# Getting rid of the Chi-square and Log-likelihood tests for analysing vocabulary differences between corpora

## Analizar las diferencias de vocabulario entre corpus sin los tests Chi-cuadrado y Log-likelihood

Yves Bestgen

Université catholique de Louvain. yves.bestgen@uclouvain.be
Received: 20/04/2017. Accepted: 11/10/2017

**Resumen:** Los tests de *log-likelihood* y *chi-cuadrado* probablemente sean las pruebas estadísticas más populares utilizadas en la lingüística de corpus, especialmente cuando la investigación tiene como objetivo describir las variaciones léxicas entre corpus distintos. Sin embargo, dado que este uso específico del chi-cuadrado no es válido, produce demasiados resultados significativos. Esta contribución explica el origen del problema (es decir, la no independencia de las observaciones), los motivos por los cuales las soluciones habituales no son aceptables y qué clase de pruebas estadísticas deben ser utilizadas en su lugar. Se ha realizado un análisis de corpus sobre las diferencias léxicas entre el inglés británico y el inglés americano para mostrar el problema y confirmar la adecuación de la solución propuesta. La última sección presenta las órdenes que pueden darse a *WordSmith Tools*, un programa informático muy popular en el procesamiento de corpus, a fin de obtener los datos necesarios para las pruebas adecuadas, así como un procedimiento muy fácil de usar en R, un paquete estadístico gratuito y fácil de instalar, que realiza estas pruebas.

**Palabras clave:** diferencias léxicas entre corpus; test de remuestreo; Wordsmith tools; inglés británico y americano.

---

**Abstract:** Log-likelihood and Chi-square tests are probably the most popular statistical tests used in corpus linguistics, especially when the research is aiming to describe the lexical variations between corpora. However, because this specific use of the Chi-square test is not valid, it produces far too many significant results. This paper explains the source of the problem (i.e., the non-independence of the observations), the reasons for which the usual solutions are not acceptable and which kinds of statistical test should be used instead. A corpus analysis conducted on the lexical differences between American and British English is then reported, in order to demonstrate the problem and to confirm

» Bestgen, Yves. 2017. "Getting rid of the Chi-square and Log-likelihood tests for analysing vocabulary differences between corpora". *Quaderns de Filologia: Estudis Lingüístics* 22: 33-56. doi: 10.7203/qf.22.11299

the adequacy of the proposed solution. The last section presents the commands that can be used with WordSmith Tools, a very popular software for corpus processing, to obtain the necessary data for the adequate tests, as well as a very easy-to-use procedure in R, a free and easy to install statistical software, that performs these tests.

**Keywords:** lexical differences between corpora; resampling test; WordSmith Tools; British and American English.

## 1. Introduction

Many studies in corpus linguistics aim to analyse lexical differences between corpora of different genres (Tribble, 2000), their regional and diatypic varieties (Oakes & Farrow, 2007), their oral or written modalities (Rayson, Leech & Hodges, 1997), the period of writing (Laviosa, Pagano, Kemppanen & Ji, 2017) or certain sociological characteristics of the speaker or writer, such as gender, age and socio-economic status (Brezina & Meyerhoff, 2014; Marquilhas, 2015), to cite a few examples. This kind of study immediately raises the question of how to decide whether a difference observed when comparing two given corpora (i.e., more occurrences of *towards* or *male* in an American English as opposed to a British English corpus) is purely accidental, or whether it reflects a real difference in the way English is used. The answer is typically provided through the use of the Pearson's Chi-square ($Chi^2$) test or its close neighbour, the log-likelihood (LL) test (Biber & Jones, 2008; Rayson & Garside, 2000).

These statistical tests are applied to a contingency table made up of the frequency of a word in the two corpora to be compared and the total number of words in each corpus. Table 1 shows the contingency tables for the words *towards* and *male* in the British English corpus FLOB and in the American English corpus FROWN, which are used in the empirical analyses reported in section 4.

| | *British* | *American* |
|---|---|---|
| Towards | 17 | 293 |
| ~Towards | 1016832 | 1018360 |

| | *British* | *American* |
|---|---|---|
| Male | 89 | 177 |
| ~Male | 1016760 | 1018476 |

Table 1. Frequency counts for two words in the FLOB and FROWN corpora

The null hypothesis tested is that the difference between the frequency of use in the two corpora is only the result of random variations, the two samples compared being randomly extracted from a single population. The statistics used are:

$$Chi^2 = \sum \frac{(O\text{-}E)^2}{E} \qquad LL = 2 \sum O \ ln \left(\frac{O}{E}\right)$$

in which *O* represents the observed frequency and *E* the expected frequency, computed on the basis of the marginal totals, and the summation is over the four cells (and not over the first two as in Brezina and Meyerhoff (2014)).

Under *H0*, these statistics are approximately distributed as a Chi-square with one degree of freedom, which makes it possible to calculate the probability of obtaining a statistic at least as high as that which would be observed if the differences were due to chance alone. Applied to the words *towards* and *male*, these two tests return probabilities of less than 0.00000001.

As noted by Sampson (2003), the use of these tests has, for example, expanded our understanding of the lexical differences between British and American English by Hofland and Johansson (1982). These authors showed that masculine words, such as *he, boy* and *man*, are significantly more frequent in American English, while feminine words are significantly more frequent in British English.

