Extending the Cochran rule for the comparison of word frequencies between corpora

Paul Rayson[1], Damon Berridge[2], Brian Francis[2]

[1]Computing Department, Lancaster University, Lancaster, LA1 4YR, UK.
paul@comp.lancs.ac.uk
[2]Centre for Applied Statistics, Lancaster University, LA1 4YF, UK.

Abstract

We first describe a number of inter-related issues that need to be considered by the researcher when comparing frequencies of linguistic features in two or more corpora. We then describe the chi-squared and log-likelihood tests used in previous research for the comparison of word frequencies. Our focus, in this paper, is on the issue of reliability of the statistical tests, and we describe simulation experiments to compare the reliability of the chi-squared and log-likelihood statistics under conditions of different-sized corpora and probability of a word occurring in text. We observe that the Cochran rule provides a good guide to accuracy of both statistics in general, but in some cases it needs to be extended. We conclude by recommending higher cut-off values for the Cochran rule at the 5%, 1% and 0.1% levels. In order to extend applicability of the frequency comparisons to expected values of 1 or more, use of the log-likelihood statistic is preferred over the chi-squared statistic, at the 0.01% level. The trade-off for corpus linguists is that the new critical value is 15.13.

Keywords: word frequency, chi-squared test, log-likelihood test, corpus linguistics.

1. Introduction

In recent years corpus-based techniques have increasingly been used to examine issues in language variation, that is, to compare language usage across corpora, users, genres, etc. Comparison of one-million word corpora is becoming common even for beginners in corpus linguistics, with the increasing availability of corpora and the reasoning that one million words gives sufficient evidence for mid- to high-frequency words. However, with the production of large corpora such as the British National Corpus (BNC) containing one hundred million words (Aston and Burnard, 1998), frequency comparisons are available across several millions of words of text (Leech, Rayson and Wilson, 2001). Sufficient data for the investigation of relatively infrequent phenomena is still problematic. However, research is still continuing on small corpora (Ghadessy, Henry and Roseberry, 2001).

There are two main types of corpus comparison:

- A: comparison of a sample corpus with a large(r) standard corpus (e.g. Scott, 2000)
- B: comparison of two (roughly-) equal sized corpora (e.g. Granger, 1998)

There are also a number of inter-related issues that need to be considered when comparing two (or more) corpora[1]: representativeness, homogeneity within the corpora, comparability of the corpora, and choice and reliability of statistical tests (for different sized corpora and other factors).

[1] Alongside practical issues such as the cost and time taken in obtaining, or collecting such corpora.

In the first type (A), we refer to the large(r) corpus as a 'normative' corpus since it provides a text norm (or general language standard) against which we can compare. These two main types can be extended to the comparison of more than two corpora. For example, we may compare one normative corpus to several smaller corpora at the same time, or compare three or more equal sized corpora with each other. In general, however, this makes the results more difficult to interpret. Biber (1993: 243) states that 'a corpus must be representative in order to be appropriately used as the basis for generalizations concerning a language as a whole'. Representativeness, in this sense, is seen as a particularly important attribute for a normative corpus when comparing a sample corpus to a 'general language' corpus (such as the BNC) that contains sections from many different text types and domains. To be representative of the language as a whole, a corpus should contain samples of all major text types (Leech, 1993) and, if possible, be in some way proportional to their usage in 'every day language' (Clear, 1992). This first type of comparison (A) is intended to discover features in the research corpus which have significantly different usage (i.e. frequency) to that found in 'general' language. Representativeness can also apply to specialised corpora, whenever the researcher wants to make a claim about language use in a particular genre or domain, rather than the language as a whole.

The second type of comparison (B) is one that views corpora as equals. It aims to discover features in the corpora that distinguish one from another. Homogeneity (Stubbs, 1996: 152) within each of the corpora is important here since we may find that the results reflect sections within one of the corpora that are unlike other sections in either of the corpora under consideration (Kilgarriff, 1997). Comparability is of interest too, since the corpora should have been sampled for in the same way; in other words, the same stratified sampling method and with, if possible, randomised methods of sample selection. This is the case with the Brown and LOB corpora (introduced below), since LOB was designed to be comparable to the Brown corpus, and neither corpus was designed to be homogeneous.

The final issue, which we address in this paper, is the one regarding the reliability of the statistical tests in relation to the size of the corpora under consideration. In the next section, we will describe two of the statistical tests previously used, chi-squared and log-likelihood.

