent

Inter-arrival time test	Spaces between occurrences of the same word are statistically independent
Bootstrap test	All texts are statistically independent

3.9 Test for Uniformity of p-Values

All of the previously discussed methods yield p-values for the hypothesis that the frequencies of a word q in S and T are equal. Several studies, including Kilgarriff (2001), Rayson *et al.* (2004), and Paquot and Bestgen (2009), have previously compared some of these methods. These studies have shown that p-values in the same setting are not equal: there are differences in the significance of a given frequency difference between one method and another. This finding is alarming because we do not know which test yields the best results.

We study the utility of these tests based on the criterion that if the data follow the distribution that is assumed in the null hypothesis and the test is unbiased, then the p-values given by the method should be uniformly distributed in the [0, 1] range. This criterion is applicable according to the definition of p-values: the probability of encountering a p-value of x or less is x itself. For example, there is 10% chance of observing a p-value of 0.1 or less, and a 1% chance of observing a p-value of 0.01 or less. If this criterion is not fulfilled, then the test is either anti-conservative (the probability of encountering a p-value of x or smaller is more than x) or conservative (the probability of encountering a p-value of x or smaller is less than x). See, for example, Blocker *et al.* (2006).

When assessing a statistical testing procedure, testing for uniformity of p-values, either visually or by a statistical test, is a common practice in many disciplines such as

particle physics; see e.g. Figures 2–6, 8-9, and 11–12 in Beaujean et al. (2011). A similar kind of experiment has been published in Lijffijt (2013), while for example Schweder and Spjøtvoll (1982) study the uniformity of p-values for multiple-hypotheses adjustment procedures, and L'Ecuyer and Simard (2007) use the Kolmogorov-Smirnov test (also used here) to measure the uniformity of random number generators.

Numerous statistical tests can be utilized to determine whether a distribution is uniform. We employ the *Kolmogorov-Smirnov test* (Massey, 1951), which can be used to compare two distributions. The reference distribution *f(x)* that we use is the uniform distribution on [0, 1]. The test is based on a simple statistic: the maximum distance between the empirical cumulative distribution $F_n(x)$, which is based on the observed data, and the theoretical uniform cumulative distribution function $F(x)$:

$$D_n = \sup_x \left| F_n(x) - F(x) \right|. \tag{8}$$

The quantity $\sqrt{n} D_n$ follows a Kolmogorov distribution. The associated p-value can be found by comparing $\sqrt{n} D_n$ to a table containing critical values for the Kolmogorov distribution. Implementations of this test are available in statistical software programs, including *R*.

3.10 Measure of Dispersion: DP_{norm}

Gries (2008) presents an overview of several dispersion measures and the disadvantages of each measure, and proposes a simple alternative that is reliable and easy to interpret: deviations of proportions (*DP*). The measure is based on the difference between observed and expected relative frequencies. The expected relative frequency is equal to the relative size of a text. Let $v_1,\ldots,v_n$ be the relative frequencies that are observed in texts $S_1,\ldots,S_n$, and let $s_1,\ldots,s_n$ be the relative sizes of the texts. *DP* is defined as

$$DP = \frac{\sum_{i=1}^{n} |s_i - v_i|}{2}, \quad (9)$$

and the normalized measure DP_{norm} is given by

$$DP_{norm} = \frac{DP}{1 - \min_i(s_i)}. \quad (10)$$

The normalized measure, as presented by Lijffijt and Gries (2012), has a minimum value of 0 and a maximum value of 1, regardless of the corpus structure, whereas *DP* also has a minimum of 0, but its maximum depends on the corpus structure. Because the dispersion is quantified as the difference between the expected and observed frequencies, a dispersion of 0 indicates that a word is dispersed as expected, whereas a dispersion of 1 indicates that the word is minimally dispersed. A word is minimally dispersed when it occurs only in the shortest text.

4. Data

For the purposes of our study, we require a relatively large and homogeneous data set containing information on the gender of the authors of the texts. To fulfil this requirement, we have selected a subcorpus of the British National Corpus (BNC, 2007), namely the prose fiction genre. Categorized by Lee (2001), the genre excludes drama but includes both novels and short stories. Lee (2001, p. 57) notes that 'where further sub-genres can be generated on-the-fly through the use of other classificatory fields, they are not given their own separate genre labels, to avoid clutter'—thus, e.g. children's prose fiction is not separated from adult prose fiction because these two types of fiction can be distinguished through the 'audience age' field. As the sub-genres of

prose fiction may differ from one another considerably, our material can be regarded as homogeneous only in relation to other super-genres, such as academic prose.

The prose fiction subcorpus contains 431 texts or c. 16 million words of present-day British English. According to Burnard (2007, Section 1.4.2.2), most of the texts are continuous extracts with a target sample size of 40,000 words, but several texts are included in their entirety. The gender of the authors is known for 409 texts or c. 15.6 million words, which are divided fairly evenly between male and female authors: 203 texts were written by men, and 206 texts were written by women (c. 7.2 and 8.4 million words, respectively). These 409 texts form our data set. For the uniformity experiments in the following section, we use the first 2,000 words of each text, while for the gender study, we analyse the full texts. We preprocess the data set by lowercasing all text; furthermore, punctuation, lemmatization, parts of speech, and multi-word tags are ignored, and only the word forms (i.e. running words) are considered.

5. Uniformity of p-Values

5.1 Randomly Assigning the Texts to Two Sets

The first experiment that we have conducted involves testing the uniformity of the p-values for each method. We have employed the following procedure. We randomly assign 200 texts to corpus *S* and 200 texts to corpus *T*, such that the corpora do not overlap. We then apply each method to all words with a frequency of 50 or greater in the fiction corpus (there are 3,302 such words). The entire process is repeated 500 times.

Due to the fact that the corpus is split into two parts at random, the null hypothesis, that there is no difference between these parts, is by definition true. Notice

that two random samples from a population are almost always different, as long as there is variation in the population the samples are drawn from. That means we expect that there will be differences between the two samples. However, since the assignment is random, any observed structure is fully explained by the artefacts of random sampling, and there is no true discriminative structure present in the data. This procedure is very similar to permutation testing, see for example Good (2005).

For example, assume that we have drawn two samples, and we observe that the word *would* is more frequent in *S* than in *T*. If we also find it has a low p-value, we may think that there is a real difference between the two populations. However, since *S* and *T* are drawn from the same population, we know that there is no true difference between the two populations with respect to the frequency of *would*. Doing many comparisons aggravates this problem, because then we are liable to find many large differences, while there are in fact none.

A statistical test quantifies how likely an observation is under the null hypothesis. Perhaps counter-intuitively, this does not mean that a p-value is always 1 when there is no true difference between the populations; it means that the distribution of a p-value should be approximately uniform on the range [0, 1]. That is, there is a 50% probability that a p-value is 0.5 or lower, 10% probability that it is 0.1 or lower, 1% probability that it is 0.01 or lower, and so on.

In that case, the test is neither conservative, nor anti-conservative. When we do multiple tests, we can use Bonferroni correction, or a more powerful alternative, to ensure that the smallest p-value of a set of tests has a uniform distribution. The probability distribution of the minimum corresponds to the family-wise error rate. Other post-hoc corrections may also have different aims.

Due to the random sampling, the p-values will not be exactly uniform, but—as discussed in Section 3.9—we can employ the Kolmogorov-Smirnov test to quantify the uniformity of the 500 p-values for one word for one test in a single p-value. We repeat this experiment for each word, and obtain for each of the 3,302 words six p-values that express the uniformity of the p-values for each of the six tests. This results in a total of $3{,}302 \cdot 6 = 19{,}812$ p-values.

