# "Inadequacy of the chi-squared test to examine vocabulary differences between corpora"

Bestgen, Yves

**Abstract**

Pearson's chi-squared test is probably the most popular statistical test used in corpus linguistics, particularly for studying linguistic variations between corpora. Oakes and Farrow (Literary and Linguistic Computing, 2007, 22, 85-99) proposed various adaptations of this test in order to allow for the simultaneous comparison of more than two corpora, while also yielding an almost correct Type I error rate (i.e. claiming that a word is most frequently found in a variety of English, when in actuality this is not the case). By means of resampling procedures, the present study shows that when used in this context, the chi-squared test produces far too many significant results, even in its modified version. Several potential approaches to circumventing this problem are discussed in the conclusion.

Document type : *Article de périodique (Journal article)*

## Référence bibliographique

Bestgen, Yves. *Inadequacy of the chi-squared test to examine vocabulary differences between corpora.* In: *Literary and Linguistic Computing : the journal of digital scholarship in the humanities*, Vol. 29, no. 2, p. 164-170 (2014)

DOI : 10.1093/llc/fqt020

# Inadequacy of the chi-squared test to examine vocabulary differences between corpora

Yves Bestgen

Center for English Corpus Linguistics - Université catholique de Louvain

Yves Bestgen
Université catholique de Louvain
Place du cardinal Mercier, 10 B-1348 Louvain-la-Neuve Belgium
yves.bestgen@uclouvain.be
(+32) 10 473005

## Abstract

Pearson's chi-squared test is probably the most popular statistical test used in corpus linguistics, particularly for studying linguistic variations between corpora. Oakes and Farrow (Literary and Linguistic Computing, 2007, 22, 85-99) proposed various adaptations of this test in order to allow for the simultaneous comparison of more than two corpora, while also yielding an almost correct Type I error rate (i.e. claiming that a word is most frequently found in a variety of English, when in actuality this is not the case). By means of resampling procedures, the present study shows that when used in this context, the chi-squared test produces far too many significant results, even in its modified version. Several potential approaches to circumventing this problem are discussed in the conclusion.

## 1. Introduction

Pearson's chi-squared test is probably the most popular statistical test used in corpus linguistics, especially when research aims to describe the lexical variations between corpora (Rayson *et al.*, 2004). This test is applied to a contingency table that is built from the frequency with which a word occurs in relation to all other words in two corpora. The null hypothesis being tested is that the difference between the frequencies with which a word is used in the two corpora results only from random variations, with the two compared samples having been extracted randomly from a single population. The test formula is:

[1] $$\sum \frac{(O-E)^2}{E}$$

where *O* refers to the observed frequency, *E* to the expected frequencies computed from the marginal totals, and the sum is calculated across the four cells of the table. This statistic approximates a chi-squared distribution when the null hypothesis is true. When comparing two corpora using the chi-squared test, there are as many contingency tables that are built and analyzed as there are words to be tested.

Hofland and Johansson (1982) took advantage of this test in order to identify the difference between the frequencies with which words occur in British English versus American English based on the comparison of two corpora consisting of one million words. This test was also used to compare corpora which are differentiated according to their sources (Baker, 2004), their genres (Tribble, 2000) or their oral or written modalities (Rayson *et al.*, 1997). More recently, Oakes and Farrow (2007) proposed an extension of this procedure to allow for the simultaneous comparison of more than two corpora, while also yielding an almost correct Type I error rate (i.e. claiming that a word occurs more frequently in a particular variety of the English language than in others, even though this is not the case). In their first experiment, they compared five corpora, each containing hundreds of English excerpts from five countries: the ACE corpus (Australia), the FROWN corpus (the United States), the FLOB corpus (Britain), the Kolhapur corpus (India), and the Wellington corpus (New Zealand). The methodology proposed by Oakes and Farrow consisted of building a contingency table with the five corpora listed in columns, and each of the different words being listed in rows. To determine whether a word is significantly more frequent in a corpus than in the others, they derived the standardized residual (Haberman, 1973) from the participation of each cell in the total chi-squared statistic (see Formula [1]). The formula for the standardized residual is:

[2] $$(O-E)/\sqrt{E}$$

This residual is positive when a word occurs too frequently in a corpus when compared to the expected frequency; the residual is a negative value when the opposite is true. Considering that its distribution is approximately normal, Oakes and Farrow determined the degree of significance of the residual by comparing it to this distribution.