2. Background

One of the largest early studies was the comparison of one million words of American English (Brown corpus) with one million words of British English (LOB corpus) by Hofland and Johansson (1982). A statistical goodness-of-fit test, the chi-squared test ($\chi 2$), was used to compare word frequencies across the two corpora. The chi-squared test statistic was calculated as follows:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \quad \text{where} \quad E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

where Oi is the observed frequency, Ei is the expected frequency, and Ni is the total frequency in corpus i (i in this case takes the values 1 and 2 for the LOB and Brown corpora respectively). Hofland and Johansson marked any resulting chi-squared values that indicated that a statistically significant difference at the 5%, 1%, or 0.1% level had been detected between the frequency of a word in American English and in British English. In some cases the expected frequency in at least one of the two corpora was too low for their calculation. The null hypothesis of the test is that there is no difference between the observed frequencies

of a word in the two corpora. Note that even if the null hypothesis is not rejected, we cannot conclude that it is true. The cut-off value corresponding to the chosen degree of confidence may not be exceeded, but this only indicates there is not enough evidence to reject the null hypothesis. Critical values for the chi-squared statistic are listed in statistical tables such as those in Barnett and Cronin (1986). For example, the critical value for the 5% level, usually shown as 0.05 in the tables, is 3.84 at 1 degree of freedom. Pearson (1904) originally suggested the chi-squared test using the statistic described above for testing the independence of two variables. It is applicable to a general two-dimensional contingency table with r rows and c columns. The number of degrees of freedom (d.f.), which is used when looking up critical values, is the number of independent terms given that the marginal totals in the table are fixed. In corpus linguistics, we usually use a 2×2 table to compare frequencies of words or other linguistic features between two corpora, so d.f. as calculated by (r-1)×(c-1) is equal to 1. In this specific case, the 2×2 contingency table is as shown in Table 1.

Table 1. Contingency table for the chi-squared test

	Corpus one	**Corpus two**	**Total**
Frequency of feature	a	b	a+b
Frequency of feature not occurring	c	d	c+d
TOTAL	a+c	b+d	N=a+b+c+d

Hence, we can calculate the chi-squared statistic (χ^2) as follows:

$$\chi^2 = N(ad-bc)^2/((a+b)(c+d)(a+c)(b+d))$$

Kilgarriff (1996a, 1996b and 2001) points out that in the Brown versus LOB comparison many common words are marked as having significant chi-squared statistics, and that because words are not selected at random in language we will always see a large number of differences in two such text collections. He selects the Mann-Whitney test that uses ranks of frequency data rather than the frequency values themselves to compute the statistic. The drawback of the Mann-Whitney test is that it omits 92% of the types in the joint corpus, as reported in Rayson (2003). Numerous other authors have used the chi-squared test to determine significant frequency differences of individual words or other linguistics features, between two corpora (for example Woods *et al*., 1986: 140; Virtanen, 1997; Oakes, 1998: 26; Roland *et al.*, 2000). It is widely understood that the chi-squared statistic becomes unreliable when the expected frequency is too small, i.e. less than 5 (Butler 1985: 117; Woods *et al.*, 1986: 144), and sometimes the same limit is applied to the observed frequencies (De Cock, 1998; Nelson *et al*., 2002: 277). Cochran (1954) suggested a rule that 4 in 5 (80%) of the expected values in an r×c table should be 5 or more. In the 2×2 table case, this means all cells should have expected values of 5 or more. Everitt (1992: 39) cites other more recent work than Cochran which suggests that this rule is too conservative. Butler (1985: 117) suggests a solution to this is to combine frequencies until the combined classes have an expected frequency of 5 or more, likewise Nelson et al (2002: 277) for the observed frequencies, but Everitt (1992: 41) argues against this practice.