We use a minimum frequency of 50 because the frequency influences the uniformity of the p-values and the influence differs per method. We do not claim that the significance tests are inapplicable to lower frequencies (in fact, we would argue the opposite), but this experiment is not meaningful using lower frequency words. We have not optimized the frequency threshold, and, as shown below, a frequency of 50 is often too low. Further details regarding why the experiments are not meaningful with less frequent words can be found in the discussion of the experimental results below.

A low p-value for the Kolmogorov-Smirnov test indicates that the p-value distribution over the random corpus assignments is not uniform. However, due to testing 19,812 hypotheses, we do *not* expect all p-values of the Kolmogorov-Smirnov test to be high. To correct for multiple hypotheses, we apply the Bonferroni correction by multiplying each p-value by the number of hypotheses. If a p-value is greater than one after multiplication, then we set the value to one. The Bonferroni correction ensures that there is only α probability of rejecting *any* sample. The correction is conservative, but we also prefer to be cautious and not reject any samples as being non-uniform unless we are certain of their lack of uniformity. For a review of multiple hypothesis correction methods see Shaffer (1995) or Dudoit *et al*. (2003).

Figure 2 shows an overview of the performance of each method. In the following discussion, we write, for brevity, that samples or words are rejected in the uniformity test, where we actually mean that the null hypothesis that the p-values follow a uniform distribution is rejected.

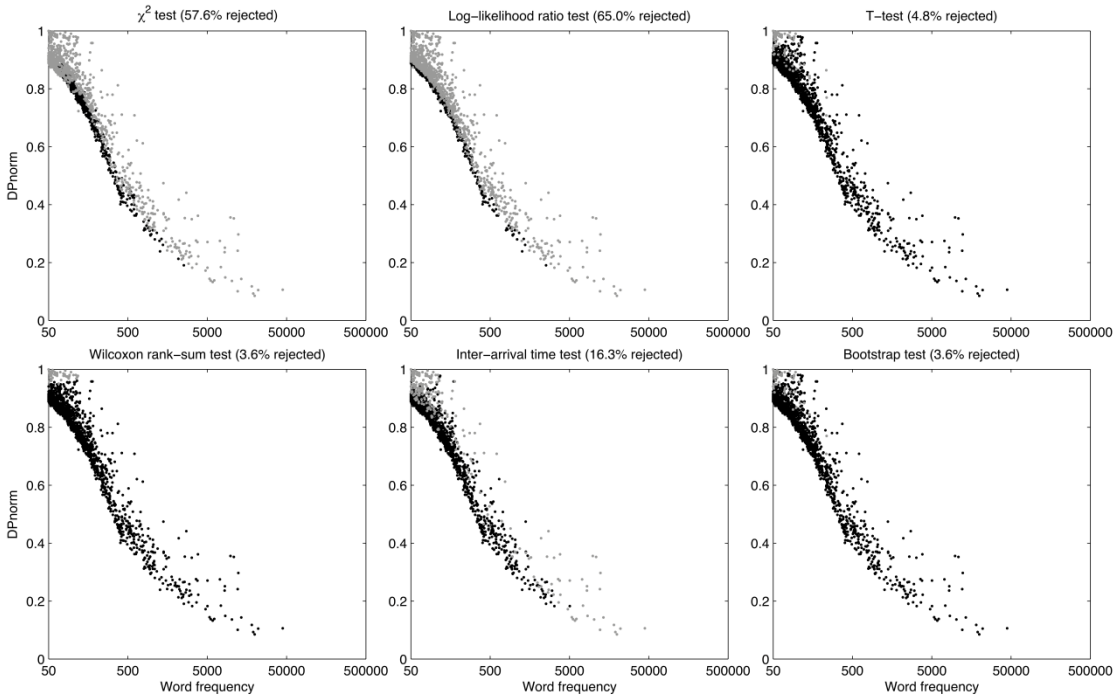

Fig. 2 The results of the uniformity test for all six methods based on random text assignments. Each dot corresponds to a word, which has a frequency (x-axis) and dispersion (y-axis). Light grey dots correspond to rejected samples. A sample is rejected if the corrected p-value of the Kolmogorov-Smirnov test for uniformity is < 0.01. The Wilcoxon rank-sum and bootstrap tests demonstrate the best performance with 3.6% rejected samples.

We observe that 57.6% of the samples are rejected for the χ^2 test, even for the highest frequency, well-dispersed words. The log-likelihood ratio test performs even worse: 65% of the words are rejected, and these also include the most frequent and best dispersed words. The difference is probably caused by Yates' correction for the χ^2 test.

The t-test, Wilcoxon rank-sum test, and bootstrap test perform much better: although 3.6% to 4.8% of the samples are rejected, we observe that these rejected samples consist of infrequent, poorly dispersed words. Thus, testing words with sufficient frequency and/or dispersion yields appropriate results. Because of Zipf's law, we know that the number of infrequent words greatly exceeds the number of frequent words, and thus, if we had selected a lower frequency threshold, then the percentage of rejected samples would have been much higher.

The inter-arrival time test has more rejected samples (16.3%), but these samples again include frequent and well-dispersed words. This result indicates that the test does not capture all of the structure that is present in the texts. This result may have occurred because inter-arrival times have correlations within texts and these are not captured by the test.

The Wilcoxon rank-sum and bootstrap tests demonstrate the best performance. Frequent and well-dispersed words always yield a uniform distribution. When comparing the bootstrap and t-tests, we observe that the samples for which the t-test does not provide a uniform distribution are all instances in which the bootstrap test does not provide a uniform distribution plus a few more. Especially for infrequent but relatively well-dispersed words, the bootstrap test appears to outperform the t-test. In contrast, the Wilcoxon rank-sum test appears to provide a tighter boundary for the rejected samples.

Finally, we have also tested the performance of all tests on words with frequencies between 20 and 50. Figure 3 displays the results. We observe that the χ^2 and log-likelihood ratio tests fail to yield uniform p-values in almost all cases. The t-test and Wilcoxon rank-sum test fail in nearly half of the instances; almost all words that have

frequencies below 30 or that are poorly dispersed are rejected. The inter-arrival time and bootstrap tests are more successful in yielding uniform p-values for low frequency words, with the bootstrap test being the most successful.

