Use of the Chi-Squared Test to Examine Vocabulary Differences in English Language Corpora Representing Seven Different Countries

Michael P. Oakes
University of Sunderland, UK

Malcolm Farrow
University of Newcastle upon Tyne, UK

Abstract

The chi-squared test is used to find the vocabulary most typical of seven different ICAME corpora, each representing the English used in a particular country. In a closely related study, Leech and Fallon (1992, Computer corpora – what do they tell us about culture? *ICAME Journal*, **16**: 29–50) found differences in the vocabulary used in the Brown Corpus of American English and that the Lancaster–Oslo–Bergen Corpus of British English. They were mainly interested in those vocabulary differences which they assumed to be due to cultural differences between the United States and Britain, but we are equally interested in vocabulary differences which reveal linguistic preferences in the various countries in which English is spoken. Whether vocabulary differences are cultural or linguistic in nature, they can be used for the automatic classification according to variety of English of texts of unknown provenance. The extent to which the vocabulary differences between the corpora represent vocabulary differences between the varieties of English as a whole depends on the extent to which the corpora represent the full range of topics typical of their associated cultures, and thus there is a need for corpora designed to represent the topics and vocabulary of cultures or dialects, rather than stratified across a set range of topics and genres. This will require methods to determine the range of topics addressed in each culture, then methods to sample adequately from each topical domain.

Correspondence:
Michael Oakes,
St Peters Campus,
St Peter's Way,
Sunderland,
SR6 0DD,
UK.
E-mail:
michael.oakes@sunderland.ac.uk

1 Introduction

The first study to use an objective measure, the chi-squared test, to observe which words were more typical of a sample of American English as opposed to a sample of British English, was performed by Hofland and Johansson (1982). Leech and Fallon (1992) later used the same measure to examine differences in the vocabulary of the Brown Corpus (one million words of American English) and the

Literary and Linguistic Computing, Vol. 22, No. 1, 2007.
doi:10.1093/llc/fql044 Advance Access Published on 30 September 2006

Lancaster–Oslo–Bergen corpus (LOB, consisting of one million words of British English) in fifteen different categories, such as sport, transport and travel, and business. They regarded some of the differences in vocabulary they found as indicative of differences in spelling (e.g. *color* and *colour*), lexical choice (e.g. *gasoline* and *petrol*), and proper nouns associated with the two countries (e.g. *Chicago* was more common in the US texts and *London* was more common in the British texts). Such differences they referred to as linguistic contrasts. All other differences they referred to as non-linguistic contrasts, which they regarded as indicators of socio-cultural differences between the two countries. The LOB and Brown corpora contain English used in the 1960s. More recently corresponding corpora called Freiburg–LOB (FLOB) and Freiburg–Brown (Frown) have been compiled from English of the 1990s. The work of Leech and Fallon has been repeated using the FLOB and Frown corpora by Oakes (2003). In this study, the comparison of a corpus of American English with a corpus of British English has been extended to include samples of written English from other countries, namely Australia, India, Kenya, New Zealand, and Tanzania. In the US, Britain, Australia, and New Zealand, English is spoken as a first language, while in India and East Africa it is mainly spoken as a second language (Crystal, 1997). English is used as a lingua franca in both Kenya and Tanzania. In both countries, English is used in secondary and tertiary education, and in the High Court. In Tanzania, Kiswahili is used in Parliament and Government institutions, but in such situations in Kenya, English would be used. (Hudson-Ettle and Schmied, 1999). All the samples may be found on the ICAME CD (Hofland *et al.*, 1999), as shown in Table 1. The total number of words in all seven corpora combined is 5,380,477. Five of the corpora are balanced with respect to each other, containing approximately the same proportions of fiction, newspaper text, and so on, as shown in Table 2. This data was obtained from the ICAME disc manuals and the Wellington corpus website. The reason the Australian corpus contains fewer words than the other four, in Table 2, is that only those sections for which copyright permission letters could be obtained were placed on the ICAME disc (Hofland, personal communication).

Table 1 Samples of written English from the ICAME CD used in this study

Corpus	Country	Words
Australian Corpus of English (ACE)	Australia	746,372
Freiburg–Lancaster–Oslo–Bergen (FLOB)	Britain	1,009,765
Kolhapur corpus	India	1,006,315
International Corpus of English (ICE, part of the EA component)	Kenya	299,792
Wellington corpus	New Zealand	1,016,623
International Corpus of English (ICE, part of the EA component)	Tanzania	292,012
Freiburg–Brown (Frown)	United States	1,009,598

As shown in Table 3, the Kenyan and Tanzanian corpora are balanced with respect to each other, but not with respect to the other five corpora used in this study. Although available on the ICAME disc, they were originally produced for the ICE (see www.ucl.ac.uk/english-usage/ice). For this study, the collections of East African printed texts were used.

2 Method: The Chi-Squared Test

We construct a contingency table with rows corresponding to words and columns corresponding to corpora. In the cell corresponding to word W in corpus C, we enter the number of occurrences of W in C. We wish to compare the frequency distributions of words between corpora. In particular, we will test the null hypothesis that the underlying probability distributions for English from the five different countries are equal.

The usual test to compare frequency distributions in a contingency table is the chi-squared test. Usually, a random sample is taken from each of a number of populations and the frequency, within each sample, of each of a number of categories is recorded. The set of frequencies for one sample then has a multinomial distribution (Plackett, 1981). For large samples, the usual test statistic has approximately a chi-squared distribution under the null hypothesis that the

Table 2 Number of sections of ~2,000 words in five comparable corpora

		ACE	FLOB	FROWN	Kolhapur	WC
A	Press: reportage	44	44	44	44	44
B	Press: editorial	27	27	27	27	27
C	Press: reviews	17	17	17	17	17
D	Religion	17	17	17	17	17
E	Skills, trades, and hobbies	38	38	36	36	38
F	Popular lore	44	44	48	48	44
G	Belles lettres, biography, and essays	77	77	75	75	77
H	Miscellaneous, e.g. government documents	30	30	30	30	30
J	Learned and scientific writings	80	80	80	80	80
K	General fiction	29	29	29	29	126
L	Mystery and detective fiction	15	24	24	24	Fiction, not sub-categorised
M	Science fiction	7	6	6	6	
N	Adventure and western fiction	8	29	29	29	
P	Romance and love story	15	29	29	29	
R	Humour	15	9	9	9	
S	Historical fiction	22	0	0	0	
W	Women's fiction	15	0	0	0	
Total		500	500	500	500	500