Dunning (1993) reports that we should not rely on the assumption of a normal distribution when performing statistical text analysis and suggests that parametric analysis based on the binomial or multinomial distributions is a better alternative for smaller texts. Hence Dunning

proposes the log-likelihood ratio as an alternative to Pearson's chi-squared test, and he demonstrates this for the extraction of significant bigrams from text. Everitt (1992: 72) also mentions that the chi-squared statistic is "easily shown to be an approximation to" the log-likelihood for large samples. The two statistics take similar values for many tables. Cressie and Read (1984) show that Pearson's χ^2 (chi-squared) and the likelihood ratio G^2 (Dunning's log-likelihood) are in fact two statistics in a continuum defined by the power-divergence family of statistics. Scott (2001) uses the log-likelihood statistic in his keywords procedure. For the 2×2 case (in Table 1), the log-likelihood ratio is calculated as follows:

$$G^2 = 2 (alna + blnb + clnc + dlnd + NlnN - (a+b)ln(a+b) - (a+c)ln(a+c) - (b+d)ln(b+d) - (c+d)ln(c+d))$$

Other statistical tests are described in the literature, for example using the Yates' correction to the chi-squared test and Fisher's exact test, and these and other tests are applied in the information retrieval domain. For reasons of space, we omit descriptions of them here, but they are reviewed in more detail in Rayson (2003). In a different application area, that of extracting groups of associated words, Weeber et al (2000) use a combination of the log-likelihood ratio and Fisher's exact test for the full word frequency range. The 2×2 tables for comparing word frequencies across corpora are, in general, not as heavily skewed as for bigram data, so the log-likelihood ratio is suitable for our use. We can now examine situations where the contingency table may become skewed.

The research question we attempt to answer in this paper is as follows. Can we rely on the generally accepted Cochran rule for large variations in corpus size, frequency of words, and ratio of size of the two corpora under comparison? The usual approach for a linguist is to avoid comparisons for low frequency words and in a non-normative test to make the corpora the same size. We wish to test whether we can perform statistically reliable comparisons of corpora of different sizes, and include a larger portion of the frequency profile in our tests.

3. Methodological approach

In order to test the reliability of the chi-squared and the log-likelihood statistics and the prediction ability of the Cochran rule we carried out a large number of simulation experiments (i.e. with simulated contingency tables)[2]. For each experiment, there were three experimental conditions which determined the characteristics of the comparison between the corpora. These were:

1. The ratio of the sizes of the two corpora (R). We assume that one corpus is the normative corpus, and therefore the comparison corpus is less than or equal in size to the normative corpus. Seven different ratio values were used, from 1:1 to 200:1.
2. The probability of the word occurring in text (p). This probability was allowed to vary from 1 in 500 to 1 in 1,000,000 to reflect a good range observed in the BNC. To give context to these values, the probability of 'the' occurring in standard British English is 1/16, the probability of 'and' is 1/37 and the probability of 'reliable' is 1/44823 (estimated from the BNC). Ten different probability values were used. We assumed that the probability of the target word was the same in the normative text and in the com-

[2] The GLIM4 statistical software package was used to do this (Francis, Green and Payne, 1993) but the simulations can be coded in any programming language with access to a reliable random-number generator.

parative text, as we wished to test the distribution of the test statistics χ^2 and G^2 under the null hypothesis of no difference in probabilities.

3. The size of the normative corpus (C). This was allowed to vary from 100,000 to 100 million words, taking eight different values. The BNC is at the top end of this range, with 100 million words.

We assumed that the probability of the target word was the same in the normative text and in the comparative text. For each experiment, C, R, and p were all fixed, and 10,000 simulated 2×2 tables of word frequencies were generated. Table 2 below shows the expected values of the 2×2 table. For each simulated table, the chi-squared test statistic χ^2 and the log-likelihood test statistic G^2 were calculated, and at the end of each experiment the χ^2 statistics were ranked in ascending order; the G^2 statistics were similarly ranked. The 95th, 99th, 99.9th and 99.99th percentiles were then determined. The experiment was replicated one hundred times to obtain empirical estimates of the standard deviation of the percentiles.

Table 2. Expected values of the 2×2 tables considered in the experiment

	Normative corpus	Comparative corpus	Total
No. of target words	pC	pRC	p(1+R)C
Number of non-target words	(1-p)C	(1-p)RC	(1-p)(1+R)C
Total number of words	C	RC	(1+R)C

Finally, for each experiment, we determined whether the usual Cochran rule would determine if the simulated values were unreliable. As the rule is based on expected values, and not actual simulated values, this could be determined once at the beginning of each experiment. From Table 2, the smallest expected value is pRC (as $P < 1$ and $R \leq 1$ by design). Thus, if pRC was less than 5, the Cochran rule was held to be true. A total of 560 (7×10×8) experiments were carried out, covering each possible combination of the three experimental factors. With one hundred replicates of each experiment, the entire study generated 560,000,000 (10,000×100×560) simulated 2×2 tables.