The procedure described above is a simple extension of the chi-squared test, which is classically used in corpus linguistics, to a contingency table containing more than two rows and two columns. As highlighted by Oakes and Farrow, this procedure (like the classical chi-squared test that is applied to two corpora) runs up against two pitfalls regarding the Type I error rate it produces. The first pitfall results from the large number of tests (often more than 10,000) that are conducted in these kinds of studies (Gries, 2005). The rejection threshold ($\alpha$) of the null hypothesis (typically 0.05) is valid for one test; it corresponds to the error rate (i.e. the probability of committing a Type I error, or in other words, rejecting the null hypothesis given that it is true) of each test. If two independent tests are carried out at the threshold of 0.05, the error rate for each test remains at 0.05, but the familywise error rate (FWER) (i.e. the probability of wrongly rejecting the null hypothesis in at least one of the two tests) is much higher since it is equal to $1-(1-0{,}05)^2$ (or 0.0975). For three tests, this probability is 0.1426. To counteract the fact that the probability of committing a Type I error increases with each additional test, Oakes and Farrow (2007) recommended the use of the Bonferroni correction for multiple comparisons, which works equally as well irrespective of whether the tests are independent or not. Based on the principle of

Bonferroni inequality, which states that the probability of the occurrence of one or more events can never exceed the sum of their individual probabilities (Howell, 2007, p. 356), this method consists of dividing the desired $\alpha$ by the number of tests carried out to get $\alpha'$, and this value is used as the significance threshold. This $\alpha'$ guarantees (all other things being correct) that the probability that one or more of the tests (out of all the tests that are carried out) will yield significant results by chance alone is at most $\alpha$. For the 509,920 tests that Oakes and Farrow performed in their study, the $\alpha'$ was set at $1.961 \times 10^{-9}$ to get a FWER of 0.001.

The second pitfall is rooted in the sampling unit at the origin of the contingency table being tested. In order to ensure that the chi-squared test is valid, each observation included in the table must have been selected from the corresponding population through a random process. In other words, the unit of analysis must be the same as the unit of sampling (Baroni and Evert, 2009; Evert, 2006; Oakes and Farrow, 2007); this is not the case in this kind of corpus comparison since the analyzed unit is the actual word, whereas the sampling unit used to build the corpus is a complete text or an excerpt. Why is this discrepancy between the unit of sampling and the unit of analysis problematic? It has been known for a long time that the frequency with which a word occurs varies greatly from text to text (Church, 2000; Lafon, 1980). It follows that the presence or absence of a specific text in a corpus may be sufficient enough to significantly increase the frequency with which certain words occur.

This is perfectly illustrated in the following example provided by Oakes and Farrow (2007). These authors observed that one of the most typical words of British English, according to the chi-squared test, is *thalidomide*; however, the authors point out that each of the 55 occurrences of this word in the corpus in question occurs within a single text. Contrary to what is suggested by the findings of the chi-squared test, *thalidomide* is not typical of British English; it was only typical within one text of the British corpus. If, at the time of the constitution of this corpus, the unit of sampling had coincided with the unit of analysis (the word), *thalidomide* would have had (almost) no chance of being declared typical.

To account for the discrepancy between the unit of sampling and the unit of analysis, Oakes and Farrow (2007) proposed to control the dispersion of significant words in order to eliminate those words whose dispersion in the corpus (in which they typically occur) is insufficient. To do this, they divided each corpus into five sections and combined two indices: the number of sections in which the word appears, and Juilland's *D*, which is based on the coefficient of variation (V) calculated by the occurrence frequencies of a given word in each section. Its formula is:

[3] $$D = 1 - \frac{V}{\sqrt{n-1}}$$

where n corresponds to the number of sections delimited in the corpus, and *D* varies between 0 and 1; the closer it is to 1, the better the dispersion of the word. Oakes and Farrow (2007) proposed a threshold for each one of these indices: in order for a word to be regarded as sufficiently dispersed, it must appear in at least three of the five sections, and Juilland's *D* must be at least 0.30.

## 2. Does Oakes and Farrow's Procedure Guarantee the Correct Type I Error Rate?

If Oakes and Farrow’s modifications of the classical chi-squared test used in corpus linguistics seem useful, it is unclear whether these modifications make it possible to attain the correct Type I error rate, as claimed. More specifically, one may wonder whether the dispersion control method they proposed is sufficient to correct the problems created by the discrepancy between the unit of sampling and the unit of analysis. If this is not the case, the Type I error rate could be much higher than desired. Several authors who have drawn attention to this problem, suggested that this issue is attenuated when very large corpora are compared (Baroni and Evert, 2009; Oakes and Farrow, 2007). However, one can hypothesize that the more text a corpus contains, the more often the problem is likely to occur since each text can bring its own batch of typical words, which may unduly increase their occurrence frequencies; empirical verification is therefore necessary. To my knowledge, only Kilgarriff (1996, 2005) attempted such a demonstration by compiling two arbitrarily pseudocorpora by randomly sampling texts (and not words) from a single corpus (the