Table 3 Composition of the Kenyan and Tanzanian printed components of the ICE-EA corpus in words

	Kenyan	Tanzanian
Creative	40,212	40,201
Instructional: administrative/regulatory	18,880	20,129
Learned: humanities	20,096	20,182
Learned: natural science	20,020	20,123
Learned: social science	20,080	10,098
Learned: technology	20,081	20,162
Persuasive: column	20,081	20,112
Persuasive: editorial	20,041	20,088
Popular: general	0	13,797
Popular: humanities	20,096	20,160
Popular: natural science	20,117	6,549
Popular: social science	20,022	20,155
Popular: technology	20,012	20,081
Reportage: features	20,092	20,149
Reportage: splash	20,006	20,027
Total	299,792	292,013

underlying populations are equal. For the purpose of this approximation, a sample is usually regarded as sufficiently large if every cell in the table has an expected frequency of at least 5 under the null hypothesis (see, for example, Altman 1991, Section 10.6.8). In this study, words giving rise to expected frequencies of less than 5 are omitted.

The sampling process in our case falls short of this ideal because the sampling units are not individual words but documents of about 2,000 words each. The five balanced corpora each contain 500 such documents. This causes two problems:

(1) The rates of occurrence of particular words may differ from document to document.
(2) Words in sequence in a document are not independent random samples. Rather, we might regard each document as the realisation of a stochastic process.

We intend to examine both of these difficulties in more detail in a future work. For now we argue that our main conclusions are unlikely to have been significantly affected. The large number of documents in a corpus tends to reduce any effect of either problem. We give further consideration to problem (1) in Section 2.1 below. The main implication of problem (2) is that if, for example, two words tend to occur together, the frequencies of occurrence of these two words are not independent. This affects the true significance level of the test. However, the effect is to make the test conservative. That is, when the null hypothesis is true, the probability of getting a result which is apparently significant at the α% level is actually $<\alpha$%. This has

been confirmed in simulation studies by the authors. The same is true if occurrences of two words are negatively correlated, perhaps because they are used as alternatives.

A third potential difficulty is the fact that each corpus contains literature in several different categories. It may be that the distinguishing characteristics of the different national varieties of English literature differ between categories, suggesting that it might be better to sum chi-squared statistics calculated separately for each category. Again, we will discuss this point in a future work but, since our present aim is to identify overall typical words, we do not disaggregate in this way in this study.

In the first experiment, a comparison of the vocabulary in the five closely balanced corpora of Australia, Britain, India, New Zealand, and United States was made. In the second experiment, a comparison of the vocabulary present in all seven corpora was made, although in this case some of the differences found in the Kenyan and Tanzanian might be due to the fact that they were compiled using a different sampling strategy to the other five corpora. In both experiments, first an overall chi-squared value was found to test the null hypothesis that there was no significant difference in the vocabulary distribution of each corpus. Then the standardised residual for each word type in each corpus was examined to find the words that occurred significantly more often in any of the corpora.

To determine the overall chi-squared value for the five balanced corpora, a computer program was written to construct a contingency table with five columns (one for each corpus) and 101,984 rows (one for each distinct word type found in any of the corpora). In these experiments, a word was defined as a string of consecutive alphabetic and numeric characters, and all upper case characters were converted to lower case. The contingency table entries for three word types, *a*, *commonwealth*, and *zzzzooop*, are shown in Table 4.

The values in the contingency table are the number of occurrences of each word token in each corpus, which are called the observed frequencies (O). The column totals are the number of words in the entire corpus, and the row totals are the number of times the word appears across all five corpora. There is no requirement that either all the column totals or all the row totals must be identical. The chi-squared test makes no assumption of equal sample sizes and is valid even if the sample sizes are very different, sample sizes being taken into account in the calculation of the chi-squared statistic (see, for example, Bain and Engelhardt, 1987, Section 13.5); so we can work with different sized corpora and words of different overall frequencies. The grand total is the total number of word tokens across the five corpora, and is both the sum of the column totals and the sum of the row totals. The next stage is to calculate a corresponding expected frequency (E) for each observed frequency. The expected frequencies take into account that the corpora are of different sizes and that the words are of different frequencies, and are calculated using the formula *expected frequency* = *row total* × *column total/grand total*. For example, the expected frequency of *commonwealth* in the ACE corpus is $249 \times 767{,}579/4{,}893{,}056 = 39.06$. The expected values for each of the observed values shown in Table 4 are shown in Table 5.

For a word type to be considered further in this analysis, its expected frequency in each corpus must

Table 4 Observed frequencies for three word types in the five balanced corpora

Word	ACE	FLOB	Frown	Kolhapur	WC	Row total
a	17,547	23,116	23,269	21,493	23,464	108,889
:	:	:	:	:	:	:
commonwealth	164	12	11	20	42	249
:	:	:	:	:	:	:
zzzzooop	1	0	0	0	0	1
Column total	746,372	1,030,811	1,034,722	1,025,237	1,034,707	Grand total = 4,893,056

be at least 5. Thus the word *zzzzooop* makes no contribution to the final chi-squared value.

The third step is to calculate for each cell the quantity $(O-E)^2/E$. For the word *commonwealth* in the ACE corpus, this gives $(164-39.06)^2/39.06 = 399.63$. The greater the difference between the observed and expected frequency, the more evidence we have that some words are relatively more frequent in some corpora than others. The $(O-E)^2/E$ values corresponding to each cell in Table 5 are shown in Table 6.

The $(O-E)^2/E$ values for every word type in each corpus are added together to produce an overall chi-squared value of 414,916.81. The number of degrees of freedom for a contingency table is the (number of rows − 1) multiplied by the (number of columns − 1); in this case $101{,}983 \times 4 = 407{,}932$. For this overall chi-squared value with this number of degrees of freedom, we can reject the null hypothesis that there is no difference in the vocabulary distribution of the five balanced corpora at the 0.1% level of significance.

Having determined that there is strong evidence for the existence of genuine differences in the frequencies of the words in the five varieties of English literature, the next stage is to look for individual cases where an individual word occurs more often in a corpus than would be expected under the null hypothesis of equal distributions. The approach we take here is to examine the standardised residual $(O-E)^2/\sqrt{E}$ in each cell (Haberman, 1973). The square of the standardised residual, $(O-E)^2/E$, is the contribution of that cell to the overall statistic, as shown in Table 6. It is easily shown that, under the null hypothesis, the mean of the distribution of the standard residual is 0 and its standard deviation is 1. Further, in moderately large samples, where E is at least 5, the distribution is approximately normal. Although the standardised residuals are not strictly independent of each other, in a large table, the dependence is very slight, and in any case, does not invalidate our procedure described below.