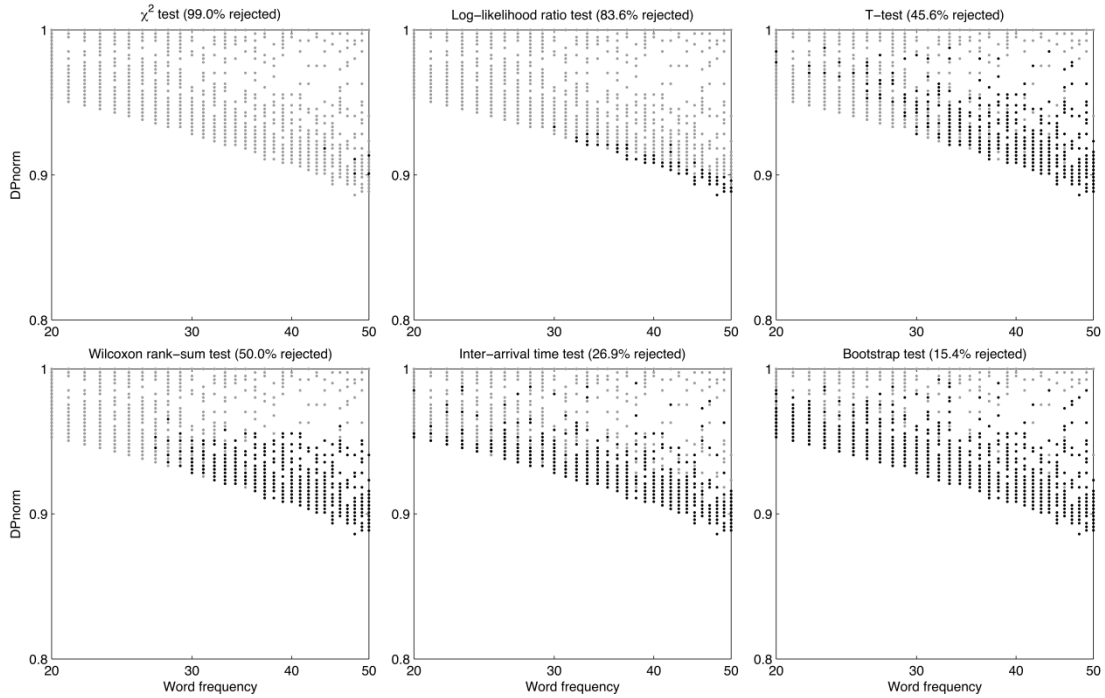

Fig. 3 The results of the uniformity test for all six methods, based on random text assignments, for low frequency words. Each dot corresponds to a word, which has a frequency (x-axis) and dispersion (y-axis). Light grey dots correspond to samples for which the null hypothesis that the p-values follow a uniform distribution has been rejected. The null hypothesis is rejected if the corrected p-value of the Kolmogorov-Smirnov test for uniformity is < 0.01.

5.2 Randomly Assigning the Words to Two Sets

The second experiment that we conducted is based on the random assignment of individual words to two sets rather than the assignment of entire texts. This approach should lead to a smoother distribution of frequencies, and we expect all methods to yield unbiased (i.e. uniform) p-values in this setting. We have used the following procedure to test this hypothesis: we randomly assign half of the 810,000 words to corpus *S* and assign the other half of the words to corpus *T*. We then apply each method

to all words with a frequency of 50 or greater in the fiction corpus (i.e. the same 3,302 words that were used in the previous experiment). The entire process is repeated 500 times. Again, we expect the p-value distribution for each word to be approximately uniformly distributed over the 500 repetitions. We use the Kolmogorov-Smirnov test as discussed above to obtain 3,302 · 6 = 19,812 p-values. We use the Bonferroni correction for multiple hypotheses to compute the final p-values.

Figure 4 shows an overview of the performance of each method.

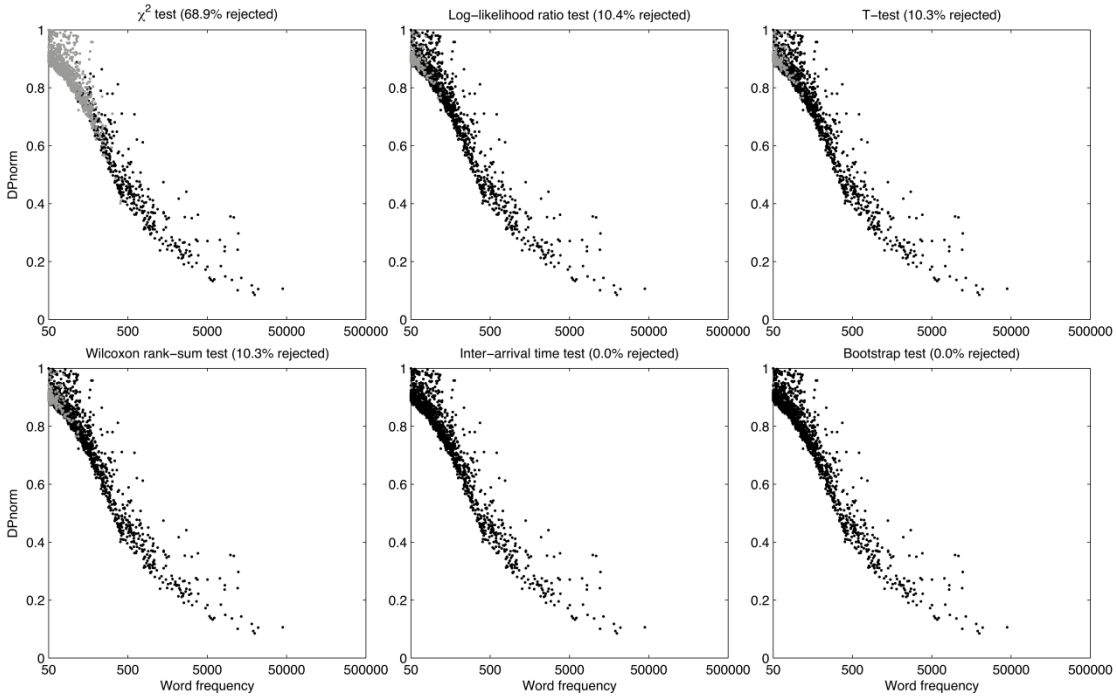

Fig. 4 The results of the uniformity test for all six methods based on random word assignments (rather than texts, as in Fig. 2). Each dot corresponds to a word, which has a frequency (x-axis) and dispersion (y-axis). Light grey dots correspond to samples for which the null hypothesis has been rejected. The null hypothesis is rejected if the corrected p-value of the Kolmogorov-Smirnov test for uniformity is < 0.01.

Surprisingly, we observe that the χ^2 test fails to yield uniform p-values for nearly 70% of the words. This result may have occurred because the test statistic only asymptotically follows a χ^2 distribution, and another contributing factor could be Yates' correction, which makes the p-values more conservative (perhaps excessively

conservative). The latter reason is easy to verify because the Kolmogorov-Smirnov test can also be employed as a one-tailed test. We computed the p-values again by testing only whether the p-values for the frequency test are excessively low. Table 7 presents the results. We now observe that 0% of the samples are rejected; this result confirms that Yates' correction leads to conservative p-values, which is not necessarily a disadvantage.

Table 7 For each method, the percentage of samples for which the null hypothesis under the one-tailed Kolmogorov-Smirnov test is rejected, based on random word assignments as in Fig. 4. The alternative hypothesis is that p-values are anti-conservative.

Test	χ^2 test	Log-likelihood ratio test	Welch's t-test	Wilcoxon rank-sum test	Inter-arrival time test	Bootstrap test
Percentage of rejected samples	0.0%	3.9%	3.9%	3.9%	0.0%	0.0%