British National Corpus). As Kilgarriff underlines it, one must expect that any differences between the frequencies of the words in the pseudocorpora exclusively result from random variations, and therefore the chi-squared test rejects the null hypothesis of equal frequencies in 0.5% of cases if a probability threshold of 0.005 is used. Kilgarriff observes that 'For very many words, including most common words, the null hypothesis is resoundingly defeated' (2005, p. 269). However, as pointed out by Gries (2005), Kilgarriff 's experiment has two important limitations.[1] Having been carried out only once, it is possible that the two pseudocorpora are quite different from one another only by virtue of chance. In addition, Kilgarriff (2005) does not apply the Bonferroni correction. It can be further argued that since this experiment was conducted prior to Oakes and Farrow's study, Kilgarriff's experiment could not take into account the control of dispersion as proposed by the latter researchers to avoid this pitfall.

The study presented here aims to answer the question of whether the Oakes and Farrow methodology can achieve the nominal Type I error rate through simulations based on a resampling procedure inspired by Kilgarriff's study (2005). These simulations were carried out on Oakes and Farrow's data in order to evaluate the extent to which the proposed solutions were effective. If we can show that their solutions are effective, further research of this type in corpus linguistics would benefit from using these methods. If these solutions were insufficient, the consequences would be even more problematic for studies that do not apply them.

## 3. Method, Analyses, and Results

### *3.1 Reproducibility of the Original Results of Oakes and Farrow (2007)*

To replicate Oakes and Farrow's analyses (2007), I used the same ICAME CD-ROM (Hofland *et al.*, 1999), which contains the raw texts of the five corpora. A series of pretreatments had to be applied to the texts such as word segmentation and special characters removal. These pretreatments were likely different from those made by Oakes and Farrow, and so preliminary analyses were carried out in order to ensure that the results from their data were reproducible. The most significant words obtained in the present study were compared to the 50 most significant words from each corpus reported in Table 9 by Oakes and Farrow (2007, p. 96). Overall, the correlation between the two sets of values is 0.998, and the average deviation between those is less than 4%. Some differences were nevertheless observed, but they were almost always limited to words close to one of the thresholds used in the control of dispersion.

### *3.2 Estimation of the Type I Error Rate by Means of Simulations*

To determine whether the methodology proposed by Oakes and Farrow can obtain the expected Type I error rate, a computer program was written in C to perform a series of simulations based on a resampling procedure of the permutation type. A simulation consists of building five pseudocorpora, each one containing the same number of texts as the original corresponding corpus; these texts were selected randomly and without replacement of the whole set of texts from the five original corpora. Together, the five pseudocorpora contained exactly the same words and occurred in the same frequencies as the five original corpora; the only difference was that the texts from the original corpora were randomly permuted to create the pseudocorpora. A contingency table was then built (as was the case with the original corpora) and submitted to the hypothesis testing procedure proposed by Oakes and Farrow. The number of words that could lead to the rejection of the null hypothesis at the $\alpha'$ threshold (i.e. $1.961 \times 10^{-9}$) and which passed the control of dispersion successfully was recorded. As the pseudocorpora were generated in a completely random way, it is expected that all null hypotheses are true. Given the very low value of $\alpha'$, the null hypothesis should never be rejected, but as pointed out by Gries (2005) in his comments about Kilgarriff's study, the effects of chance can be misleading. The simulation was thus replicated 10,000 times by varying the seed used in the random generator to assign the texts to the five pseudocorpora. Of these 10,000 simulations, and while using $\alpha'$ as the significance threshold, a dozen simulations are expected to include one or more significant tests. If this is the case, then the Oakes and Farrow methodology produces the correct FWER.

After carrying out the numerous simulations, it became apparent that the results were quite different than those expected. All 10,000 simulations contained at least one statistically significant test and even much more since there were an average of 170 significant tests in a simulation (min = 119, max = 223). It is important to note that the results would be much worse if the control of the word dispersion, as proposed by Oakes and Farrow, were not applied. In this case, no less than an average of 577 words (min = 510, max = 640) per simulation were statistically significant. This underlines the importance of conducting a control of dispersion in these types of corpus studies.