In this study, we consider only those words which are typical of corpora, where the observed frequency of that word in a corpus is greater than the expected frequency. We can produce a ranking of all the words in each corpus, according to the standardised residuals. We check whether each standardised residual is statistically significant by assuming that, under the null hypothesis, they have approximately a standard normal distribution, and applying the Bonferroni correction for multiple comparison procedures, as will be described in Section 2.1.

When all the distinct words (types) in the Australian corpus are placed in order of greatest chi-squared value first, we obtain the list of the top 20 words most typical of the Australian corpus given in Table 7. The word itself is in the first column, its frequency in each corpus is given in columns 2–6, and its standardised residual in the ACE corpus is given in the final column.

Toju, a Japanese name, appears because it appears 47 times in one story excerpt, and never

Table 6 $(O-E)^2/E$ values for three words in the five balanced corpora

Word	ACE	FLOB	Frown	Kolhapur	WC
a	12.68	1.36	2.55	76.65	8.33
:	:	:	:	:	:
commonwealth	399.63	31.20	32.95	19.84	2.16
:	:	:	:	:	:
zzzzooop	–	–	–	–	–

Table 5 Expected frequencies of three words in the five balanced corpora

Word	ACE	FLOB	Frown	Kolhapur	WC
a	17,081.54	22,939.44	23,026.48	22,815.40	23,026.14
:	:	:	:	:	:
commonwealth	39.06	52.46	52.66	52.17	52.65
:	:	:	:	:	:
zzzzooop	0.16	0.21	0.21	0.21	0.21

Table 7 The 20 words most typical of Australian English

Word	ACE	FLOB	Frown	Kolhapur	WC	$(O-E)^2/E$ (ACE)	Standardised residual (ACE)
Australia	973	40	10	37	252	2859.7	53.48
Australian	832	22	6	10	166	2757.9	52.52
Sydney	339	12	3	4	75	1081.8	32.89
Aboriginal	210	1	1	3	5	892.3	29.87
Queensland	202	1	0	0	8	861.9	29.36
Melbourne	203	8	3	4	19	739.6	27.20
NSW	138	0	0	0	1	619.2	24.88
Aborigines	101	0	1	0	1	445.5	21.11
Brisbane	105	3	0	0	10	404.1	20.10
Commonwealth	164	12	11	20	42	399.6	19.99
Hawke	114	0	0	0	24	394.0	19.85
Australians	121	4	1	1	47	321.7	17.94
Canberra	73	0	0	0	10	276.3	16.62
Perth	74	0	1	0	11	271.4	16.47
Adelaide	68	1	0	1	8	254.1	15.94
Macquarie	58	0	0	0	2	250.8	15.84
Tasmania	59	0	0	0	3	249.6	15.80
Kakadu	53	0	0	0	0	240.2	15.50
Ewes	64	14	0	0	0	219.0	14.80
Toju	47	0	0	0	0	213.0	14.59

in any of the other corpora. A method of filtering out such terms is to reject all words with low dispersion (where most occurrences of the word are clumped together) as estimated by a statistical measure of dispersion such as Juilland's *D* measure (Lyne, 1986). The use of this measure, which also filters out the word *ewes*, will be described further in Section 2.2.

2.1 Controlling the false discovery rate

In examining the table of standardised residuals for significant values, we are effectively performing as many comparisons as there are cells in the contingency table. If we use a significance level of 0.001 to determine if each word is typical of each corpus, one in a thousand comparisons will appear positive purely by chance. With a contingency table of 509,920 cells, where, in fact, there are no underlying deviations from the null hypothesis, we might expect about 500 occurrences where a word is spuriously deemed typical of a corpus. This problem of multiple testing can be addressed by adjusting the significance levels of the individual tests. A commonly used method is the Bonferroni correction (Miller, 1981). This controls the 'familywise error rate' (FWER), that is the probability of rejecting at least one out of m multiple hypotheses when all are, in fact, true. We can obtain an FWER $\leq 100\ \alpha\%$ by testing each of the individual hypotheses at the $100\ \alpha/m$ level.

The Bonferroni correction is 'conservative' in that it errs on the side of non-significance. Further, the FWER is not the most appropriate measure for our purpose. More appropriate is the 'false discovery rate' (FDR) (Benjamini and Hochberg, 1995). This is the expected value of the proportion of the rejected null hypotheses which are in fact, true. In our case it is the expected value of the proportion of the words identified as typical, which have been erroneously identified. Benjamini and Hochberg proposed a method for controlling the FDR. It is easily seen that any hypothesis rejected using the Bonferroni correction for an FWER of $100\ \alpha\%$ would also be rejected under the Benjamini and Hochberg procedure for an FDR of $100\ \alpha\%$. (The converse is not true.) A modification of the FDR procedure, called the 'positive false discovery rate', which gives greater power, is described by Storey (2002).

To identify the words listed in Section 3 below, we simply used the Bonferroni correction to give an FWER of not more than 0.001 (i.e. 0.1%). Therefore the FDR will also not be >0.1%. In comparing the vocabulary of the five balanced corpora, we effectively perform 509,920 tests because there are 101,984 unique word types across the corpora, each tested once for each corpus. To find the appropriate critical value of the standard normal distribution, we divided 0.001 by 509,920 to give an adjusted significance level of 1.961×10^{-9}. For a one-sided test, since we are only considering positive residuals, the critical value of the standard normal distribution is 5.89. The corresponding value from the chi-squared distribution of one degree of freedom for the individual contributions to the chi-squared statistic is 34.66, which is 5.89^2. We then identify words with positive standard residuals greater than 5.89 or chi-squared contributions above 34.66. This procedure ensures that, on average, not >0.1% of the words selected using that will have been incorrectly identified, subject, of course, to the proviso that the normal distribution is a good approximation in this case.

The WordSmith corpus analysis package (Scott, 1999) has a 'keywords' feature, which uses the chi-squared test to identify words which occur with an unusually high frequency in an analysis corpus with respect to a reference corpus. It would have been possible to compare each national corpus against a reference corpus consisting of all the remaining national corpora. The method used in this study allows all seven national corpora to be simultaneously compared, and overcomes the problem that with the WordSmith method, the tests for individual words are not strictly independent of one another. For example, if we had two 100-word corpora, we might compare the use of Word A across the corpora, and find that it occurred 50 times in Corpus 1 and 20 times in Corpus 2. Then, in a second statistical test, we compare the frequency of Word B across the two corpora. This second comparison is not altogether independent of the first, because there are only 50 slots left in Corpus 1 where Word B could possibly occur (given what we know from the first statistical test). Thus there must be some interdependence among the statistical tests for each word. A much greater problem is the FDR. When making a large number of comparisons (one for each word in each corpus), we will inevitably get some false positives purely by chance. We implement Bonferroni's correction to control the FDR.