The popularity of these tests has undoubtedly been reinforced by their implementation in a software as frequently used in corpus linguistics as *WordSmith Tools* (Scott, 1997), one of its main functions being the identification of *Keywords*, i.e., all words that successfully pass the $Chi^2$ or LL tests for a probability threshold of 0.000001. This same function is also available in other software, such as *AntConc* (Anthony, 2012). These two tests are also very frequently used to test specific hypotheses in corpus linguistics (Lee & Chen, 2009; Lubbers Quesada & Blackwell, 2009; see Gablasova, Brezina & McEnery (2017) for illustrations and a discussion). For example, Siyanova-Chanturia (2015) used the $Chi^2$ test to confirm that Chinese beginner learners of Italian used more strongly associated collocations at the end of an intensive course than they did at the beginning.

However, these tests, according to the way they are used to analyse lexical differences between corpora, are inadequate, as has already been pointed out by several authors, and should no longer be used (Bestgen, 2012, 2014; Brezina & Meyerhoff, 2014; Kilgarriff, 1996, 2005; Lijffijt, Nevalainen, Säily, Papapetrou, Puolamäki & Mannila, 2016). The aim of this paper is to help researchers to abandon them by explaining in detail the problem they pose and its origin, by showing why several possible solutions are ineffective and by recommending two valid and efficient statistical tests. To make the use of these adequate tests as sim-

ple as possible, the last section provides the commands to obtain the necessary data by means of *WordSmith Tools* and a very easy-to-use script in *R*, a free and easy to install multi-operating system statistical software, to perform them.

## 2. The problem

The use of these two tests in corpus linguistics has been criticized for the very large number of significant differences they claim to detect (Baker, 2004; Gries, 2005; Kilgarriff, 1996, 2005). For example, Paquot and Bestgen (2009) observed, when comparing a literary corpus and an academic corpus of 15 million words each, that more than 90% of the 10,333 words tested were significantly more frequent in one of the two corpora for a probability threshold of 0.000001. The origin of this problem was most often explained by the very large sample size under analysis (Kilgarriff, 2005) or in the large number of tests performed (Gries, 2005). The problem is, in fact, much deeper and does not arise only in linguistics. It was mentioned by Lewis and Burke as early as 1949 as the main misuse of the $Chi^2$ test in psychology, and has been repeatedly emphasized since then: "Chi-square may be correctly used only if all N observations are made independently" (Kurtz & Mayo, 1979: 366); that is, each observation must be "taken from the population at random, and the selection of each member of the sample is independent from the next" (Wallis, 2013: 352). In other words, for the test to be valid[1], the unit analysed must be the sampling unit (Bestgen, 2014; Gablasova *et alii.*, 2017). This is (almost) never the case in corpus linguistics. The unit analysed is often a word, or sometimes a sentence, while the sampling unit used to construct the corpus is a text (or an extract from a text).

Why does this discrepancy between the sampling unit and the unit of analysis so strongly affect the number of significant words in corpus comparison? It has long been known that the frequency of word occurrences varies greatly between texts (Church, 2000). It follows that the presence of some very specific texts, or even a single one, in a corpus may be sufficient to increase the frequency of certain words and thus

[1] This problem arises for all statistical tests that can be applied to a contingency table, including Fisher's exact test, which also requires the observations to be independent.

to modify the words considered as being significantly more frequent in this corpus according to the $Chi^2$ and LL tests. This phenomenon is perfectly illustrated in the following example reported in Oakes and Farrow (2007). These authors observed that one of the most typical words in British English, according to the $Chi^2$ test, is *thalidomide*. They note, however, that all of the 55 occurrences of this word in the British corpus appear in one single text. Contrary to what the $Chi^2$ test seems to indicate, *thalidomide* is not typical of British English, just of one text in the British corpus. It is because this text has been selected in its entirety for inclusion in the corpus that *thalidomide* appears as typical. If the sampling unit had coincided with the unit of analysis (the word), *thalidomide* would have had (virtually) no chance of being declared typical. Thus, each selected text may cause a series of false positives. It is important to note that it is not just such extreme cases that invalidate the $Chi^2$ and LL tests. The simple fact that the probability of a word occurring in a text for a second time is far higher than that of having it for the first time, shows that non-independence is general and not occasional (Church, 2000).

## 3. The solutions

A first solution consists of disregarding the probability derived from the inferential test (Bestgen, 2014; Gabrielatos & Marchi, 2011; Leedham, 2012). The $Chi^2$ (or LL) values (called *Keyness* in *WST*) are interpreted as indicators of the potential interest of each of the numerous vocabulary differences between the corpora: the larger it is, the more interesting the word. This solution has the major drawback of only masking the problem without solving it because there is an inverse monotonic relationship between the p-value and the test statistic. A word such as *thalidomide* is extremely significant, because it has a very high $Chi^2$ value. Pretending to only look at the $Chi^2$ or the LL scores does not solve anything.

A second solution is to use a dispersion measure to eliminate words that only occur in a part of a corpus (Baker, 2004; Oakes & Farrow, 2007). The first problem with this solution is that the threshold used to decide that a word is insufficiently dispersed is necessarily arbitrary, which is all the more annoying since the main measures proposed in the literature are difficult to interpret (Oakes & Farrow, 2007). More-

over, Bestgen (2014) showed that taking dispersion into account made it possible to reduce the problem posed by very badly dispersed words (like *thalidomide*), but not to eliminate it. The LL and Chi[2] tests remain inadequate.