4. Results and discussion

Table 3 shows some typical output from the simulation experiments, giving the results for the 95th percentile of the χ^2 test statistic where p=1/16000. For each combination of normative corpus size C and ratio R, the table entries give the mean and standard deviation (in brackets) of the 100 simulated values of the test statistic. We define the test statistic to be accurately represented by the chi-squared distribution if the 95th percentile of the chi-squared distribution on 1 degree of freedom (3.840) is contained in the interval defined by the mean plus or minus two standard deviations. Cells of the table in bold font show where this condition does not hold. It can immediately be observed that the test statistic is accurate in most of the table, but not in the top left hand corner. This corner is characterised by small expected values in the 2×2 table, and examination of Table 4, which gives the smallest expected value, indicates that the Cochran condition provides a good guide to accuracy in this case. Cells of the table in bold font show where the Cochran rule is true.

Table 3. Means and standard deviations of the 95th simulated percentile of the chi-squared test statistic under independence for various 2×2 tables with p=1/16000[3]

p=1/16000	Ratio R						
Corpus size C	**1 / 200**	**1 / 100**	**1 / 20**	**1 / 10**	**1 / 5**	**1 / 2**	**1 / 1**
100,000	**0.060**	**8.772**	3.718	4.068	**3.410**	**3.528**	3.752
	(0.000)	**(0.664)**	(0.206)	(0.192)	**(0.098)**	**(0.069)**	(0.062)
500,000	**5.0945**	**2.615**	**3.400**	**3.605**	3.763	3.836	3.831
	(0.070)	**(0.100)**	**(0.119)**	**(0.037)**	(0.074)	(0.081)	(0.070)
1,000,000	**2.253**	**3.505**	**3.523**	3.736	3.805	3.828	3.823
	(0.069)	**(0.064)**	**(0.055)**	(0.059)	(0.070)	(0.081)	(0.074)
5,000,000	**3.707**	**3.324**	3.801	3.811	3.831	3.850	3.845
	(0.048)	**(0.034)**	(0.075)	(0.075)	(0.068)	(0.076)	(0.084)
10,000,000	**3.244**	3.739	3.813	3.834	3.840	3.842	3.837
	(0.019)	(0.101)	(0.067)	(0.065)	(0.079)	(0.065)	(0.082)
20,000,000	3.720	3.752	3.813	3.842	3.851	3.847	3.854
	(0.073)	(0.077)	(0.073)	(0.076)	(0.070)	(0.070)	(0.078)
50,000,000	3.773	3.809	3.829	3.833	3.849	3.836	3.845
	(0.047)	(0.071)	(0.075)	(0.072)	(0.081)	(0.074)	(0.080)
100,000,000	3.820	3.826	3.835	3.843	3.844	3.845	3.831
	(0.068)	(0.072)	(0.071)	(0.067)	(0.074)	(0.071)	(0.077)

Table 4. Smallest expected values in the 2×2 tables when p=1/16000

p=1/16000							
Corpus size	**Ratio R**						
C	**1 / 200**	**1 / 100**	**1 / 20**	**1 / 10**	**1 / 5**	**1 / 2**	**1 / 1**
100,000	**0.031**	**0.063**	**0.313**	**0.625**	**1.250**	**3.125**	6.250
500,000	**0.156**	**0.313**	**1.563**	**3.125**	6.250	15.625	31.250
1,000,000	**0.313**	**0.625**	**3.125**	6.250	12.500	31.250	62.500
5,000,000	**1.563**	**3.125**	15.625	31.250	62.500	156.250	312.500
10,000,000	**3.125**	6.250	31.250	62.500	125.000	312.500	625.000
20,000,000	6.250	12.500	62.500	125.000	250.000	625.000	1250.000
50,000,000	15.625	31.250	156.250	312.500	625.000	1562.500	3125.000
100,000,000	31.250	62.500	312.500	625.000	1250.000	3125.000	6250.000

[3] Standard deviations are in parentheses. Cells where the 95% critical value of the chi-squared distribution on 1 degree of freedom (3.84) lies outside interval defined by the mean plus or minus two standard deviations are defined as inaccurate and shown in bold.