2.2 Dispersion

Dispersion measures show how evenly or otherwise a word is distributed throughout a corpus (Lyne, 1985, 1986). In a study of this nature, it is important to consider only those words which are relatively evenly spread throughout the corpus. Considering the chi-squared measure alone, it would seem that the 15th most typical word of British English is *thalidomide*. However, all 55 occurrences of this word in the FLOB corpus are found in a single article entitled 'Thalidomide treatment for chronic graft-versus-host disease'. This suggests that the appearance of *thalidomide* so high in the chi-squared list is more a function of sampling decision to include that particular article than the fact that the word is genuinely typical of British English.

The simplest measure of dispersion is the range, which is the number of subsections of the corpus the word appears in. For example, if the FLOB corpus is divided into five consecutive subsections of equal length, *thalidomide* appears 55 times in the fourth subsection, but no times at all in any of the other subsections. The range is the number of subsections that contains the word at least once, which is just one for *thalidomide*. The word *England*, which, considering chi-squared value alone, would be the 14th most typical word of British English, occurs 106, 61, 78, 27, and 27 times, respectively, in each subsection, and thus has a range of 5. Since this word appears in all five subsections of the corpus, we can be more confident that it really does occur more often in British English than in corpora of other Englishes.

A more sophisticated measure of dispersion is Juilland's *D* (Juilland *et al.*, 1970), originally developed for Spanish texts, which takes into account not only the presence or absence of a word in each subsection of the corpus, but the exact number of times it appears. In Lyne's (1985)

experiment, using a corpus of French Business Correspondence, he divided the texts into five equal subsections, which is the approach followed in this study. The subsections are based on splitting the texts in their category-based order—for example, Section 1 contains texts A01, A02, A03, etc. Taking the word *Commonwealth* in the Australian corpus, we find that it occurs 164 times altogether: 31 times in the first subsection, 8 times in the second, 32 times in the third, 88 times in the fourth, and 5 times in the final one. Thus, although the word appears in all subsections of the corpus, it is not altogether evenly distributed, as the majority of occurrences are found in the fourth subsection. The mean number of times the word occurs in each subsection is $164/5 = 32.8$. To find the standard deviation of the number of times the word is found in each subsection, we use the formula

$$s = \sqrt{\frac{\sum (x_i - \text{mean})^2}{n}}$$

where n is the number of corpus subsections $= 5$.

$$\begin{aligned}\sum (x_i - \text{mean})^2 &= (31 - 32.8)^2 + (8 - 32.8)^2 \\ &\quad + (32 - 32.8)^2 + (88 - 32.8)^2 \\ &\quad + (5 - 32.8)^2 \\ &= 3.24 + 615.04 + 0.64 \\ &\quad + 3047.04 + 772.84 \\ &= 4438.8 \\ s &= \sqrt{\frac{4438}{5}} = 29.79\end{aligned}$$

To account for the fact that the standard deviation tends to be higher for more frequent words, the standard deviation is divided by the mean frequency to give a coefficient of variation called V:

$$V = \frac{s}{\text{mean}} = \frac{29.79}{32.8} = 0.908$$

The coefficient of dispersion, D, is derived from V, and is designed to fall in the range 0 (where all occurrences of a word are clumped in the same subsection) to 1 (where the word is distributed perfectly, evenly throughout the corpus).

$$D = 1 - \frac{V}{\sqrt{n-1}} = 1 - \frac{0.908}{\sqrt{4}} = 0.546$$

Juilland also describes a usage coefficient U, which combines dispersion and frequency. For *Commonwealth*, this would be $0.546 \times 164 = 89.5$.

The chi-squared test can also be used as a measure of evenness of distribution (Lyne, 1985, p. 117). The observed values (one for each subsection of the corpus) are the subfrequencies of the word in each subsection. The expected value (which we would find if the word was distributed perfectly evenly throughout the corpus) for each subsection is the total frequency of the word divided by the number of subsections. Since, when using the chi-squared test, there is a requirement that all the expected values are at least five, the test can only be used for words with a total frequency in the corpus of at least five times the number of subsections. Chi-squared value will be 0 if the word is evenly dispersed throughout the corpus, and a positive value otherwise. If five subsections are used, we have $5 - 1 = 4$ degrees of freedom, and a chi-squared value greater than 13.28 will lead to rejection of the hypothesis that the word is evenly distributed at the 1% level of significance.

Whether we use range, D, or U, our cut-off point for discriminating between well and poorly dispersed words must be arbitrary. To estimate suitable cut-off points or thresholds, consider the following table of the 25 proper nouns with the greatest chi-squared values in the FLOB corpus, shown in Table 8.

We have placed an asterisk in the second column for each proper noun we subjectively would not expect to be typical of a British English corpus. The columns headed S1–S5 are the number of times each word occurred in each subsection of the corpus. With the exception of *Gorbachev* and *Kinnock*, where intermediate results were obtained, the proper nouns could be automatically classified using dispersion measures. All the proper nouns

Table 8 The 25 proper nouns with greatest chi-squared values in the FLOB corpus of British English