The only acceptable solution is to use an inferential test that reconciles the sampling units and the units of analysis and that is therefore based on the frequency of the words[2] not in the corpus, but in the texts making up the corpus. Several statistical tests are possible. The most obvious choice is the Student's t-test for comparing two means. This test, however, is problematic, because it is based on a postulate of normality which is very difficult to sustain in the case of data made of word frequencies. For this reason, Kilgarriff (1996) and several authors after him (Brezina & Meyerhoff, 2014; Lijffijt *et alii.*, 2016; Paquot & Bestgen, 2009) proposed the use of a distribution-free test (also called a nonparametric test). The test recommended by Kilgarriff is the Wilcoxon-Mann-Whitney test (WMW), which is carried out on the relative frequency of each word in each text after they have been transformed into ranks. When simplified a little, it calculates the probability of having, under the null hypothesis that the two corpora were drawn at random from identical populations of texts, a difference which is at least as important between the average ranks as that observed.

This proposal was strongly criticized by Rayson and colleagues (Rayson, Berridge & Francis, 2004; Rayson & Garside, 2000) because this test neglects to take into account some important information available in the data due to the transformation of the frequencies into ranks. However, it is easy to remedy this problem because there is a WMW-equivalent test that can be applied to the non-ranked values: the Fisher-Pitman (FP) test (Berry, Mielke & Mielke, 2002; Neuhauser & Manly, 2004). It calculates the probability, under the same null hypothesis, of obtaining a difference between the mean frequencies in texts as large as the difference actually observed. The only difference between these two tests is therefore that one is calculated on the basis of ranked data and the other on raw data.

---

[2] Since the texts in a corpus are rarely of exactly the same length, the analyses must be carried out on the relative frequencies (number of occurrences divided by the length of the text).

These two tests have some properties which are important to know in order to use them adequately. First, because they free us from the normality assumption, they test a more general null hypothesis than that tested by the Student's t-test. They also detect differences in the variability and even in the shape of the distributions. However, they are particularly sensitive to differences in mean or medians (Howell, 2008; Hesterberg, Moore, Monaghan, Clipson & Epstein, 2006).

Second, the p-value that they provide when analysing large samples, as is almost always the case in corpus linguistics, is obtained using a Monte-Carlo resampling procedure[3]. This type of test is gaining more and more attention in statistics (Good, 2005) as well as in corpus linguistics (Gries, 2006). However, its weakness is that the degree of precision of the probability depends on the number of resamplings performed and it is therefore time-consuming to obtain probability estimates for many words. This limitation is especially important when estimating extremely small probabilities, since they cannot be smaller than one divided by the number of resamplings done.

Finally, replacing the relative frequencies by ranks in the case of the WMW has the consequence that the corpus containing the fewest occurrences of a word may be the one whose texts have the highest average rank. This will be the case, for instance, if one of the two corpora only contains a single text containing many occurrences of the word, while the other corpus contains a sufficiently large number of texts containing a small number of occurrences of it. The first corpus will have the highest frequency, but the second will have the highest average rank. This difference between the two tests is not a defect. It points out words showing an atypical profile.

## 4. Empirical evaluation of the different tests

So far, studies which have stressed the inadequacy of the Chi$^2$ and LL tests for analysing lexical differences between corpora presented ar-

[3] Lijffijt *et alii.* (2016) proposed an ad hoc resampling procedure of the bootstrap type that differs from the usual practices in statistics since the resampling is done in a manner that is not consistent with the null hypothesis (Hesterberg *et alii.*, 2010) and since, when the two samples are unequal in size, the smallest sample size is used in the resampling procedure (see Efron and Tibshirani [1993, Chap. 16] for a significance test based on bootstrap).

guments using the fact that these tests declare too many words to be significant even when extremely strict probability thresholds are used (Bestgen, 2014; Brezina & Meyerhoff, 2014; Kilgarriff, 2005; Lijffijt *et alii.*, 2016). Such demonstrations have obviously not been sufficient, since these tests continue to be used in corpus linguistics and they are still the only statistical tests available in *WST* and *AntConc*. We are thus proposing another proof of the problem. We will evaluate the effectiveness of a statistical test based on what it is really used for, that is, the conclusion derived from a significant difference. If a test claims that a given word is more frequent in one variety of English than it is in another because it finds a significant difference between the frequency of this word in the two corpora, it is expected that if two other corpora that differ on the same dimension are analysed, that difference will also be observed. One can immediately see the problematic consequences resulting from a test that is not very effective according to this criterion: nobody can trust the conclusions to which it leads. This evaluation procedure is used in the analyses reported below, which were conducted on the distinction between American and British English. The statistically significant differences were determined on the basis of two corpora of one million words each, and the verification on the basis of two very large corpora, frequently used as reference corpora for the varieties in question.

## 4.1. *Materials*

### 4.1.1. Corpora for finding the significant differences

We made use of the FLOB (Freiburg LOB Corpus of British English) and the FROWN corpus (Freiburg Brown Corpus of American English), both compiled at the University of Freiburg to be as similar as possible except, of course, in terms of the variety of English. Each corpus contains a million words, corresponding to 500 extracts from texts[4] published in the early 90s. Each contains approximately 2000 words, and they comprise 15 genres of written texts, such as press texts, scientific writing, romantic fiction and science fiction. They are available on the ICAME CD-ROM (Hofland, Lindebjerg & Thunestvedt, 1999).