Of course, Table 3 is a small portion of the output generated. Similar tables were generated for G^2 and χ^2; for the 99th, 99.9th and 99.99th percentiles as well as the 95th percentile, and for each of the ten values of the proportion p. Table 5 summarises the results into a single display. The table consists of an array of ten rows and five columns. The first four columns show the accuracy of the tests at the 5%, 1%, 0.1% and 0.01% significance level, and the final column the standard Cochran rule. Each of the ten rows represents a different proportion. Within a cell of the array, an 8×7 grid shows the simulated accuracy. Rows of each grid represent the eight corpus sizes in ascending order, and the columns of the grid represent the seven ratios, again in ascending order. The symbols used are: ■ both tests inaccurate; ◘ chi-squared test inaccurate; ▣ likelihood ratio test inaccurate; □ both tests accurate. A ✖ indicates that the smallest expected value of the generated table is less than 5. The critical values we used were 3.84 (5% level), 6.63 (1% level), 10.83 (0.1% level) and 15.13 (0.01% level which is not usually listed in published tables).

It can be seen from Table 5 that the overall pattern in each 8×7 grid is similar to that in Table 3. The chi-squared and likelihood ratio test statistics are accurate in much of each grid apart from the top left hand corners. There is also an observable trend as we move down the ten rows from the top to the bottom of the overall table. The trend moving down the overall table is for fewer cells in each 8×7 grid to be marked indicating fewer inaccurate tests as the proportion increases. A trend moving from left to right (from 5%, 1% to 0.1% and 0.01%) in the overall table is less easy to detect. For all the proportions, there are slightly fewer inaccurate tests observed at the 99.99th percentile (0.01%) than at the 95th percentile (5%) level.

Let us now contrast the accuracy of the two tests. When one test rather than both is inaccurate, and this occurs in 246 grid cells, the likelihood ratio test is inaccurate 81 times, and the chi-squared test is inaccurate 165 times. There is an overall trend from left to right which distinguishes the two tests. It is most clear at the 0.01% level in which, with the exception of 25 cells, the likelihood ratio is accurate in all the cells. This compares with 120 cells where the chi-squared test is inaccurate. Examining the data underlying the results at the 0.01% level in particular, the simulations for the likelihood ratio showed much smaller standard deviations than for the chi-squared test, as well as the mean statistic for the chi-squared test being skewed. The likelihood ratio test remains accurate for very heavily skewed tables at the 0.01% level, even for example, in one of the cells in the first grid at the 0.01% level represents a simulation of a word occurring once in a 1,000,000 word corpus as compared to the word occurring 100 times in a 100 million word corpus. The cells are the top right of the grids at the 0.01% level are false positives since the simulated critical values fall below the listed one.

If for each row we compare the four grids showing the accuracy of the tests to the grid showing the Cochran rule, we can see that the Cochran rule provides a good guide to accuracy in most cases. However, there are some cases (marked with a shaded background) which show inaccurate tests that are not covered by the Cochran rule. These are cases where the smallest expected value in the generated table is 5 or greater. The cells in question show inaccuracy in both chi-squared and likelihood ratio tests. At the 5% level the largest expected value showing an inaccurate test is 12.5. At the 1% level we observe a value of 10, at the 0.1% level it is 7.8125 and at the 0.01% level the largest expected value coinciding with an inaccurate test is 6.25. This suggests that the Cochran rule needs to be extended.

The statistics have more evidence from larger corpora and can therefore detect smaller differences in frequency. It is a feature of the tests that values are greater in larger corpora. So, for example, a χ^2 value of 500 obtained from a 1 million versus 9 million word normative test is

not comparable with the same χ^2 value of 500 obtained from a 10,000 versus 90,000 word comparison. This does not mean the tests are flawed, but that we must be careful when considering their results.

5. Conclusion

There are several conclusions we can draw from our experiments. The statistical tests are accurate for the most part with various combinations of corpus size, word probability and ratio of corpora. From the point of view of both of the statistical tests there are no problems comparing unbalanced sized corpora as long as we avoid low expected values in the contingency table. At the 5% level, the Cochran rule should be extended to ensure expected values are 13 or more. At the 1% level, the Cochran rule should be extended to ensure expected values are 11 or more. At the 0.1% level, the Cochran rule should be extended to ensure expected values are 8 or more. The usual Cochran rule is sufficient at the 0.01% level for the chi-squared test. However (ignoring false positives), we can safely lower the Cochran rule at the 0.01% level for the log-likelihood test to expected values of 1 or more. The trade-off is that the critical value is higher than at the usual 5% level at 15.13.