Word	Estim	χ^2	S1	S2	S3	S4	S5	Range	JD	U
UK		526.6	37	68	77	38	3	5	0.71	157.5
London		506.6	156	104	109	33	65	5	0.78	362.5
Britain		350.6	114	83	55	50	17	5	0.74	237.5
Thatcher		222.6	69	20	9	4	6	5	0.44	47.2
Kinnock		189.3	60	2	0	2	0	3	0.08	5.0
Scotland		184.4	18	24	25	12	18	5	0.88	85.2
James		163.0	33	20	89	13	80	5	0.67	156.5
Glasgow		146.3	10	8	37	1	3	5	0.45	26.5
Edward		142.2	23	12	21	5	52	5	0.64	72.9
England		140.7	106	61	78	27	27	5	0.75	223.0
Gorbachev	*	132.6	31	6	35	0	0	3	0.47	33.5
Lara	*	122.7	2	0	0	0	41	2	0.06	2.5
Bristol		117.9	8	29	6	1	4	5	0.48	23.1
Oxford		118.4	27	20	25	10	12	5	0.82	77.0
Minton	*	112.4	0	0	38	0	0	1	0	0
Chrissie	*	108.0	0	0	0	0	38	1	0	0
Jacko	*	103.8	0	0	0	0	38	1	0	0
Tissee	*	103.5	0	0	0	0	35	1	0	0
Manchester		103.1	29	18	4	2	6	5	0.57	33.4
Tansy	*	100.6	0	0	0	0	34	1	0	0
Caruso	*	100.6	1	0	33	0	0	2	0.04	1.2
Kent		97.7	16	13	1	1	10	5	0.62	25.6
Abby	*	97.6	0	0	0	0	33	1	0	0
Clive		96.2	9	0	3	9	26	4	0.52	24.5
Clementina	*	94.6	0	0	5	0	27	2	0.18	5.8

we felt were typical of British texts had a range of 4 or 5, while the others had a range of 1 or 2. Juilland's *D* was 0.44 or more for all the proper nouns we felt were typical of British texts, while for the other proper nouns it was 0.18 or less. *U* was greater than 23 for the proper nouns we considered typical of a British text, and less than six for the others. As a result of this small study, we felt that a combined cut-off point of 3 for range and 0.3 for Juilland's *D* should be used to determine the well-distributed words from the rest for the remainder of this work. The WordSmith corpus analysis tool has a 'plot' feature, which shows graphically how words are distributed throughout a corpus. In this study, we quantify the degree of dispersion with numeric methods, and also estimate a cut-off point below which we can say that a word is insufficiently dispersed throughout a corpus to be considered typical of the corpus as a whole.

3 The Vocabulary Most Typical of Each of the Corpora

The 50 words most typical of each of the five balanced corpora studied as ordered by their contributions to the overall chi-squared value (where Juilland's dispersion measure *D* is at least 0.3 and the word appears in at least three out of five subsections) are shown in Table 9. Altogether, the number of words with a chi-squared contribution above 34.66 with sufficient range and dispersion were 166 for the ACE corpus, 127 for FLOB, 249 for Frown, 348 for Kolhapur, and 160 for the Wellington corpus.

In the Australian sample, 18 of the 19 words with highest chi-squared values are the names of Australian people and places. The exception is *Commonwealth* ($\chi^2 = 399.6$)—Australia is the only Commonwealth country in this study to have

Table 9 The 50 words most typical of each of the five balanced corpora

Australia	Australia (2859.7), Australian (2757.9), Sydney (1081.8), Aboriginal (892.3), Queensland (861.9), Melbourne (739.6), NSW (619.2), Aborigines (445.5), Brisbane (404.1), Commonwealth (399.6), Hawke (394.0), Australians (321.7), Canberra (276.3), Perth (271.4), Adelaide (254.1), Macquarie (250.8), Tasmania (249.6), Kakadu (240.2), Victoria (215.0), mining (189.1), ALP (181.3), Senator (162.7), Federal (146.6), unions (146.1), Whitlam (146.0), Premier (143.4), wool (141.1), WA (140.0), Hobart (137.8), govern (136.5), Wales (134.2), unemployed (133.6), Pope (127.5), Labor (126.5), ABC (123.0), BHP (119.0), 1985 (109.0), Fraser (104.7), kilometres (103.2), commission (102.2), government (96.4), industry (93.5), ski (92.0), superannuation (89.1), library (89.0), bread (86.0), fringe (84.5), Mr (83.2), card (81.9), Victorian (79.6).
Britain	UK (526.6), London (506.6), Britain (350.6), British (232.4), Thatcher (222.6), NHS (192.2), Scotland (184.4), BBC (184.3), Labour (181.6), Scottish (177.7), Tory (165.7), 1990 (163.0), James (163.0), local (155.6), pounds1 (153.8), royal (150.0), Glasgow (146.3), which (142.8), Edward (142.2), England (140.7), charter (138.4), Gorbachev (132.6), century (142.5), EC (123.8), pounds2 (121.3), Oxford (118.4), Bristol (117.9), authority (112.2), Commons (108.0), Lord (105.9), Manchester (103.1), Prince (102.9), Lords (102.1), European (101.5), Kent (97.7), Clive (96.2), 1989 (93.2), she (92.8), Duke (89.4), cent (85.2), Essex (85.2), the (84.5), he (84.4), Churchill (83.6), Marie (83.3), Olivier (82.7), Frances (82.1), Southampton (80.5), poll (77.2), eighteenth (76.0).
India	India (3834.7), Indian (1874.4), Gandhi (1146.6), Rs (1141.1), Delhi (933.9), the (904.1), Bombay (838.5), Singh (835.7), of (795.4), Nehru (534.6), caste (523.4), Congress (520.9), Bengal (510.3), castes (427.9), Indira (393.6), Ram (388.0), crores (340.0), Pradesh (338.4), Punjab (318.2), sabha (307.2), Hindi (307.2), Maharashtra (301.2), Kerala (301.2), Hindu (299.5), Calcutta (295.1), village (292.6), etc. (291.9), scheduled (288.0), Krishna (279.0), Shri (274.4), Lok (259.7), Bihar (265.5), Sanskrit (254.9), Desai (253.5), Rupees (252.0), temple (248.8), Patel (246.8), Tamil (246.2), Bengali (244.6), seeds (232.2), Madras (225.4), upto (214.7), Raja (214.7), Kumar (210.5), lakhs (202.8), Chandra (201.3), constitution (199.9), sari (196.9), Raj (195.3), villages (193.4).
NZ	Zealand (5440.3), Maori (2749.3), Auckland (1739.0), New (1462.9), Wellington (1427.3), Te (831.2), Christchurch (713.0), pakeha (620.4), Canterbury (417.5), Zealanders (394.5), Lange (355.7), Otago (329.3), Pacific (264.6), Dunedin (249.9), Waikato (247.0), rugby (235.9), Maoris (233.7), 1984 (228.8), Bay (225.6), island (222.3), Waitangi (197.0), NZ (183.9), Mum (172.6), overseas (171.6), Hutt (167.6), Aotearoa (167.6), beech (157.5), kiwi (156.6), harbour (153.8), Ngati (152.9), Tasman (151.4), GST (147.1), Rotorua (138.2), forest (138.2), islands (136.5), Dad (127.4), players (126.1), were (124.1), NZPA (123.5), Zealander (122.0), marae (120.6), Pa (119.9), Bryce (117.9), Hamilton (116.6), Landscape (116.6), Nelson (115.0), TVNZ (114.7), Wanganui (110.3), Waitaki (108.8), tussock (107.3).
USA	percent (807.1), S (759.4), toward (739.9), U (733.3), programs (526.7), defense (524.8), American (514.6), program (491.7), center (475.2), President (445.7), Washington (385.6), behaviour (332.8), color (311.3), Americans (289.9), California (279.4), black (278.0), labor (268.1), fiber (255.8), United (218.0), gray (213.4), 1991 (213.3), York (211.8), that (201.0), gender (199.4), presidential (199.2), Republican (196.6), federal (181.9), 1992 (173.9), States (171.2), theater (167.7), favourite (164.3), disclosure (164.0), Florida (161.8), white (159.1), Miami (158.6), San (156.9), 1989 (151.8), Los (150.8), favor (150.5), colors (150.5), organization (141.6), county (141.3), Texas (139.2), Francisco (135.8), America (132.2), Illinois (131.3), Angeles (130.2), beans (128.3), Joe (127.9), said (127.1).