[4] The resampling tests do not require both corpora to contain the same number of texts.

### 4.1.2. Corpora for evaluating the test decisions

Two large reference corpora for these varieties of English were used:

- The British National Corpus (BNC), a 100-million-word collection of samples of written and spoken language designed to represent a wide cross-section of British English from the late 20th century.
- The Corpus of Contemporary American English (COCA) is a very large and balanced corpus of American English. The version we used contains more than 425 million words of text (20 million words for each year between 1990 and 2011) and is equally divided between speech, fiction, popular magazines, newspapers and academic texts.

In the following analyses, a word is considered typical of an English variety according to the reference corpus when its relative frequency is higher in the corresponding corpus.

## 4.2. *Procedure*

A series of pre-treatments had to be applied to the texts, such as word segmentation and special character removal. The same pre-processing steps were carried out on the analysed corpora (FLOB and FROWN) and on the reference corpora (BNC and COCA).

The $Chi^2$, LL, WMW and FP tests were applied to all words with a total frequency of at least 10 in the two corpora, so as to analyse only words with a sufficient expected frequency (a requirement for using the $Chi^2$ test). To estimate the p-values for the two resampling tests, one million permutations were made. The probability threshold for deciding that a word is significantly more frequent in one of the two compared corpora was set at 0.000001, which is the default value in *WST*.

The analyses were carried out twice: the first time without taking into account the dispersion criterion and the second time only considering words occurring at least in 5% of the texts of the corpus in which they have the highest relative frequency. This is the dispersion criterion, the range, which is used in *WST*, and it is set to its default value in this software. This threshold of 5% corresponds to 25 texts in these

corpora and therefore implies a minimum frequency of 25 occurrences of the word in the corpus. An advantage of the range over many other measures of dispersion is that it is easily interpretable. It is important to compare the performance of the tests with and without a dispersion threshold because few studies use them, whereas Oakes and Farrow (2007) have shown that it is useful for filtering uninteresting words when using the $Chi^2$ test.

### 4.3. *Results*

Table 2 summarizes the main results of the analyses. For each statistical test, and with or without taking the dispersion criterion into account, the number of words considered as significant at the threshold of 0.000001 is given, as well as the proportion of these words validated by the reference corpora and the number of words not validated. As can be seen, many more words are selected by the $Chi^2$ and LL tests than by the two adequate tests, confirming the criticism raised by Kilgarriff (2005). Without a control on dispersion, a non-negligible percentage of these words is not validated by the reference corpora. When the dispersion threshold is taken into account, 8% of the words selected by the two inappropriate statistical tests are rejected. For both appropriate statistical tests, the results are very different. These tests clearly select fewer words, but all of them are validated when dispersion is taken into account, and only one word is not validated when this criterion is not considered

| | *Without Range* | | | *With Range* | | |
|---|---|---|---|---|---|---|
| *Test* | *Nbr. Sig* | *%OK* | *Nbr. KO* | *Nbr.Sig* | *%OK* | *Nbr. KO* |
| CH | 577 | 83.36 | 96 | 280 | 92.14 | 22 |
| G | 805 | 81.24 | 151 | 288 | 92.01 | 23 |
| WMW | 122 | 99.18 | 1 | 113 | 100.00 | 0 |
| FP | 104 | 99.04 | 1 | 99 | 100.00 | 0 |

Table 2: Results for the four statistical tests

From a qualitative point of view, the words selected by the $Chi^2$ and the LL tests that were not validated by the corpus of reference when ap-

plying the dispersion criterion are as follows (ordered according to their keyness score): *t, i, japan, have, st, m, ai, children, opera, last, male, stress, performance, poll, has, relations, okay, legal, mental, d, yeah* and *prison*. The LL adds the word *patient* to this list. This list includes the word *male*, which has been used as an example in Table 1 and which is considered by the inadequate tests as being typical of American English. It is interesting to compare this list with the 25 words validated by the corpus of reference with the highest keyness scores: *percent, which, cent, labour, toward, program, clinton, bush, president, programs, towards, american, uk, per, states, london, labor, british, was, defense, centre, center, britain, united* and *washington*. This list includes the other example in Table 1 (*towards*). There is no doubt that the words on the second list are clearly more easily interpretable, in the sense that it is easy to guess the variety of English in which they occur most frequently, whereas it is much more difficult for the first list. The term selected by the WMW and FP tests that is not validated by the corpus of reference is *DC*, which is more frequent in the FROWN corpus than the FLOB corpus, but less frequent in the COCA than in the BNC, where it appears not only as expected after *Washington*, but also as the abbreviation of direct current and in an extract of *The Dickens Index* book.

The objective of this analysis is to illustrate the problems posed by the classical $Chi^2$ and LL tests and to show that the proposed tests do not encounter these difficulties. It is not possible to analyse the two inappropriate tests in detail, in order to determine whether it is possible to make them more efficient, by using more extreme probability thresholds or by using other dispersion measures. Such analyses would require a variation in the size of the corpora to determine whether or not an efficient solution for comparing two one-million-word corpora is also appropriate for smaller and larger corpora, or for corpora of different sizes.

## 5. How can the adequate test be obtained?

The previous section very concretely shows the problems caused by using inadequate statistical tests when analysing lexical differences between corpora. However, to persuade researchers to adopt the adequate tests, it is necessary to simplify their use as much as possible. This section presents instructions for both *WST* and *R*, which make it easy to use these tests.