Should we conclude that the results obtained by Oakes and Farrow are worthless? Definitely not. Based on the actual data, the authors achieved significance among 1050 words at $\alpha'$ for the five corpora. In the simulations presented here, I attained an average of 170 significant tests by simulation (that is 6.2 times fewer than those noted by the original authors). Therefore, there are clear differences between the actual corpora; however, a substantial proportion (16%) of the words reported as statistically significant may have occurred by chance alone.

## 4. Discussion and Conclusion

The simulations reported above show that the chi-squared test when used to examine vocabulary differences in corpora produced a Type I error rate that was far too high. Concluding in such a negative way would be unfortunate. This section discusses several potential approaches to circumvent the inflated Type I error rate problem.

Since the dispersion control method proposed by Oakes and Farrow is very beneficial (albeit insufficient), and given that the thresholds they proposed are necessarily arbitrary (Oakes and Farrow, 2007, p. 92), one might consider making these thresholds stricter. I therefore determined the FWER and the corresponding number of significant words for certain parameter values involved in this control. If one requires that the word appear in each of the five sections (not in at least three), the FWER stays at 1; on average, there were 92 significant words that occurred by simulation (min = 54; max = 137). The parameter $D$ has a much greater impact. Set at 0.60 instead of 0.30, it leads to a FWER of 1, but the average number of significant tests that resulted from the simulations went down to 36 (min = 13; max = 70). When $D$ was set at 0.90, 35% of the 10,000 simulations contained at least one significant test. This is still too high since the value should have been 0.1%.[2] The major problem with this approach is that it strongly affects the number of significant words that occur in the original data. Setting $D$ at 0.90 makes it possible to declare that only 27 words were significant for the five corpora (instead of 1070), and that these words are often very frequent words like *and*, *any*, *in*, *it*, *not*, *the*. In an attempt to avoid being too liberal, the procedure becomes too conservative and produces far too many Type II errors (i.e. not rejecting the null hypothesis when it is false).

A second possible solution exhibits the same weakness as the previous one. It consists in using an even more extreme probability threshold to effectively achieve a Type I error rate consistent with that of Oakes and Farrow expectations (a FWER of 0.001). The simulations reported in Section 3.2 render it possible to see what this approach could produce. To execute this method, the maximum standardized residual of each of the 10,000 simulations was identified, and the eleventh largest value among them was recorded (22.80). If we set the critical value for the test just above that number, only 10 of 10,000 simulations would have led to at least one significant test, and this would have corresponded to a FWER of 0.001, which was desired. The impact of such a critical value on the actual results obtained by Oakes and Farrow would have meant that only 31 significant words would have been obtained in the five corpora, instead of the 1070 significant words that were obtained using the original threshold. Among the terms that ceased to be statistically significant, there are many which appeared to be typical of a variety of English dialects such as *center*, *color,* and *behavior* for American English, *caste* and *upto* for Indian English, or *Aborigines* for Australian English. This observation simply confirms a classical observation that being too strict on the Type I error rate dramatically increases the Type II error rate.

A third solution is to use an inferential test adapted to the actual sampling unit (i.e. the text). Kilgarriff (1996) recommends using the nonparametric Wilcoxon-Mann-Whitney test which is carried out, not on the total frequency with which words occur in each corpus, but on the occurrence frequency (transformed into rank) of each word in each text. This proposal has not been successful in corpus linguistics (however, see Paquot and Bestgen (2009) for a comparison with the chi-squared test), and was criticized by Rayson and colleagues (Rayson *et al.*, 2004; Rayson and Garside, 2000) because this test would only allow for the analysis of the most frequently occurring words and because it neglects to take into account some important information available in the data due to the transformation of the frequencies into ranks. A permutation test can also be called upon to preserve the use of inferential statistics while ensuring a correct Type I error rate. This type of test, which has been gaining more and more attention in statistics (Howell, 2007) as well as in corpus linguistics, rests on the very same simulation procedure reported above, but with the specific aim of deciding whether a word occurs more frequently in one corpus than in the others. To do this, the standardized residual of any word obtained from an original corpus is compared to all the standardized residuals obtained for the same word across a large number of permutations carried out as explained in Section 3.2. The proportion of permutations that produces, for each word, a value higher than that actually observed is indicative of statistical significance.[3] Despite these benefits, this approach suffers from a major drawback: it requires a large number of simulations to reach a significance level as strict as the one required by Oakes and Farrow (i.e. one that is equal to $1.961 \times 10^{-9}$). One must carry out 509,920,000 simulations in order to check whether no more than one of these residuals is greater than the actual residual.