this term in its list of the 50 most significant words, probably due to the use of the phrase *Commonwealth of Australia*. Other significant terms in the Australian corpus are *mining* (189.1) and *wool* (141.1). Two terms relate to the political system, *Premier* (143.4) and *Senator* (162.7), as well as the names of the individual politicians *Hawke* (394.0) and *Whitlam* (146.0) and the political parties *ALP* (181.3), *Labor* (126.5), and *BHP* (119.0). A number of terms refer to employment rights: *unions* (146.1), *unemployed* (133.6) and *superannuation* (89.1). The Australian list also includes a local broadcaster, *ABC* (123.0), and lower down the list, the (Sydney) *Herald* (44.9). An artefact of the ACE corpus layout was that a number of words were split across two lines, resulting in certain prefixes and suffixes (con-, com-, -tion, -ing, and -ment) having high $(O-E)^2/E$ values, since they occurred only in the ACE corpus. These were not included in Table 9.

The most significant terms typical of British English were mainly the names of British people and places. *NHS* (National Health Service) (192.2) and *BBC* (British Broadcasting Company) (184.3) are British institutions. Other terms, *Tory* (165.7), *royal* (150.0), and *Labour* (181.6) refer to government, where the Tory party is a nickname for the Conservative Party. Individual politicians found in

the list are *Thatcher* (222.6) and *Churchill* (83.6), and although not a British politician, *Gorbachev* (132.6) is mentioned widely in the British texts. *EC* (123.8) stands for *European Community*. Two terms related to historical epochs, *century* (142.5) and *eighteenth* (76.0) appear in the British corpus. The most striking feature of the British list, also found in the study by Leech and Fallon, was the presence of so many aristocratic titles. *Duke* (89.4), *Earl* (55.0), *Lord* (105.9), *Lords* (102.1), *Prince* (102.9), *Princess* (40.6), *and Royal* (150.0) all had $(O-E)^2/E$ values of over 34.66. No other corpus produced any significant terms for aristocratic titles. An artefact of the corpus mark-up conventions used was the presence of *pounds1* (153.8) and *pounds2* (121.3) for *£1* and *£2* in the list of significantly British terms.

The terms typical of Indian English were again mostly names of people and places, but the list also includes *Rs* (1141.1), abbreviation for the currency (Rupees), *mn* (563.6) for millions, *crores* (340.0) and *lakhs* (202.8), indigenous Indian words for ten million and ten thousand, respectively. Three entries in the list refer to the caste system: *caste* (523.4), *castes* (427.9), and *dalit* (98.4). The list for Indian English also contains four high-frequency function words: *the* (904.1), *and* (58.2), *of* (795.4), and *in* (179.5). Kilgarriff (1996) writes that the appearance of such words in such lists is an artefact of the chi-squared test rather than a distinguishing feature of a particular word sample. The Indian list of significant words is the only one to contain terms for clothes, being the indigenous term *sari* (196.9), and also *cloth* (54.6) and *shawl* (49.4). An important feature of the list of terms significantly typical of the Indian corpus was the sheer number of religious terms, and the variety of religions represented by them. These terms were: *Buddha* (86.0), *Buddhism* (45.4), *divine* (150.6), *Gita* (119.3), *God* (37.8), *Gods* (78.6), *Goddess* (44.4), *Hindu* (299.5), *Hindus* (148.1), *Karma* (61.4), *Muslim* (151.8), *Muslims* (42.2), *mystic* (53.1), *Mystics* (100.7), *pandit* (104.4), *puja* (122.3), *Saints* (35.6), *Sikh* (80.0), *Swami* (131.2), *temple* (248.8), *temples* (104.2), *Vedas* (101.4), *Vedic* (102.9), *yoga* (97.7). No other corpus had more than three significant terms pertaining to religion. The Indian corpus was the only one where *upto* (214.7) was spelt as a single word.

The terms most typical of New Zealand English are nearly all place names. *Te* (831.2) occurs in a number of place names, such as *Te Kuiti* and *Te Anga*. *Lange* (355.7) is a former Prime Minister, and *pakeha* (620.4) is a Maori word for someone of European descent. *Rugby* (235.9) is the sixteenth most typical word of the WC, where other significant sporting terms are *batsman* (41.2), *golf* (34.2), *players* (126.1), and *team* (55.2). The Australian corpus had the significant sporting terms *ski* (92.0), and *skiing* (49.6). The Indian list has *tournament* (47.3) and the US list has *baseball* (125.2), but the British list has no significant sporting terms. Another characteristic of the list of terms typical of New Zealand English was the large number of terms describing the natural world: *bay* (225.6), *beach* (36.0), *cliff* (46.4), *earthquake* (61.9), *forest* (138.2), *harbour* (158.3), *island* (222.3), *islands* (136.5), *lake* (96.2), *lakes* (74.0), and *landscape* (116.6) were all significant. The next greatest number of terms pertaining to the natural world, found in the Australian corpus, was four. Three of the corpora had significant terms for the flora and fauna found in those countries. For New Zealand there were *beech* (157.5), *fern* (57.2), and *snail* (63.3). For Australia there was *crocodile* (42.4), and for India there were *bamboo* (54.9), *deer* (52.5), *elephant* (112.9), and *tiger* (44.4).