### 5.1. *Getting the necessary data with WST*

The first step is to create a wordlist for each corpus containing the frequencies of all of the words in all of the texts by supplying a file per text to *WST* (after, if necessary, using the *Split* function in *File Utilities*). In the Wordlist function, use the *Make a batch now* option with *One file with all individual results in it* in zip format. Then, use the *Detailed consistency* function, where you select the zip file containing all Wordlists (one per file). Finally, save the results displayed on the screen in a *.txt* file with tab as column separator and uncheck the *Separate thousands* box. These steps must be performed separately for each corpus.

### 5.2. *The R script for computing the statistical tests*

The *R* (R Core Team, 2013) script provided in the appendix requires a complementary package, called *Coin* (Hothorn, Hornik, van de Wiel & Zeileis, 2008), which performs the resampling tests. If it is not already installed, the script tries to do so. To use this script, just copy the whole code (the *CorpLexTests* function) and paste it into the *R* console window and press *Enter*. Then, it is necessary to adapt the command line provided below to the files to be analysed and the parameter values to be used.

```
CorpLexTests(file1="E:/FLOBList.txt", file2="E:/FROWNList.txt",
minfreq = 10, minrange = 0.05, maxpll = 0.0001, niter1 = 10000, pperm =
0.0003, niter2 = 1000000)
```

The parameters are as follows:

- *file1* and *file2* provide the paths and filenames for the two files obtained from *WST* (on Windows, "E:\\FLOBList.txt" works as well).
- *minfreq* indicates the minimum total frequency of the word in both corpora for the analysis to be conducted. The default value is 0 and corresponds to no threshold. However, it seems meaningless to try to determine whether a very rare word is more frequent in one corpus than in another. Moreover, in addition to the problem of non-independence described above, it is known that, in order to be valid, the Chi-square test imposes a condition on

the expected frequencies in the contingency table cells (usually at least five). It should be noted that this condition does not apply to the permutation tests, but that a rare word is unlikely to be significant enough to merit a thorough linguistic analysis.

- *minrange* gives the minimum threshold of the number of texts in which the analysed word should occur. This value is given in proportion to the number of texts in the corpus containing the most occurrences (in terms of relative frequency) of this word. The default value, taken from *WST*, is 0.05.
- *maxpll* indicates the maximum p-value from the LL test for the analysis to be conducted. It must be between 0 and 1, 0 allowing only a very small number of tests and 1 allowing all tests. This function makes it possible to reduce the duration of the analysis by only performing the resampling tests on words which would have been declared significant by the usual (but problematic) LL test. The default value is set to 0.000001, as in *WST*.
- *niter1* indicates the number of resamplings to be performed for any word that successfully passes the three conditions (*minfreq*, *minrange* and *maxpll*). It is desirable not to go below 1000. The chosen value will determine the smallest probability that can be given to a word by the resampling tests. For example, 1000 corresponds to a probability of 0.001. The greater the number of resamplings requested, the more time the analyses will take. For this reason, it is possible to request additional resamplings for the most significant words using the last two parameters. This function is activated by the parameter *pperm*, which gives the maximum p-value for performing a series of complementary resamplings. It is applied independently to each of the two resampling tests. Thus, for each of these tests, if the p-value resulting from the first *niter1* resamplings is less than or equal to this parameter, a total of *niter2* resamplings is performed. *Niter2* must necessarily be greater than *niter1* (since it includes these iterations). Setting *niter2* to 1000000 by default will yield probabilities as small as 0.000001, the default threshold for *WST*. The default value of *pperm* is set to 0, and this option is thus not used.

The only required parameters are the two file paths, since all other parameters have acceptable default values. This script works both on Windows and Mac OS X (but not *WST*).

### 5.3. *R script output*

The results are displayed on the *R* Console and saved in a file named *CorpLexTestsRes.txt* in the folder where corpus 1 is located. The first fourth lines give general information about the analyses performed. The first two lines show the file path and name of each corpus, as well as the number of texts and the number of words they contain. This line thus serves as a reminder of which corpus corresponds to *corpus 1* in the results. The third row contains the values of the parameters used in the analysis. The fourth line gives the names of the variables provided in the results.

```
# Corpus 1: File=E:\FLOBList.txt NbrText=500 NbrWord=1016849
# Corpus 2: File=E:\FROWNList.txt NbrText=500 NbrWord=1018653
# minfreq=12 minrange=0.05 maxpll=0.05 niter1=10000 maxpperm=0.0003 niter2=1000000
Word        FreqC1 FreqC2    Chi2     Chi2_p      LL     LL_p      Range WMW_p     FP_p
#            11961  12452    9.13  2.5097e-03    9.13 2.5088e-03   434
A            23113  23248    0.19  6.5906e-01    0.19 6.5906e-01   500
ABANDONED       20     38    5.55  1.8433e-02    5.65 1.7477e-02    32
ABILITY         98    128    3.93  4.7439e-02    3.94 4.7108e-02    72
AFTERWARDS      54      6   38.49  5.5140e-10   44.25 2.8842e-11    49 0.000001 0.000001
AMERICAN       219    660  220.57  6.8032e-50  230.93 0.0000e+00   205 0.000001 0.000001
AMERICANS       43    206  106.43  5.9437e-25  115.77 0.0000e+00    77 0.000001 0.000001
*BABY          109     78    5.19  2.2657e-02    5.22 2.2349e-02    32 0.076800 0.468400
```

Figure 1 : Output of the R script for the FLOB vs. FROWN comparison (partim)

The printed results are as follows:

- The analysed word.
- FreqC1 gives the frequency of the word in corpus 1 and FreqC2 its frequency in corpus 2. The relative frequencies can be calculated using the total frequencies of the two corpora given in the first row.
- Chi2 gives the Chi-square statistic and Chi2_p the corresponding p-value. LL and LL_p do the same for the LL test.
- Range gives the number of texts containing this word in the corpus in which the word is the most frequent (in terms of relative frequency).
- WMW_p gives the obtained p-value from the WMW test and FP_p does the same for the FP test.