A final alternative is to interpret the chi-squared values (or the standardized residuals) as indicators of the potential interest of each of the numerous vocabulary differences between the corpora. In this way, the chi-squared test is applied as it should be in these kinds of studies: as a statistical tool for exploratory research that allows for the identification of words that deserve deeper analysis, and not as an instrument of confirmatory analysis, whose objective is to accept or reject specific null hypotheses (Gries, 2005). This is probably the most reasonable approach since it only requires that researchers refrain from making reference to inferential statistics in these types of studies.

*Notes*

[1] Gries (2005) reports a series of simulations, but those relate to the comparison of individual texts and not to the corpora made up of numerous texts. Therefore, they cannot be used in this discussion.

[2] By making the two parameters more strict (simultaneously) does not help in this instance because a very high *D* can only be obtained for words that occur in every section.

[3] The main differences between this approach and the one proposed above to set α' at a level that ensures the right FWER is that the present approach does not require the use of the control of dispersion method and, most importantly, that the probability is calculated specifically for each word.

*Author Note*

Yves Bestgen is a Research Associate with the Belgian Fund for Scientific Research (F.R.S-FNRS). A preliminary version of this study has been presented at the *$11_{es}$ Journées Internationales d'Analyse Statistique des Données Textuelles*, Liège (Belgique), June 2012. [Bestgen, Y. (2012). Analyse des Différences Lexicales Entre des Corpus: Test ou distance du Khi-2? *Actes des JADT'12* (pp. 150-161) [http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Bestgen,%20Yves%20-%20Analyse%20des%20differences%20lexicales%20entre%20des%20corpus.pdf].

*References*

Baker, P. (2004). Querying Keywords: Questions of Difference, Frequency and Sense in Keyword Analysis, *Journal of English Linguistics,* 32: 346-359.

Baroni, M. and Evert, S. (2009). Statistical Methods for Corpus Exploitation. In Lüdeling, A. and Kytö, M. (eds), *Corpus Linguistics. An International Handbook*, Mouton de Gruyter, pp. 777-803.

Church, K. (2000). *Empirical Estimates of Adaptation: The Chance of Two Noriegas is Closer to p/2 than p2, Proceedings of the 17th Conference on Computational Linguistics*, pp. 180-186.

Evert, S. (2006). How Random is a Corpus? The Library Metaphor, *Zeitschrift fu Anglistik und Amerikanistik*, 54: 177-190.

Gries, S. (2005). Null Hypothesis Significance Testing of Word Frequencies: a Follow-up on Kilgarriff, *Corpus Linguistics and Linguistic Theory*, 1: 277-294.

Groom, N. (2010). *Closed-Class Keywords and Corpus-Driven Discourse Analysis*. In Bondi, M. and Scott, M. (eds.), *Keyness in Texts*, John Benjamins, pp. 59-78.

Haberman, S. J. (1973). The Analysis of Residuals in Cross-Classified Tables, *Biometrics*, 29: 205-220.

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

Hofland, K., Lindebjerg, A. and Thunestvedt, J. (1999). *ICAME Collection of English Language Corpora*. Norway: The HIT Centre, University of Bergen.

Howell, D. (2007). *Statistical Methods for Psychology*. Belmont, CA: Duxbury Press.

Kilgarriff, A. (1996). Comparing Word Frequencies Across Corpora: Why Chi-Square Doesn't Work, and an Improved LOB-Brown Comparison, *Proceedings of ALLC-ACH Conference*, pp. 169-172.

Kilgarriff, A. (2005). Language is Never, Ever, Ever Random, *Corpus Linguistics and Linguistic Theory*, 1: 263-275.

Lafon, P. (1980). Sur la Variabilité de la Fréquence des Formes dans un Corpus, *Mots*, 1: 127-165.

Oakes, M. 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: 85-99.

Paquot, M., and Bestgen, Y. (2009). Distinctive Words in Academic Writing: A Comparison of Three Statistical Tests for Keyword Extraction. In Jucker, A. H., Schreier, D. and Hundt, M. (eds), *Corpora: Pragmatics and Discourse*, Rodopi, pp. 247-269.

Rayson, P., Berridge, D. and Francis, B. (2004). *Extending the Cochran Rule for the Comparison of Word Frequencies Between Corpora, Proceedings of the 7th International Conference on Statistical analysis of textual data*, pp. 926-936.

Rayson, P. and Garside, R. (2000). Comparing Corpora Using Frequency Profiling, *Proceedings of the workshop on Comparing Corpora*, 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: 133-152.

Tribble, C. (2000). Genres, Keywords, Teaching: Towards a Pedagogic Account of the Language of Project Proposals. 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, pp. 75-90.