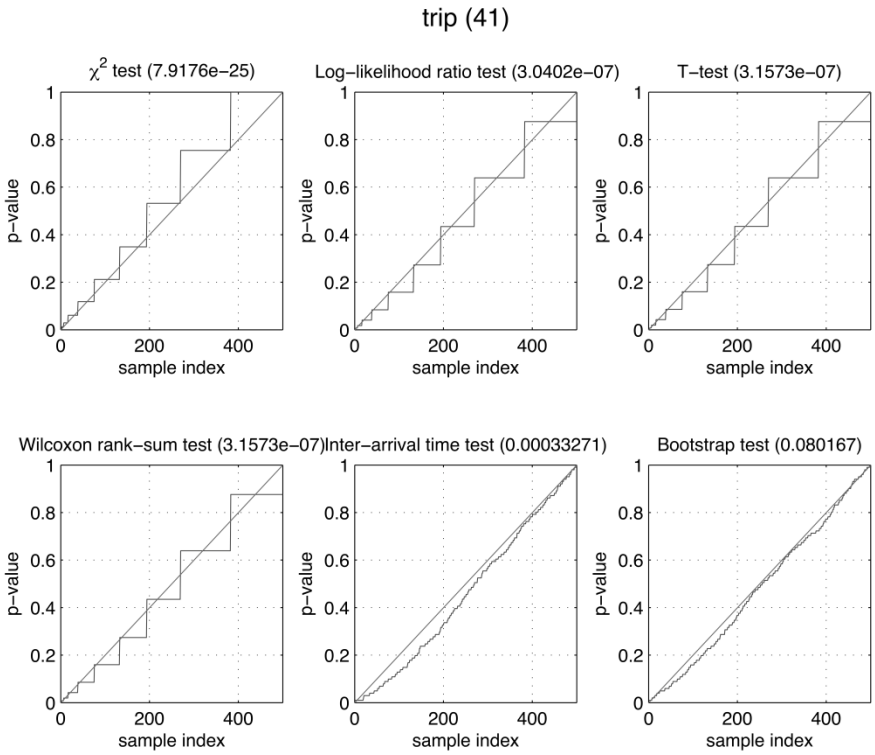

Fig. 5 Cumulative distribution of p-values for each method for the word *trip*. The diagonal line indicates the uniform distribution, which we expect to be close to the actual distribution. The p-values of the uniformity tests are presented in parentheses. The first four tests show a jagged pattern because of the deterministic nature of these tests, i.e. the limited number of different inputs leads to a limited number of different output values. This behaviour causes the uniformity test to yield low p-values. The inter-arrival time and bootstrap tests are less affected by this limitation.

Table 7 also shows that 3.9% of the samples are rejected for the log-likelihood ratio test, t-test, and Wilcoxon rank-sum test despite our use of the conservative Bonferroni correction. Perhaps surprisingly, the inter-arrival time and bootstrap tests have no rejected samples; thus, we can conclude that these tests consistently yield reasonably uniformly distributed p-values. Figure 4 shows that all of the rejected samples are infrequent words. Because this difference is unexpected, let us examine an example of the p-values that are given by each method for an infrequent word.

Figure 5 illustrates the p-values for the word *trip*. We notice a problem here: the first four tests do not yield the expected uniform distributions. The cause is visible in

the figure: the number of unique p-values that these tests yield is limited, and the tests give a similar p-value for many of the randomized inputs, because the number of distinct inputs is also very low. This behaviour is not necessarily unfavourable; if we assume that only a certain number of p-values are possible, then the observed distribution may be 'as uniform as possible' under the constraints. The reference distribution in our test—which is the uniform distribution on [0, 1]—does not assume a finite set of possible values. This distribution could have caused the uniformity test to be slightly inappropriate and to reject many samples, especially those corresponding to infrequent or very poorly dispersed words. Thus, we should not necessarily interpret the smoother curves given by the inter-arrival time and bootstrap tests as superior. However, we are not aware of any significance tests that would be more appropriate in this situation, and we leave this issue for further research.

Figure 6 illustrates a comparison of the p-values for the frequent word *would*. We continue to observe the jagged pattern, but the pattern is now less severe. The high p-values for each of the tests demonstrate that the uniformity test now functions properly. This result corroborates the evidence in Fig. 4 that in this randomization setting (assigning each word in the subcorpus randomly to *S* or to *T*) none of the frequent words is rejected.

We conclude that all of the methods yield uniform p-values in this setting, in which we randomly sample words rather than texts. Thus, the differences between the methods in the first experiment are fully explained by the additional structure of the texts. This finding is important because, when creating a corpus, one usually samples texts from various sources rather than individual words. As a note of caution, the jagged patterns provide the first four tests with a disadvantage in the uniformity test; thus, we

cannot conclude that these four methods are all inferior. Nonetheless, the evidence does not suggest that any test is superior to the *bootstrap test* either. Based on the experiments that have been discussed thus far, we can conclude that under the assumption of randomly sampled texts the χ^2 and log-likelihood ratio tests may lead to spurious conclusions, and we therefore recommend the use of a representation of the data and a statistical test that takes into account the distribution of the word within the corpus.

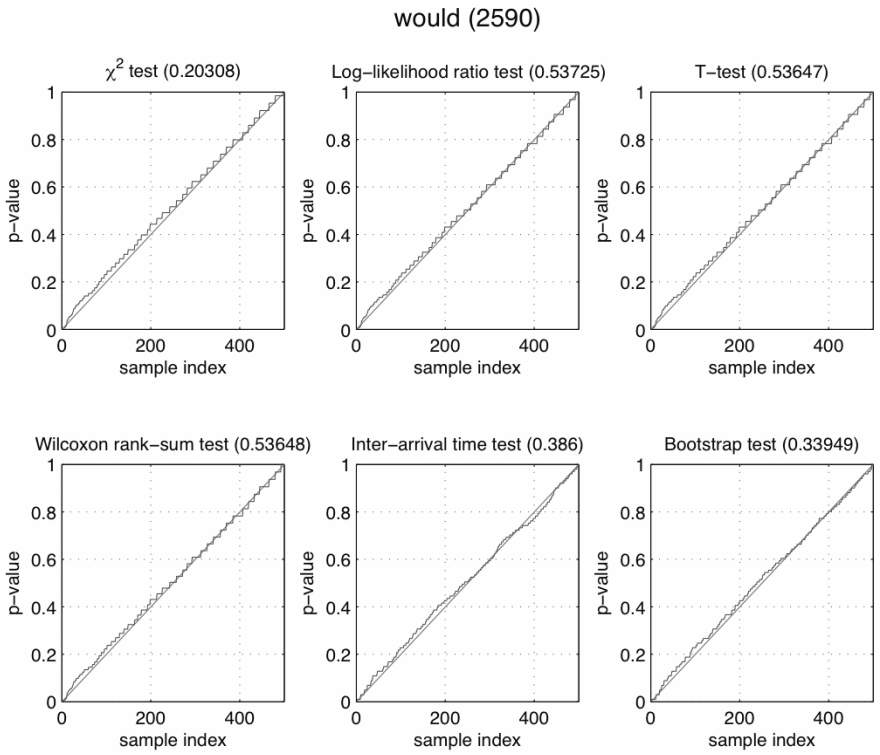

Fig. 6 Cumulative distribution of p-values for each method for the word *would*. The diagonal line indicates the uniform distribution, which we expect to be close to the actual distribution. The p-values of the uniformity tests are presented in parentheses. The first four tests show a jagged pattern because of the deterministic nature of these tests, i.e. the limited number of different inputs leads to a limited number of different output values. Nonetheless, at this frequency, the uniformity test works properly.

6. Differences between Male and Female Writing

6.1 The Bootstrap Test

Past research on the BNC reports statistically significant gender differences in word-frequency distributions in conversation (e.g. Rayson *et al*., 1997) and in both the fiction and non-fiction genres (e.g. Argamon *et al*., 2003). We next consider the extent to which word-frequency distributions display statistically significant gender differences in the BNC prose fiction texts using the bootstrap test.

After we control for a false discovery rate (FDR; Benjamini and Hochberg, 1995) at $\alpha = 0.05$, which controls the expected relative number of false positives over all positives, the bootstrap test returns 74 words (occurring 5,000 times or more in both corpora) whose frequency differs significantly between the male- and female-authored subcorpora. The minimum frequency of 5,000 was chosen for ease of illustration, as the list of significant words would have been considerably longer if lower frequencies had been considered (cf. Fig. 7, below). Tables 8 and 9 list the words that are most significantly overrepresented in male and female prose fiction, respectively.