As can be seen in the above extract of a comparative analysis of the FLOB and FROWN corpora, only those words which pass the *minfreq*,

the *minrange* and the *maxpll* thresholds are printed. The word *BABY* is preceded by an asterisk because the corpus that contains the most occurrences of this word is the one whose texts do not have the highest average rank. In this case, caution should be taken when interpreting the results, as explained in section 3. However, it is unlikely that this kind of result will be observed for words that are very significant for both the WMW and the FP tests.

## 6. Conclusion

This paper deals with statistical tests used in corpus linguistics for analysing lexical differences between corpora. Its most important contributions are the following:

- To explain in detail why the $Chi^2$ and LL tests are inadequate in this research field. It is important to emphasize again that the problem raised applies as much to the statistics resulting from the tests as it does to the probability that it is derived from them, and therefore also affects the keyness score or other any effect size measures such as that proposed by Gabrielatos and Marchi (2011). The problem raised affects any use of these tests to analyse corpora regardless of the linguistic unit counted: words, but also lexical bundles, collocations, syntactic constituents... It follows that, for instance, using these tests to analyse the use of the passive voice in different corpora is also inappropriate.
- To concretely demonstrate the seriousness of the erroneous conclusions reached when they are used.
- To propose two tests that are adequate and effective.
- To provide a concrete solution, which we hope is easy to put into practice, to use the appropriate tests.

An important question that has so far gone unanswered is which of the two appropriate tests is preferable. The main difference between them is that the WMW test, based as it is on ranks, is more sensitive than the FP test to small differences in frequency within a relatively large number of texts, whereas the FP test is more sensitive to the presence of a relatively small number of texts containing relatively high frequencies. Ideally, both tests should be significant. If only one of them

is clearly not significant for the chosen probability threshold, it is necessary to be very careful in interpreting the results and, in any case, to analyse the distribution of this word in the texts of the two corpora using *WST*.

## 7. Bibliography

Anthony, Laurence. 2012. *AntConc Version 3.3.5.* [Computer Software]. Tokyo: Waseda University. http://www.antlab.sci.waseda.ac.jp/.

Baker, Paul. 2004. Querying keywords: questions of difference, frequency and sense in keywords analysis. *Journal of English Linguistics* 32(4): 346-359.

Berry Kenneth J.; Mielke, Paul W. & Mielke, Howard W. 2002. The Fisher-Pitman permutation test: an attractive alternative to the F test. *Psychological Reports* 90: 495-502.

Bestgen, Yves. 2012. Analyse des différences lexicales entre des corpus : test ou distance du Khi-2? Dans *Actes de JADT 2012 : 11es Journées internationales d'Analyse statistique des Données Textuelles*, 150-161.

Bestgen, Yves. 2014. Inadequacy of the chi-squared test to examine vocabulary differences between corpora. *Literary and Linguistic Computing* 29: 164-170

Biber, Doug & Jones, James. 2009. Quantitative methods in corpus linguistics. In Ludeling, Anke & Kytö, Merja (ed.) *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter, 1286-1304.

Brezina, Vaclav & Meyerhoff, Miriam. 2014. A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics* 19(1): 1-28.

Church, Kenneth W. 2000. Empirical estimates of adaptation: The chance of two Noriegas is closer to p/2 than $p^2$. In *Proceedings of the 17th International Conference on Computational Linguistics*, 180-186.

Efron, Brad & Tibshirani, Rob. 1993. *An introduction to the bootstrap*. New York: Chapman & Hall.

Gablasova, Dana; Brezina, Vaclav & McEnery, Tony. 2017. Exploring learner language through corpora: Comparing and interpreting corpus frequency information. *Language Learning* (advance access). doi: 10.1111/lang.12226.

Gabrielatos, Costas & Marchi, Anna. 2011. Keyness: Matching metrics to definitions. Paper presented at the *Corpus Linguistics in the South: Theoretical-methodological challenges in corpus approaches to discourse studies - and some ways of addressing them.* Portsmouth: 5th November 2011.

Good, Phillip I. 2005. *Permutation, parametric and bootstrap tests of hypotheses* (Third Edition). New-York: Springer.

Gries, Stefan Th. 2005. Null hypothesis significance testing of word frequencies: a follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory* 1: 277-294.

Gries, Stefan Th. 2006. Exploring variability within and between corpora: some methodological considerations. *Corpora* 1(2): 109-151.

Hesterberg, Tim; Moore, David S.; Monaghan, Shaun; Clipson, Ashley & Epstein, Rachel. 2006. Bootstrap methods and permutation tests. Supplemental chapter for Moore, David S. & McCabe, George P. *Introduction to the Practice of Statistics*. New York: W H Freeman.