There are relatively few people and places in the list of the 50 terms most typical of US English, the largest category of words arising here due to American spellings: *toward* (139.9), *percent* (807.1), *programs* (526.7), *defense* (524.8), *program* (411.7), *color* (311.3), *behavior* (332.8), *labor* (268.1), *fiber* (255.8), *gray* (213.4), *theater* (167.7), *favorite* (164.3), *favor* (150.5), *colors* (150.5), and *organization* (141.6). The *U* (733.3) and *S* (759.4) come from *US*. The US list contains terms from the politics of inclusiveness: *black* (278.0), *gender* (199.4), *white* (159.1), *diversity* (46.1), and *gay* (40.7). The US list contained several typical terms related to transport, but most of these were due to lexical differences, since other terms would be used for the same concepts in the other corpora.

Table 10 Words most typical of the two East African corpora when compared with the five balanced corpora

Kenya	Kenya (9934.3), Nairobi (2731.1), Kenyan (2567.8), Kenyans (2496.1), EA (2027.1), African (848.7), development (787.8), Africa (770.1), women (722.2), pesticides (692.3), malaria (606.3), maize (605.5), livestock (485.1), children (429.8), environmental (413.7), population (362.9), agricultural (360.4), environment (350.1), crops (323.8), farmers (322.6), crop (318.2), food (313.7), Kiswahili (294.5), education (292.5), management (284.0), colonial (246.3), soil (240.0), conservation (237.7), ford (225.8), farmer (224.9), drugs (217.5), programmes (216.0), countries (204.4), disease (197.2), programme (180.5), students (179.7), beans (174.2), curriculum (171.1), fees (164.7), indigenous (145.1), nations (144.1), chemicals (139.6), province (138.8), diseases (134.3), laws (133.8), agriculture (133.8), abortion (128.1), milk (120.4), lawyer (119.3), planting (117.0).
Tanzania	Tanzania (9767.5), EA (7633.4), Ndugu (3794.0), Dar (3196.3), Salaam (3065.1), Es (2591.6), CCM (2114.2), Zanzibar (1679.4), Tanzanian (1537.8), HIV (1072.3), African (1013.0), Kiswahili (797.3), development (758.3), Africa (647.8), AIDS (562.2), democracy (545.4), region (525.6), university (519.1), countries (494.6), district (480.1), regional (362.5), economic (362.1), shall (358.5), transformation (327.6), project (321.7), republic (317.9), education (300.6), health (300.3), 2000 (296.7), programme (294.1), ministry (289.0), malaria (278.6), institutions (278.0), registration (270.5), villages (265.9), primary (242.0), objectives (233.6), teaching (232.7), academic (219.1), independence (208.7), music (207.3), constitution (206.9), 1993 (205.0), activities (202.8), diseases (197.2), therefore (191.4), intercourse (190.3), village (178.7), faculty (168.0), consultants (168.0).

These terms were *traveled* (108.8), *pickup* (93.4), *railroad* (86.8), *highway* (90.2), *transportation* (79.4), and *truck* (45.3).

The words most typical of the East African corpora as determined by the comparison of seven corpora are listed in Table 10. In the list of words for Kenyan English, there are a number of references to agriculture: *pesticides* (692.3), *maize* (605.5), *livestock* (485.1), *agricultural* (360.4), *farmers* (322.6), *crops* (318.2), *food* (313.7), *farmer* (224.9), *beans* (174.2), *agriculture* (133.8), *milk* (120.4), and *planting* (117.0) in the list of 50 words most typical of the Kenyan corpus. Although there are no agricultural terms in the top 50 for the Tanzanian corpus, six such terms were found to be significant lower down in the list. The difference in the number of significant agricultural terms in the two East African corpora was just significant ($\chi^2 = 3.87$, $P < 0.05$). In the comparison of the balanced corpora, the Indian and Australian corpora also had a number of significant terms for agriculture (seven and five, respectively). *EA* (2027.1 in the Kenyan corpus, 7633.4 in the Tanzanian) stands for East Africa.

In the Tanzanian sample, the most significant words refer to people and places, except for the honorific *Ndugu* (3794.0), the political party *CCM* (2114.2), and the local lingua franca *Kiswahili* (797.3 in the Tanzanian corpus, 294.5 in the Kenyan). The East African corpora have most significant terms related to medicine. The Kenyan corpus has *malaria* (606.3), *drugs* (217.5), *disease* (197.2), and *diseases* (134.3) in the top 50, while the Tanzanian corpus has *HIV* (1072.3), *AIDS* (562.2), *health* (300.3), *diseases* (197.2), and *malaria* (278.6). Only one significant educational term was found in the five balanced corpora. There are very highly significant terms for education present in both the East African corpora: *education* (292.5), *students* (197.7), and *curriculum* (171.1) in the Kenyan top 50, and *university* (519.1), *education* (300.6), *teaching* (232.7), and *academic* (219.1) in the Tanzanian top 50. Both corpora contain further educational terms with lower, but still significant, chi-squared contributions. Just two significant educational terms were found in the five balanced corpora. It is possible for the same word to be typical of more than one corpus. An example of this is *Africa*, which is typical of both the Kenyan (770.1) and Tanzanian (647.8) corpora.

4 Conclusion

The goal of Leech and Fallon's study was 'using the corpora as evidence of cultural differences' between Britain and the United States. Although we observed clear differences in the vocabulary of each of our seven corpora, we cannot necessarily conclude that these are due to cultural differences between the

seven countries. Our approach has been to compare vocabulary across corpora rather than cultures. The reason is that we do not know that the sample of texts included in any of the seven ICAME corpora represents the full range of topics typical of the associated culture. What we need are corpora designed to represent the general topics and vocabulary of cultures or dialects. This would require methods to determine the range of topics addressed in each culture, and then methods to sample adequately from each topical domain. Here we have a trade-off situation. Balanced corpora, such as FLOB and Frown, are necessary for many statistical comparisons. However, two corpora designed to reflect the ranges of topics and styles typical of two different cultures, might well not contain the same range of topics and styles as each other, and hence not be balanced. A difficulty with categorising words as described in Section 4 is that some of the categories overlap. For example, science is a very broad category, which could include terms for medicine or flora and fauna. In this study, 'science' is taken to mean 'science excluding medicine or flora and fauna'. With these caveats in mind, the results of this study may be summarised as follows:

The Indian corpus had both the greatest number of significant religious terms and the greatest variety of religions represented by those terms. The New Zealand corpus had greatest number of significant terms describing the natural world. The terms found to be most typical of US English did not so much reveal vocabulary differences between the US and the other countries, but lexical differences due to alternative spellings. Trudgill and Hannah (1994: 2) describe the two main varieties of standard English as being US and British English, with the other five Englishes studied here as being more closely related to British English. Words pertaining to education and medicine appear to be more typical of the East African corpora, but we must be guarded in making this judgement since the text sampling strategy used for the East African corpora was different to that used for the other five corpora. In particular, the East African corpora, unlike the other five, had sections of newspaper texts (persuasive column and persuasive editorial) which might contain suggestions for policy in these three areas. However, these two are the only significant East African corpora currently available.