Table 8 High-frequency words that are significantly overrepresented in male-authored prose fiction in the BNC according to the bootstrap test

Word	Males	M/million	Females	F/million	DP_{norm}	Bootstrap
a	164,254	22,823.55	179,376	21,442.46	0.06	0.0001
another	5,293	735.48	5,285	631.76	0.14	0.0001
by	20,971	2,913.98	20,687	2,472.91	0.13	0.0001
first	7,211	1,001.99	7,145	854.11	0.13	0.0001

from	29,201	4,057.56	29,279	3,499.99	0.10	0.0001
in	103,423	14,370.92	113,461	13,563.04	0.06	0.0001
its	7,031	976.98	5,863	700.86	0.26	0.0001
man	11,533	1,602.54	10,626	1,270.22	0.21	0.0001
of	161,802	22,482.84	165,196	19,747.39	0.09	0.0001
on	54,122	7,520.40	58,075	6,942.24	0.07	0.0001
one	22,641	3,146.03	23,432	2,801.04	0.09	0.0001
some	11,887	1,651.73	11,839	1,415.22	0.14	0.0001
the	417,501	58,012.94	379,234	45,333.32	0.09	0.0001
their	15,044	2,090.41	13,912	1,663.03	0.20	0.0001
they	37,660	5,232.96	35,721	4,270.06	0.17	0.0001
through	9,117	1,266.83	8,300	992.18	0.16	0.0001
two	9,592	1,332.84	8,402	1,004.37	0.17	0.0001
us	6,744	937.10	5,059	604.75	0.26	0.0001
we	26,275	3,650.99	22,273	2,662.50	0.21	0.0001
were	26,899	3,737.69	27,088	3,238.08	0.12	0.0001
is	32,539	4,521.39	30,015	3,587.97	0.21	0.0003
left	5,803	806.34	5,994	716.52	0.14	0.0005
other	8,843	1,228.76	9,170	1,096.17	0.12	0.0005

there	29,585	4,110.92	30,533	3,649.89	0.13	0.0005
are	15,878	2,206.29	15,541	1,857.76	0.18	0.0007
where	9,333	1,296.85	9,596	1,147.10	0.15	0.0013
he	124,464	17,294.62	130,393	15,587.07	0.14	0.0045

Table 9 High-frequency words that are significantly overrepresented in female-authored prose fiction in the BNC according to the bootstrap test

Word	Males	M/million	Females	F/million	DP_{norm}	Bootstrap
'll	9,340	1,297.82	14,921	1,783.64	0.24	0.0001
'm	9,263	1,287.12	14,500	1,733.32	0.24	0.0001
've	8,092	1,124.41	12,258	1,465.31	0.23	0.0001
be	32,481	4,513.33	43,381	5,185.73	0.10	0.0001
come	7,742	1,075.77	10,737	1,283.49	0.15	0.0001
could	20,573	2,858.68	27,724	3,314.10	0.12	0.0001
did	19,633	2,728.06	26,923	3,218.35	0.14	0.0001
eyes	6,955	966.42	12,757	1,524.96	0.26	0.0001
face	7,206	1,001.29	10,427	1,246.44	0.21	0.0001
for	46,664	6,484.09	59,191	7,075.64	0.07	0.0001
go	9,104	1,265.03	12,736	1,522.45	0.16	0.0001
her	49,768	6,915.40	146,675	17,533.41	0.29	0.0001
how	9,714	1,349.79	13,231	1,581.62	0.13	0.0001
if	20,859	2,898.42	27,324	3,266.29	0.11	0.0001
knew	5,700	792.03	8,264	987.87	0.18	0.0001
made	7,094	985.73	9,772	1,168.14	0.13	0.0001

make	5,341	742.15	7,379	882.08	0.13	0.0001
much	6,613	918.89	9,195	1,099.16	0.15	0.0001
must	6,054	841.22	8,325	995.16	0.18	0.0001
n't	45,068	6,262.33	66,842	7,990.24	0.20	0.0001
never	6,969	968.36	10,827	1,294.25	0.17	0.0001
not	33,130	4,603.51	45,580	5,448.60	0.16	0.0001
own	5,403	750.76	8,078	965.64	0.17	0.0001
she	57,200	7,948.10	164,039	19,609.09	0.28	0.0001
should	5,417	752.71	7,962	951.77	0.16	0.0001
so	20,460	2,842.97	29,023	3,469.39	0.12	0.0001
thought	8,753	1,216.25	13,774	1,646.53	0.19	0.0001
to	178,154	24,755.00	223,827	26,756.10	0.05	0.0001
too	8,348	1,159.98	11,448	1,368.48	0.14	0.0001
want	6,050	840.66	8,956	1,070.59	0.20	0.0001
when	17,667	2,454.88	23,864	2,852.68	0.13	0.0001
with	48,613	6,754.91	62,689	7,493.79	0.07	0.0001
would	23,077	3,206.61	32,428	3,876.42	0.14	0.0001
you	79,286	11,017.01	119,301	14,261.14	0.16	0.0001
your	12,257	1,703.14	18,688	2,233.95	0.18	0.0001
had	63,597	8,836.98	85,125	10,175.77	0.15	0.0003
look	6,476	899.86	9,045	1,081.23	0.16	0.0003
take	5,467	759.66	7,181	858.41	0.13	0.0003
very	8,570	1,190.83	12,089	1,445.11	0.22	0.0003
do	28,665	3,983.08	38,382	4,588.15	0.15	0.0005
because	5,599	778.00	8,054	962.77	0.23	0.0007
put	5,415	752.43	7,195	860.08	0.18	0.0023

that	76,457	10,623.91	95,829	11,455.32	0.10	0.0029
little	7,654	1,063.54	10,360	1,238.43	0.19	0.0047
're	8,584	1,192.77	11,813	1,412.12	0.24	0.0049
have	30,736	4,270.85	38,696	4,625.69	0.11	0.0053
well	9,511	1,321.58	12,540	1,499.02	0.18	0.0057

Tables 8 and 9 are consistent with earlier research that has found gender differences based on word frequencies in prose fiction. Overall, the tables suggest that male-authored fiction is dominated by more frequent use of noun-related forms than female-authored fiction, which is verb-oriented. Male authors overuse articles (*a*, *the*) and prepositions (*by*, *from*, *in*, *of*, *on*, *through*), both of which are associated with nouns. Similarly, male-authored fiction overuses other function words that are typically associated with noun phrases and nominal functions, such as *another*, *first*, *one*, *some*, *two*, and *other*. However, it is noteworthy that the list of significant items for male authors is shorter than that for female authors.

The personal pronouns that are overrepresented in male-authored fiction are the first-person plural forms *us* and *we* and the third-person pronouns *its*, *their*, and *they*, while women's fiction overuses the second-person forms *you* and *your*, which can have singular and plural referents. Stereotypically, men tend to write about *man* and *he*, and women about *her* and *she*. These pronoun findings are consistent with those of Argamon *et al*. (2003, pp. 325–327) but deviate in that women do not significantly favour the first-person pronoun *I*, as the previous findings suggest. When the bootstrap method is used, personal pronouns do not emerge as unequivocal female-style markers in contemporary prose fiction.