Hofland, Knut & Johansson, Stig. 1982. *Word frequencies in British and American English*. Bergen: The Norwegian Computing Centre for the Humanities.

Hofland, Knut; Lindebjerg, Anne & Thunestvedt, Jorn. 1999. *ICAME collection of English language corpora*. [CD-ROM]. Bergen: The HIT Centre, University of Bergen.

Hothorn, Torsten; Hornik, Kurt; van de Wiel, Mark A. & Zeileis, Achim. 2008. Implementing a class of permutation tests: the coin package. *Journal of Statistical Software* 28(8): 1-23.

Howell, David C. 2008. *Méthodes statistiques en sciences humaines*. Bruxelles: De Boeck Université.

Kilgarriff, Adam. 1996. Comparing word frequencies across corpora: Why Chi-square doesn't work, and an improved LOB-Brown comparison. In *Proceedings of ALLC-ACH Conference*, 169-172.

Kilgarriff, Adam. 2005. Language is never, ever, ever random. *Corpus Linguistics and Linguistic Theory* 1: 263-275.

Kurtz, Albert K. & Mayo, Samuel T. 1979. *Statistical methods in psychology and education*. New York: Springer.

Laviosa, Sara; Pagano, Adriana; Kemppanen, Hannu & Ji, Meng. 2017. *Textual and contextual analysis in empirical translation studies*. Singapore: Springer.

Lee, David Y. W. & Chen, Sylvia Xiao 2009. Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing* 18: 149-165.

Leedham, Maria. 2012. Review of: "New trends in corpora and language learning" and "Keyness in texts". *System* 40(1): 162-165.

Lewis, Don & Burke, C. J. 1949. The use and misuse of the chi-square test. *Psychological Bulletin* 46(6): 433-489.

Lijffijt, Jefrey; Nevalainen, Terttu; Säily, Tanja; Papapetrou, Panagiotis; Puolamäki, Kai & Mannila, Heikki. 2016. Significance testing of word

frequencies in corpora. *Literary and Linguistic Computing* 31(2): 374-397.

Lubbers Quesada, Margaret & Blackwell, Sarah E. 2009. The L2 acquisition of null and overt spanish subject pronouns: A pragmatic approach. In Collentine, Joseph (ed.) *Selected Proceedings of the 11th Hispanic Linguistics Symposium*. Somerville, MA: Cascadilla Proceedings Project, 117-130.

Marquilhas, Rita. 2015. Non-anachronism in the historical sociolinguistic study of Portuguese. *Journal of Historical Sociolinguistics* 1(2): 213-242.

Neuhauser, Markus & Manly, Bryan F. J. 2004. The Fisher-Pitman permutation test when testing for differences in mean and variance. *Psychological Reports* 94: 189-194.

Oakes, Michael & Farrow, Malcolm. 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, Magali & Bestgen, Yves. 2009. Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In Jucker, Andreas H.; Schreier, Daniel & Hundt, Marianne (ed.) *Corpora: Pragmatics and Discourse*. Amsterdam: Rodopi, 247-269.

R Core Team. 2013. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. http://www.R-project.org/.

Rayson, Paul; Leech, Geoffrey & Hodges, Mary. 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.

Rayson, Paul; Berridge, Damon & Francis, Brian. 2004. Extending the Cochran rule for the comparison of word frequencies between corpora. In *Proceedings of the 7th International Conference on Statistical analysis of textual data*, 926-936.

Rayson, Paul & Garside, Roger. 2000. Comparing corpora using frequency profiling. In Kilgariff, Adam & Sardinha, Tony B. (ed.) *Proceedings of the Comparing Corpora Workshop*, 1-6.

Sampson, Geoffrey. 2003. Statistical linguistics. In Frawley, William J. *International Encyclopedia of Linguistics* (2 ed.). New York: Oxford University Press.

Scott, Mike. 1997. PC analysis of key words - and key key words. *System* 25(2): 233-245.

Siyanova-Chanturia, Anna. 2015. Collocation in beginner learner writing: A longitudinal study. *System* 53: 148-160.

Tribble, Chris. 2000. Genres, keywords, teaching: towards a pedagogic account of the language of project proposals. In Burnard, Lou & McEnery, Tony (ed.) *Rethinking language pedagogy from a corpus perspective: papers from the third international conference on teaching and language corpora*. Bern: Peter Lang, 75-90.

Wallis, Sean. 2013. z-squared: The origin and application of Chi-square. *Journal of Quantitative Linguistics* 20(4): 350-378.