'Simpson's paradox refers to the reversal of direction of a comparison or an association when data from several groups are combined to form a single group' (Moore and McCabe, 1993: 190). The actual reversal of direction of an association is an extreme case of the phenomenon which might affect a comparison of unbalanced corpora. In a less extreme case, the strength of association might vary between literature types. For example, the word *goal* might be more common in sports coverage and therefore if one corpus contains more sports coverage, we might inadvertently think that 'goal' is more characteristic of that corpus. To avoid this difficulty, the main problem with unbalanced data, it would be possible to calculate a separate chi-squared statistic in each literature category. First these chi-squared values should be summed to find a single overall value. The original degrees of freedom should be multiplied by the number of categories, enabling us to find whether there is a significant difference between the corpora or not. When looking at the individual contributions of each word in each category to the overall chi-squared value, we may find that some significant words might not be significant in every category.

Kilgarriff (1996) discusses the issue of 'clumpiness', i.e. where a word has occurred in a text, the chances of it occurring again increase, and hence the words in a corpus are not altogether independent. Similarly, words are not completely independent if they form collocations, such as *football* and *goal*. As discussed in Section 2, the effect of this is to make our test more conservative.

This study has examined only single word orthographic vocabulary differences between the corpora used to represent the English used in seven different countries. The chi-squared method is readily amenable to identifying sequences of two or more words typical of a corpus, possibly identified by the incorporating multi-word expression (MWE) tools, but, as noted by Oakes (2003), repeated word sequences are considerably more rare in a corpus than repeated individual words.

Other studies examine the relative frequencies of word lemmas, use stop lists or look at the relative frequencies of collocations (Granger, 1998: 31). The fact that *labour* is more typical of British English and *labor* is more typical of US English is due mainly to an orthographic difference, but also because the *Labour* party is a major political party in Britain. To examine the effect of this cultural rather than the orthographic difference, it would be possible to regard both orthographic forms as equivalent, and to obtain χ^2 values for *labor/labour* across the corpora.

This study has considered only words in isolation. It would be interesting to conduct a deeper study of potentially interesting words in context, using a concordancer. This, for example, would determine whether the names of towns such as *Dunedin* and *Invercargill* are being used in a geographic sense or in a sporting sense, where they refer to their respective towns' rugby or cricket teams. Similarly, some of the proper names might be the names of athletes, and in such cases should be considered as sporting terms. Tagging each potentially interesting word according to its context would allow the use of the chi-squared test to find which word-meanings rather than which word-types are typical of a given corpus.

In this study, using the existing ICAME corpora as far as possible, we have identified vocabulary differences which can be used as potential discriminators for an automatic classifier which classifies texts of unknown provenance according to their most likely variety of English. The methodology described in this article will enable improved lists of vocabulary differences to be derived from cultural corpora, which represent the range of topics found in a national variety, when these become available.

References

Altman, D. G. (1991). *Practical Statistics for Medical Research*. London: Chapman and Hall.

Bain, L. J. and Engelhardt, M. (1997). *Introduction to Probability and Mathematical Statistics*. Boston: Duxbury.

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* B, **57**(1): 289–300.

Cox, D. R. and Miller, H. (1965). *The Theory of Stochastic Processes*. London: Chapman and Hall.

Crystal, D. (1997). *English as a Global Language*. Cambridge: Cambridge University Press.

Granger, S. (1998). The Computer learner corpus: a versatile new source of data for SLA research. In Granger, S. (ed.), *Learner English on Computer*. London and New York: Longman, pp. 3–18.

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics*, **29**: 205–20.

Hofland, K. and Johansson, S. (1982). *Word Frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities/London: Longman.

Hofland, K., Lindebjerg, A., and Thunestvedt, J. (1999). *ICAME Collection of English Language Corpora*. The HIT Centre, University of Bergen, Norway. http://www.hit.uib.no/icame/cd.

Hudson-Ettle, D. M. and Schmied, J. (1999). Manual to Accompany the East African Component of the International Corpus of English ICE-EA. See the *ICAME Collection of English Language Corpora* disc.

Juilland, A., Brodin, D., and Davidovitch, C. (1970). *Frequency Dictionary of French Words*. The Hague: Mouton.

Kilgarriff, A. (1996). *Which Words are Particularly Characteristic of a Text? A Survey of Statistical Approaches*. Information Technology Research Institute, University of Brighton.

Lyne, A. A. (1985). *The Vocabulary of French Business Correspondence*. Geneva and Paris: Slatkine-Champion.

Lyne, A. A. (1986). *In* Praise of Juilland's D. *Methodes Quantitatives et Informatiques das l'Etude des Textes*, **2**: 587. Geneva and Paris: Slatkine-Champion.

Leech, G. and Fallon, R. (1992). Computer corpora – what do they tell us about culture? *ICAME Journal*, **16**: 29–50.

Miller, R. G. (1981). *Simultaneous Statistical Inference*, 2nd edn. New York: Springer Verlag.

Moore, D. S. and McCabe, G. P. (1993). *Introduction to the Practice of Statistics*, 2nd edn. New York: Freeman.

Nelson, G. (1996). The Design of the Corpus. In Greenbaum, S. (ed.), *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press, pp. 27–35.

Oakes, M. P. (2003). Contrasts Between US and British English of the 1990s. In Oleksy, E. H. and Lewandowska-Tomaszczyk, B. (eds), *Research and Scholarship in Integration Processes*. Łódź: University of Łódź Press, pp. 213–22.

Plackett, R. L. (1981). *The Analyis of Categorical Data*, 2nd edn. High Wycombe: Griffin.

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.

Scott, M. (1999). Wordsmith Tools, Version 3, Oxford: Oxford University Press. Version 4 available at www.lexically.net/wordsmith (accessed 29th October 2004).

Storey, J. D. (2002). A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society,* B, **64 Part 3**: 479–98.

Trudgill, P. and Hannah, J. (1994). *International English*, 3rd edn. Edward Arnold.

Wellington Corpus of New Zealand English: http://khnt.hit.uib.no/icame/manuals/wellman/INDEX.HTM.