Table 9 shows that female-authored fiction is marked by frequent verb use: there are more than twenty verb forms among the items overused by women (forms of *be*, *do*, and *have*; modals, such as *could*, *should*, *must*, and *would*; and activity and mental verbs, including *come*, *go*, *make*, *knew*, and *thought*). Only three such verb forms are overused in male-authored fiction (*were*, *is*, and *are*). Particularly salient features in women's fiction are contracted forms (*'ll*, *'m*, *'ve*, *n't*, *'re*), negative particles (*n't*, *never*, *not*), and intensifiers (*much*, *so*, *too*, *very*). These are all indicators that female-authored fiction employs a more involved, colloquial style than male-authored fiction, which, by contrast, is marked by features associated with an informational, noun-oriented style (for these distinctions, see Biber, 1995, pp. 107–120; Biber and Burges, 2000).

However, these style markers may not be a simple reflection of the gender of the authors; rather, these differences may be correlated with target audience differences. Both the male and female authors sampled for the BNC wrote for adults, and only a small minority wrote for children. However, c. 5 million of the total of 7.2 million words in the male-authored fiction subcorpus was intended for a mixed readership, whereas half of the female-authored subcorpus (c. 4.4 million of 8.4 million words) targeted female audiences and may hence include more female characters and female-oriented topics than the male-authored subcorpus. Previous research indicates that audience design is relevant in spoken interaction, and style shifting is typically a response to the speaker's audience (Bell, 1984). In weblogs, for example, the diary subgenre is reported to display more 'female' stylistic features, and the filter subgenre contains more 'male' stylistic features; in both cases the findings are independent of the gender of the author (Herring and Paolillo, 2006). It is plausible that different subgenres

of fiction and their target audiences also play a role in the word-distribution differences that are observed in the BNC prose fiction genre.

6.2 Comparing the χ^2 Test with the Bootstrap Test

The above analysis is based on words that are ranked as significant by the bootstrap test. Most of these words are also significant according to the other tests, including those based on the bag-of-words model. However, how do we evaluate words that are ranked as significant by the bag-of-words tests, such as the χ^2 test, but are considered insignificant by the more valid tests, such as the bootstrap test? Tables 10 and 11 list high-frequency words (occurring 5,000 times or more in both subcorpora) for which the difference between the χ^2 and bootstrap p-values is at least tenfold. By accounting for FDR control at $\alpha = 0.05$, we find that the χ^2 p-values are significant, but the bootstrap p-values are not significant. All of the listed words are also significant according to our other bag-of-words test, the log-likelihood ratio.

Table 10 High-frequency words that are significantly overrepresented in male-authored prose fiction in the BNC according to the χ^2 test but not according to the bootstrap test

Word	Males	M/million	Females	F/million	DP_{norm}	χ^2	Bootstrap
an	18,513	2,572.43	20,422	2,441.23	0.11	0.0000	0.1027
back	17,159	2,384.29	18,863	2,254.87	0.13	0.0000	0.0951
down	14,405	2,001.62	15,483	1,850.83	0.13	0.0000	0.0207
has	6,595	916.39	6,553	783.34	0.26	0.0000	0.0519
his	72,681	10,099.23	76,064	9,092.63	0.16	0.0000	0.0131
I	125,809	17,481.51	141,074	16,863.87	0.20	0.0000	0.5232
into	18,468	2,566.18	20,505	2,451.15	0.12	0.0000	0.1477

my	25,143	3,493.69	24,885	2,974.73	0.30	0.0000	0.0585
off	8,869	1,232.37	9,379	1,121.16	0.15	0.0000	0.0205
old	6,455	896.94	6,895	824.22	0.24	0.0000	0.1931
or	17,248	2,396.66	17,442	2,085.00	0.17	0.0000	0.0139
out	24,466	3,399.62	26,980	3,225.17	0.11	0.0000	0.0749
people	6,342	881.24	6,243	746.28	0.26	0.0000	0.0135
them	18,592	2,583.41	19,973	2,387.56	0.15	0.0000	0.0509
this	24,230	3,366.83	26,699	3,191.58	0.14	0.0000	0.1537
up	25,018	3,476.32	27,754	3,317.69	0.12	0.0000	0.1525
which	13,030	1,810.56	12,809	1,531.18	0.25	0.0000	0.0185
who	14,583	2,026.35	15,619	1,867.08	0.15	0.0000	0.0329
then	19,598	2,723.20	21,899	2,617.79	0.16	0.0001	0.3891
looked	9,904	1,376.19	10,995	1,314.33	0.21	0.0009	0.4287
something	7,457	1,036.17	8,191	979.15	0.17	0.0004	0.1911
just	13,760	1,911.99	15,393	1,840.07	0.19	0.0011	0.4473
turned	5,738	797.31	6,311	754.41	0.18	0.0025	0.2917

Table 11 High-frequency words that are significantly overrepresented in female-authored prose fiction in the BNC according to the χ^2 test but not according to the bootstrap test

Word	Males	M/million	Females	F/million	DP_{norm}	χ^2	Bootstrap
all	25,813	3,586.79	31,323	3,744.33	0.11	0.0000	0.1765
and	184,332	25,613.45	222,854	26,639.78	0.09	0.0000	0.0873
any	7,879	1,094.81	9,837	1,175.91	0.15	0.0000	0.1033
as	45,322	6,297.62	56,365	6,737.83	0.10	0.0000	0.0063
away	8,152	1,132.74	10,130	1,210.93	0.14	0.0000	0.0615
been	20,639	2,867.85	25,253	3,018.72	0.13	0.0000	0.1319

but	42,393	5,890.63	50,780	6,070.20	0.11	0.0000	0.2905
'd	12,340	1,714.68	17,259	2,063.13	0.34	0.0000	0.0565
day	5,369	746.04	6,788	811.43	0.19	0.0000	0.0899
going	7,539	1,047.57	9,628	1,150.92	0.20	0.0000	0.0753
him	34,197	4,751.77	42,555	5,086.99	0.15	0.0000	0.0883
last	5,116	710.88	6,620	791.35	0.16	0.0000	0.0077
might	5,960	828.16	7,630	912.08	0.20	0.0000	0.0655
no	21,170	2,941.63	26,348	3,149.62	0.10	0.0000	0.0093
now	14,568	2,024.26	18,450	2,205.50	0.13	0.0000	0.0141
only	10,668	1,482.35	13,320	1,592.26	0.12	0.0000	0.0239
said	35,208	4,892.25	46,938	5,610.93	0.28	0.0000	0.0681
seemed	5,036	699.77	6,518	779.16	0.23	0.0000	0.0789
think	9,406	1,306.99	12,231	1,462.08	0.17	0.0000	0.0145
time	13,072	1,816.39	16,112	1,926.02	0.10	0.0000	0.0215
told	5,509	765.49	7,455	891.16	0.20	0.0000	0.0065
was	111,268	15,461.00	132,703	15,863.21	0.10	0.0000	0.3401
why	7,034	977.39	8,955	1,070.47	0.16	0.0000	0.0433
room	5,708	793.14	7,107	849.57	0.22	0.0001	0.2215
know	14,188	1,971.46	17,191	2,055.00	0.15	0.0003	0.2985
about	18,742	2,604.25	22,573	2,698.36	0.14	0.0003	0.3357
even	8,156	1,133.30	9,947	1,189.06	0.16	0.0013	0.2625
after	8,541	1,186.80	10,371	1,239.74	0.12	0.0029	0.1553
long	6,326	879.02	7,740	925.23	0.12	0.0026	0.1111
tell	5,557	772.16	6,792	811.91	0.16	0.0057	0.2347

Some of the words in Tables 10 and 11 appear to corroborate the above analysis: the writing style of women is more verb-oriented, whereas men overuse masculine and collective personal pronouns, such as *his* and *them*. However, the list of words for female-authored fiction also includes a male personal pronoun, *him*, and men appear to significantly overuse the first-person singular pronouns *I* and *my*, which is surprising in view of our general knowledge of gendered styles (Argamon *et al*., 2003; Newman *et al*., 2008). Furthermore, men appear to overuse directional adverbs, such as *back*, *down*, *out*, and *up*; this result could be misinterpreted as an interesting discovery with regard to the focus of male prose writing on spatial orientation.