There is a difference between establishing statistical significance and practical significance. In carrying out tests of significance we should always bear in mind the other issues relevant to corpus comparison as listed in the introduction which are equally important as determining significance: representativeness, homogeneity and comparability. The final issue, that of choice and reliability of statistical tests, has been addressed directly in this study. In any comparison experiments one should keep all these issues in mind when interpreting the results. For example, if we chose to compare a written corpus with a spoken corpus, it is very likely that lexical and grammatical differences between the spoken and written language will be exposed as well as differences in domain or content that we may wish to focus on.

References

Aston G. and Burnard L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*, Edinburgh University Press.

Barnett S. and Cronin T. M. (1986). *Mathematical formulae for engineering and science students, fourth edition*. Longman.

Biber D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, vol. (8/4), Oxford University Press: 243-257.

Butler C. (1985). *Statistics in linguistics*. Blackwell.

Clear J. (1992). Corpus sampling. In Leitner G. (Ed.) *New directions in English language corpora*. Mouton-de-Gruyter: 21-31.

Cochran W. G. (1954). Some methods for strengthening the common $\chi 2$ tests. *Biometrics*, vol. (10): 417-451.

Cressie N. and Read T.R.C. (1984) Multinomial Goodness-of-Fit Tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. (46/3): 440-464.

De Cock S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics*, vol. (3/1), John Benjamins: 59-80.

Dunning T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, vol. (19/1): 61-74.

Everitt B.S. (1992). *The analysis of contingency tables, 2nd edition*. Chapman and Hall.

Francis B., Green M. and Payne C. (1993) *The GLIM4 system*. Oxford University Press.

Ghadessy M., Henry A. and Roseberry R.L. (Eds) (2001). *Small corpus studies and ELT: theory and practice*, John Benjamins.

Granger S. (1998). The computer learner corpus: a versatile new source of data for SLA research. In Granger S. (Ed.), *Learner English on Computer*. Longman: 3-18.

Hofland K. and Johansson S. (1982). *Word frequencies in British and American English*. The Norwegian Computing Centre for the Humanities.

Kilgarriff A. (1996a). Which words are particularly characteristic of a text? A survey of statistical approaches. In Evett L.J. and Rose T.G. (Eds), *Language Engineering for Document Analysis and Recognition (LEDAR), AISB96 Workshop proceedings, Brighton, England*. Faculty of Engineering and Computing, Nottingham Trent University: 33-40.

Kilgarriff A. (1996b) Why chi-square doesn't work, and an improved LOB-Brown comparison. *ALLC-ACH Conference*.

Kilgarriff A. (1997). Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Proceedings 5th ACL workshop on very large corpora*. Beijing and Hong Kong.

Leech G. (1993). 100 million words of English: a description of the background, nature and prospects of the British National Corpus project. *English Today* 33, Vol. (9/1). Cambridge University Press.

Leech G., Rayson P., and Wilson A. (2001). *Word frequencies in written and spoken English: based on the British National Corpus*. Longman.

Nelson G., Wallis S. and Aarts B. (2002). *Exploring Natural Language: Working with the British Component of the International Corpus of English*. John Benjamins.

Oakes M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press.

Pearson K. (1904). On the theory of contingency and its relation to association and normal correlation. *Biometric Series*, vol. (1). Drapers' Co. Memoirs.

Rayson P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Ph.D. thesis, Lancaster University.

Roland D., Jurafsky D., Menn L., Gahl S., Elder E. and Riddoch C. (2000). Verb Subcategorization Frequency Differences between Business-News and Balanced Corpora: the role of verb sense. In *proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*. 1-8 October 2000, Hong Kong: 28-34.

Scott M. (2000). Focusing on the text and its key words. In Burnard L. and McEnery T. (Eds) *Rethinking language pedagogy from a corpus perspective: papers from the third international conference on teaching and language corpora*. Peter Lang: 104-121.

Scott M. (2001). Mapping key words to problem and solution. In Scott M. and Thompson G. (Eds) *Patterns of Text: in honour of Michael Hoey*, Benjamins: 109-127.

Stubbs M. (1996). *Text and corpus analysis: computer-assisted studies of language and culture*. Blackwell.

Virtanen T. (1997). The progressive in NS and NNS student compositions: evidence from the International Corpus of Learner English. In Ljung M. (Ed.) *Corpus-based studies in English: papers from the seventeenth International Conference on English language research on computerized corpora (ICAME 17)*, Stockholm, May 15-19, 1996. Rodopi: 299-309.