If words of all frequencies are considered, then the most salient category of words that are ranked as significant by the χ^2 test but not by the bootstrap test is proper nouns, as in the *Matilda* example above. Some of these words are also easily misinterpreted as genuine differences between subcorpora. Even an experienced linguist cannot determine which bag-of-words results are genuinely significant; our comparisons show that such results can lead to conflicting interpretations. Therefore, it is advisable to avoid the noise that is inherent in bag-of-words methods and to use a more valid test, such as the bootstrap test.

6.3 Comparing the Tests According to Significance Threshold

Figure 7 summarizes the number of significant words that were returned by each test at varying significance testing thresholds. The t-test yields the least number of significant words, followed by the Wilcoxon rank-sum and bootstrap tests in both figures. Only the curve for the inter-arrival time test differs substantially between Figs 7a and 7b. The test appears to have difficulty with comparing zero with non-zero frequencies and always deems such cases significant. We also observe that the χ^2 and log-likelihood ratio tests

yield more words (by several orders of magnitude) as significant results than the t-test and the Wilcoxon rank-sum and bootstrap tests. Both axes have a logarithmic scale.

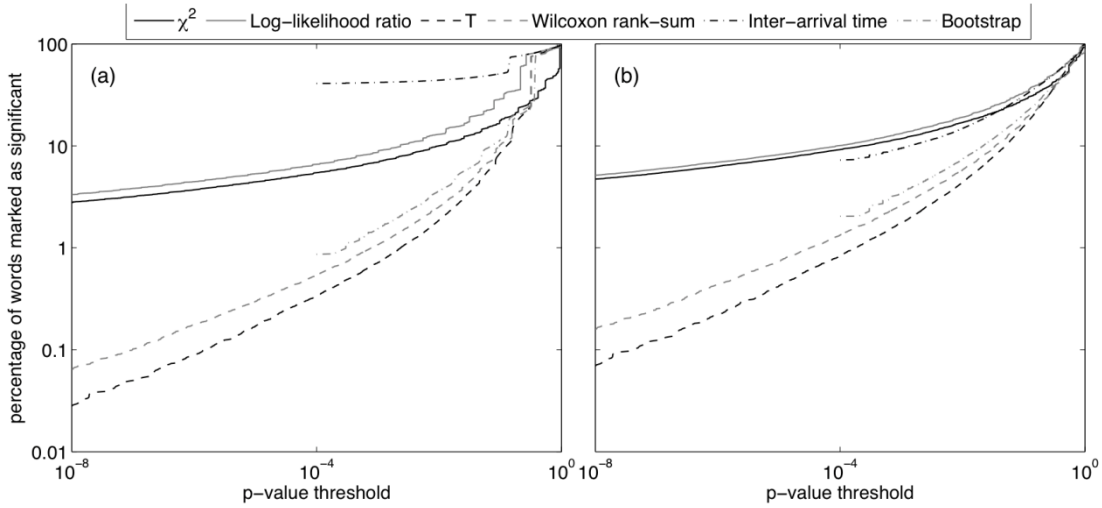

Fig. 7 Comparison of the number of significant words for the six methods. For each method, a curve demonstrates how the number of significant words increases as we increase the significance threshold in the male vs. female author comparison without correcting for multiple hypotheses. The x-axis shows the p-value threshold, and the y-axis shows the percentage of words that are marked as having significantly different frequencies between genders. The figure on the left (a) is based on all words in the prose fiction subcorpus, and the figure on the right (b) includes only those words with frequencies greater than zero for both genders.

7. Conclusion

Many current corpus tools use the χ^2 and log-likelihood ratio tests. We suggest that other tests be added to these tools for the reasons discussed in this paper. The core difference between the bag-of-words tests (the χ^2 and log-likelihood ratio tests) and the other four tests (the t-test and the Wilcoxon rank-sum, inter-arrival time, and bootstrap tests) is the representation of the data, and thus, the unit of observation: for the bag-of-words tests, the data are represented in a 2x2 table (Table 1) and the number of samples equals the number of words in a corpus, whereas for the other four tests, the data are

represented either by a frequency list (Table 2), or a list of inter-arrival times. In those cases, the number of samples is much lower than the number of words in a corpus.

For the t-test, the Wilcoxon rank-sum test, and the bootstrap test, the number of samples equals the number of texts in a corpus, and for the inter-arrival time test, the number of samples equals the number of occurrences of the word being tested rather than the total number of words. The number of samples generally determines our certainty with regard to the estimates, and the experimental results show that the bag-of-words tests have excessively high confidence in the estimates of mean word frequencies, in the context of the statistical comparison of two corpora.

By studying the uniformity of the p-values that were given by each of the tests, we have shown that the choice of how to define independent samples and how to represent the data plays a major role in the outcome of a significance test. We have shown that bag-of-words-based tests may lead to spurious conclusions when assessing the significance of differences in frequency counts between corpora. Note, however, that we are not suggesting that there is anything wrong with the χ^2 and log-likelihood ratio tests as such, but only that their application in this context is problematic. We have also shown that appropriate alternatives exist: Welch's t-test, the Wilcoxon rank-sum test, and the bootstrap test.

We have considered the choice of statistical tests for comparing moderate-sized or large corpora (at least 100 texts each). Due to space limitations, we have not include discussion on how to compare small corpora. This problem is briefly addressed in Lijffijt *et al.* (2012). It appears that the advice on which statistical test to use is not as straightforward as for large corpora. The objections made in this paper against the bag-of-words test hold for corpora of any size. However, in small corpora, counting all

occurrences of a word in the same text as one sample, i.e., a sample equals a text, may preclude the detection of many patterns. We would expect the inter-arrival time test to be a tempting alternative in that setting, but further investigation into the use of statistical tests for comparing small corpora or individual texts is warranted.

Notes

[1] Kilgarriff refers to this test as the Mann-Whitney ranks test.

[2] In Lijffijt *et al.* (2012) we set out to explore the question of lexical variation in a historical single-genre corpus of personal correspondence over time. Comparing the log-likelihood ratio and bootstrap tests, we found that the two successive half-a-million-word subperiods of the corpus that we examined were more similar to each other with regard to their lexis than a bag-of-words method might lead one to postulate. We also observed that, besides the choice of method and the size of the corpus, the observed degree of similarity depends on several other factors, notably, the type of post-hoc correction, and the frequency cut-off and significance thresholds used.

[3] Both p-values are actually 0 using double precision floating point numbers; thus, these values are much smaller than 0.0001.

Acknowledgements

We thank the anonymous reviewers for their valuable comments and suggestions.

Funding

This work was supported by the Academy of Finland [grant numbers 129282, 129350]; the Finnish Centre of Excellence for Algorithmic Data Analysis (ALGODAN); the Finnish Centre of Excellence for the Study of Variation, Contacts and Change in

English (VARIENG); the Finnish Doctoral Programme in Computational Sciences (FICS); the Academy of Finland's Academy professorship scheme; and the Finnish Graduate School in Language Studies (Langnet).

References

Altmann, E. G., Pierrehumbert, J. B., and Motter, A. E. (2009). Beyond word frequency: bursts, lulls, and scaling in the temporal distributions of words, *PLoS One*, **4**(11): e7678.

Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts, *Text*, **23**(3): 321–46.

Beaujean, F., Caldwell, A., Kollár, D., and Kröninger, K. (2011). P-values for model evaluation, *Physical Review D*, **83**(1): 012004.

Bell, A. (1984). Language style as audience design, *Language in Society*, **13**: 145–204.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society*, **57**(1): 289–300.

Berry, G. and Armitage, P. (1995). Mid-P confidence intervals: a brief review, *The Statistician*, **44**(4): 417–23.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.

Biber, D. and Burges, J. (2000). Historical change in the language use of women and men: gender differences in dramatic dialogue, *Journal of English Linguistics*, **28**(1): 21–37.

Blocker, C., Conway, J., Demortier, L., Heinrich, J., Junk, T., Lyons, L., and Punzig, G. (2006). Simple facts about p-values, Technical Report

CDF/MEMO/STATISTICS/PUBLIC/8023, Laboratory of Experimental High Energy Physics, The Rockefeller University.

BNC = The British National Corpus, version 3 (BNC XML Edition) (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/ (accessed 26 November 2012).

Burnard, L. (2007). *Reference Guide for the British National Corpus (XML Edition)*. Published for the British National Corpus Consortium by the Research Technologies Service at Oxford University Computing Services. http://www.natcorp.ox.ac.uk/docs/URG/ (accessed 26 November 2012).

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments, *Statistical Science*, **18**(1): 71–103.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence, *Computational Linguistics*, **19**: 61–74.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. New York: Chapman and Hall/CRC.

Evert, S. (2005). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für Maschinelle Sprachverarbeitung, University of Stuttgart.

Good, P. (2005). *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. 3rd edn., New York/Heidelberg: Springer.

Gries, S. Th. (2005). Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff, *Corpus Linguistics and Linguistic Theory*, **1**(2): 277–94.

Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora, *International Journal of Corpus Linguistics*, **13**(4): 403–37.

Herring, S. C. and Paolillo, J. C. (2006). Gender and genre variation in weblogs, *Journal of Sociolinguistics*, **10**(4): 439–59.

Hinneburg, A., Mannila, H., Kaislaniemi, S., Nevalainen, T., and Raumolin-Brunberg, H. (2007). How to handle small samples: bootstrap and Bayesian methods in the analysis of linguistic change, *Literary and Linguistic Computing*, **22**(2): 137–50.

Hoffmann, S., Evert, S., Smith, N., Lee, D., and Berglund Prytz, Y. (2008). *Corpus Linguistics with BNCweb—a Practical Guide*. Frankfurt am Main: Peter Lang.

Kilgarriff, A. (2001). Comparing corpora, *International Journal of Corpus Linguistics*, **6**(1): 1–37.

Kilgarriff, A. (2005). Language is never, ever, ever, random, *Corpus Linguistics and Linguistic Theory*, **1**(2): 263–76.

L'Ecuyer, P. and Simard, R. (2007). TestU01: a C library for empirical testing of random number generators, *ACM Transactions on Mathematical Software*, **33**(4): 22.

Lee, D. Y. W. (2001). Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle, *Language Learning & Technology*, **5**(3): 37–72.

Lijffijt, J. (2012). Bootstrap test for R and Matlab. http://users.ics.aalto.fi/lijffijt/bootstraptest/ (accessed 26 November 2012).

Lijffijt, J. (2013). A fast and simple method for mining subsequences with surprising event counts. In Blockeel, H., Kersting, K., Nijssen, S., and Železný, F. (eds), *Proceedings of ECML-PKDD 2013—Part I*. Berlin: Springer-Verlag, pp. 385–400.

Lijffijt, J. and Gries, S. Th. (2012). Correction to Stefan Th. Gries' "Dispersions and adjusted frequencies in corpora", *International Journal of Corpus Linguistics*, **17**(1): 147–9.

Lijffijt, J., Papapetrou, P., Puolamäki, K., and Mannila, H. (2011). Analyzing word frequencies in large text corpora using inter-arrival times and bootstrapping. In Gunopulos, D., Hofmann, T., Malerba, D., and Vazirgiannis, M. (eds), *Proceedings of ECML-PKDD 2011—Part II*. Berlin: Springer-Verlag, pp. 341–57.

Lijffijt, J., Säily, T., and Nevalainen, T. (2012). CEECing the baseline: lexical stability and significant change in a historical corpus. In Tyrkkö, J., Kilpiö, M., Nevalainen, T., and Rissanen, M. (eds), *Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources*. Studies in Variation, Contacts and Change in English, Vol. 10. Helsinki: VARIENG. http://www.helsinki.fi/varieng/journal/volumes/10/lijffijt_saily_nevalainen/ (accessed 26 November 2012).

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics*, **18**(1): 50–60.

Massey, F. (1951). The Kolmogorov-Smirnov test for goodness of fit, *Journal of the American Statistical Association*, **46**(253): 68–78.

Newman, M. L., Groom, C. J., Handelman, L. D., and Pennebaker, J. W. (2008). Gender differences in language use: an analysis of 14,000 text samples, *Discourse Processes*, **45**: 211–36.

North, B. V., Curtis, D., and Sham, P. C. (2002). A note on the calculation of empirical p-values from Monte Carlo procedures, *The American Journal of Human Genetics*, **71**(2): 439–41.

Oakes, M. P. and Farrow, M. (2007). Use of the chi-squared test to examine vocabulary differences in English-language corpora representing seven different countries, *Literary and Linguistic Computing*, **22**(1): 85–100.

Paquot, M. and Bestgen, Y. (2009). Distinctive words in academic writing: a comparison of three statistical tests for keyword extraction. In Jucker, A., Schreier, D., and Hundt, M. (eds), *Corpora: Pragmatics and Discourse*. Amsterdam: Rodopi, pp. 247–69.

Rayson, P. (2008). From key words to key semantic domains, *International Journal of Corpus Linguistics*, **13**(4): 519–49.

Rayson, P., Berridge, D., and Francis, B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In Purnelle, G., Fairon, C., and Dister, A. (eds), *Le poids des mots: Proceedings of the 7th International Conference on Statistical Analysis of Textual Data (JADT 2004)*. Louvain-la-Neuve: Presses universitaires de Louvain, pp. 926–36.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In Kilgarriff, A. and Berber Sardinha, T. (eds), *Proceedings of the Workshop on Comparing Corpora*. Stroudsburg: Association for Computational Linguistics, pp. 1–6.

Rayson, P., Leech, G., and Hodges, M. (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational component of the

British National Corpus, *International Journal of Corpus Linguistics*, **2**(1): 133–52.

Savický, P. and Hlaváčová, J. (2002). Measures of word commonness, *Journal of Quantitative Linguistics*, **9**(3): 215–31.

Schweder, T. and Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously, *Biometrika*, **69**(3): 493–502.

Scott, M. (2012). WordSmith Tools, version 6. Liverpool: Lexical Analysis Software.

Shaffer, J. P. (1995). Multiple hypothesis testing, *Annual Review of Psychology*, **46**: 561–84.

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved, *Biometrika*, **34**(1–2): 28–35.

Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics Bulletin*, **1**(6): 80–3.