Weeber M, Vos R. and Baayen R.H. (2000). Extracting the Lowest Frequency Words: Pitfalls and Possibilities. *Computational Linguistics*, vol. (26/3). MIT Press: 301-317.

Woods A., Fletcher P. and Hughes A. (1986). *Statistics in language studies*. Cambridge University Press.

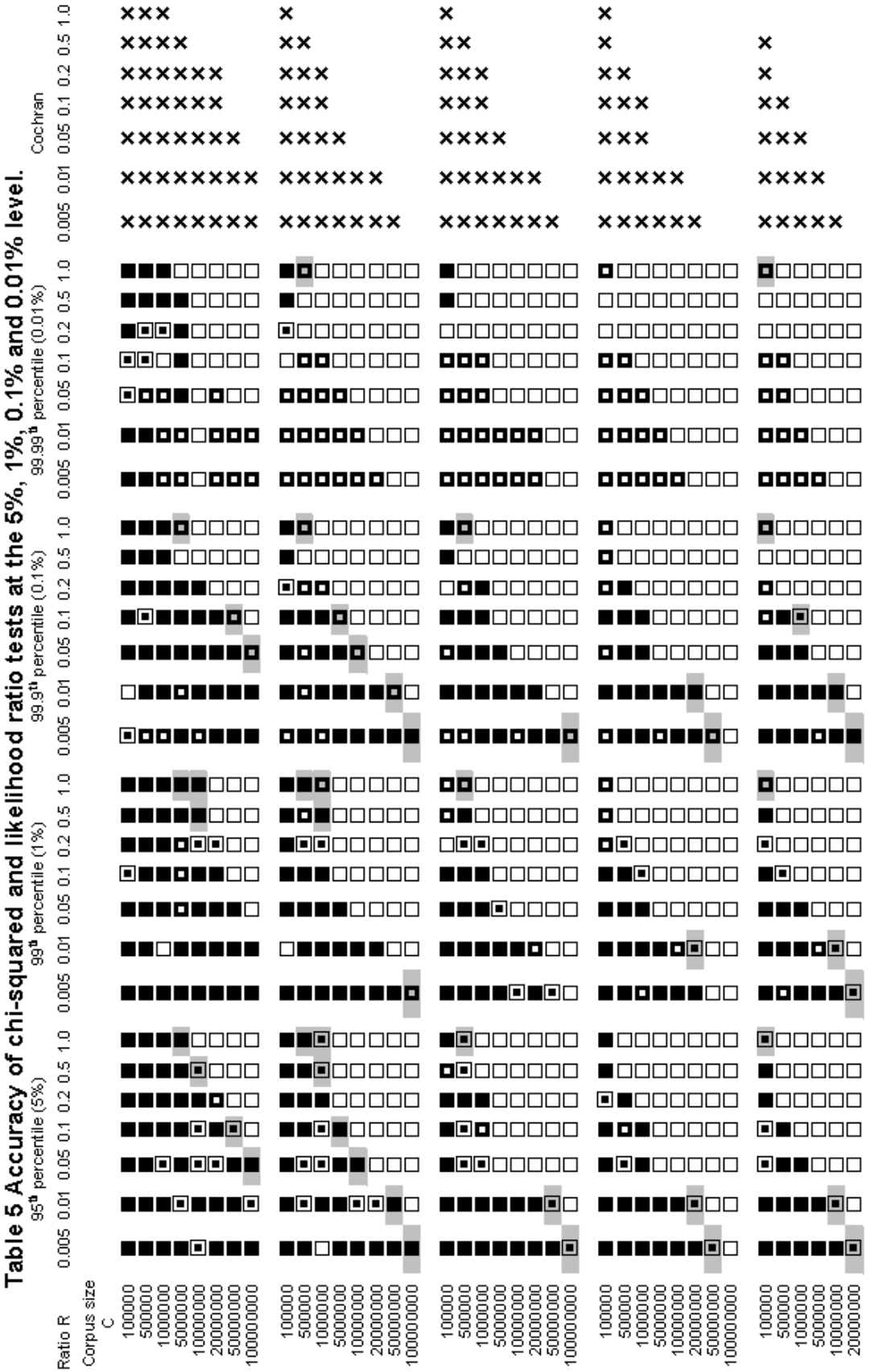
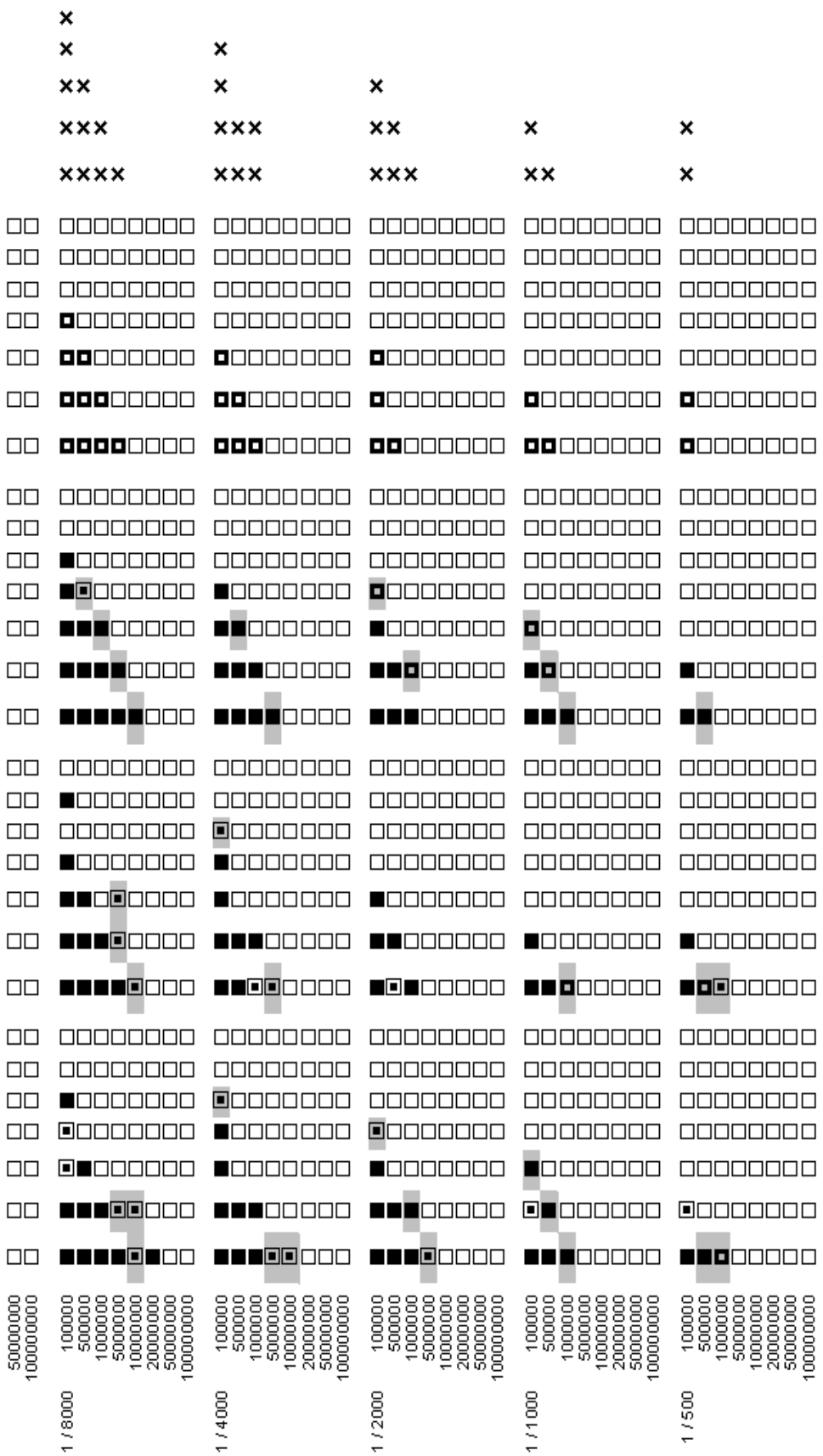


Table 5 Accuracy of chi-squared and likelihood ratio tests at the 5%, 1%, 0.1% and 0.01% level.
95th percentile (5%)
99th percentile (1%)
99.9th percentile (0.1%)
99.99th percentile (0.01%)
Cochran
Proportion p
Ratio R
Corpus size C
0.005 0.01 0.05 0.1 0.2 0.5 1.0
1/1000000
1/100000
1 / 64000
1 / 32000
1 / 16000
100000
500000
1000000
5000000
10000000
20000000
50000000
100000000


1 / 8000
1 / 4000
1 / 2000
1 / 1000
1 / 500
100000
500000
1000000
5000000
10000000
20000000
50000000
100000000