## Appendix: The R script for computing the statistical tests

```
CorpLexTests <- function(file1="no-file",file2="no-file",
        minfreq=0,minrange=0.05,maxpll=0.000001,niter1=10000,
        maxpperm=0,niter2=1000000) {
#parametres
if (maxpll<0 | maxpll>1) {cat(sprintf("\nParamater error : maxpll= %f not
      between 0 and 1\n",maxpll)); stop("Please change this parameter
      value")}
if (maxpperm<0 | maxpperm>1) {cat(sprintf("\nParamater error : maxpperm=
      %f not between 0 and 1\n",maxpperm)); stop("Please change this
      parameter value")}
if (minrange<0 | minrange>1) {cat(sprintf("\nParamater error : minrange=
      %f not between 0 and 1\n",minrange)); stop("Please change this
      parameter value")}
if (minfreq<0) {cat(sprintf("\nParamater error : minfreq= %d must be >=
      0\n",minfreq)); stop("Please change this parameter value")}
if (maxpperm>=1/niter1 & niter2<=niter1) {cat(sprintf("\nParamater error
      : niter2= %d must be > niter1= %d\n",niter1,niter2)); stop("Please
      change this parameter value")}
cat("Loading coin package\n")
if(!require(coin)){ #Try to install the Coin package if not already installed
  install.packages("coin")
}
library("coin")
cat("Reading first file\n")
d1=read.table(file1, header = FALSE,skip=1,comment.char=".",row.names =
      1,fileEncoding="UTF-16LE",sep = "\t",dec = ",", quote="\"")
cat("Reading second file\n")
d2=read.table(file2, header = FALSE,skip=1,comment.char=".",row.names =
      1,fileEncoding="UTF-16LE",sep = "\t",dec = ",", quote="\"")
#fnout = paste(dirname(file1),"Res.txt",sep="/") si on est sur de sep...
fnout = paste(substr(file1,1,nchar(file1)-
      nchar(basename(file1))),"CorpLexTestsRes.txt",sep="")
cat("Preparing data for processing\n")
#Delete some columns
d1=d1[,-(2:5)]
d2=d2[,-(2:5)]
#nbr of texts
ncol1 <- ncol(d1)-1 #the first is the word
ncol2 <- ncol(d2)-1
```

```
#merge by words, keeping all of them
da=merge(d1,d2,by.x=1,by.y=1,all=TRUE)
#transpose the data, but not the word
tda=t(da[,-1])
#remplace NA by 0
tda[is.na(tda)] <- 0
#Corpus id
corpus=c(rep(1, ncol1), rep(2, ncol2))
#number of mots in each text
rs=rowSums(tda)
#add these variables to the data
mydata=cbind(corpus,rs,tda)
lastword <- ncol(mydata) #last column number
cat("Start computing the statistical tests\n")
sink(fnout, append=FALSE, split=TRUE) #to print the output in a file
#print first lines
cat(sprintf("# Corpus 1: File=%s NbrText=%d NbrWord=%d\n# Corpus 2:
        File=%s NbrText=%d NbrWord=%d\n",
    file1,ncol1,sum(mydata[mydata[,'corpus'] ==
        1,2]),file2,ncol2,sum(mydata[mydata[,'corpus'] == 2,2])))
cat(sprintf("# minfreq=%d maxpll=%f niter1=%d  maxpperm=%f niter2=%d\
        n",minfreq,maxpll,niter1,maxpperm,niter2))
cat(sprintf("%-25s  %10s %10s  %10s %13s %10s %13s  %10s  %10s
        %10s\n","Word", "FreqC1","FreqC2","Chi2","Chi2_p","LL","LL_p"
        ,"Range","WMW_p","FP_p"))
#Loop on the words
  for (myidx in 3:lastword) {
   if (sum(mydata[,myidx])>=minfreq) {
    #compute the number of other words in the texts
    otherwords <- mydata[,2]-mydata[,myidx]
    #Compute Chi2 and LL
    ori <- as.table(rbind(tapply(mydata[,myidx], list(mydata[,'corpus']),
        FUN=sum), tapply(otherwords, list(mydata[,'corpus']), FUN=sum)))
    oriXsq <- chisq.test(ori,correct=FALSE)
    LL<-
        2*sum(oriXsq$observed[oriXsq$observed>0]*log(oriXsq$observed
        [oriXsq$observed>0]/oriXsq$expected[oriXsq$observed>0]))
    LL_pval=1-pchisq(LL,1)
    #Compute range in the corpus in which this word is the most frequent (in
        relative frequency)
```

```
if (oriXsq$observed[1]>=oriXsq$expected[1]) {
 whichcor=1
 range<-sum(mydata[mydata[,'corpus'] == 1,myidx] > 0)
 rm<-range/ncol1
}
else {
 whichcor=2
 range<-sum(mydata[mydata[,'corpus'] == 2,myidx] > 0)
 rm<-range/ncol2
}
if (rm>=minrange & LL_pval<=maxpll) { #No output if range is
      insufficient or the p-value for LL is to large
 #Relative frequency (by overwriting the original data)
 mydata[,myidx] <- mydata[,myidx]/mydata[,2]
 wt<-wilcox_test(mydata[,myidx] ~ factor(mydata[,'corpus']),distribution
      = "asymptotic")
 pwta<-pvalue(wilcox_test(mydata[,myidx] ~ factor(mydata[,'corpus']),dis
      tribution = approximate(B = niter1-1)))
 if ((pwta*(niter1-1)+1)/niter1<=maxpperm) {
  pwta<-(1+pwta+pvalue(wilcox_test(mydata[,myidx] ~ factor(mydata[
      ,'corpus']),distribution = approximate(B = niter2-niter1)))*(niter2-
      niter1))/niter2
 }
 else pwta<-(pwta*(niter1-1)+1)/niter1
 pfpa<-pvalue(oneway_test(mydata[,myidx] ~ factor(mydata[,'corpus']),
      distribution = approximate(B = niter1-1)))
 if ((pfpa*(niter1-1)+1)/niter1<=maxpperm) {
  pfpa<-(1+pfpa+pvalue(oneway_test(mydata[,myidx] ~ factor(mydata
      [,'corpus']),distribution = approximate(B = niter2-niter1)))*(niter2-
      niter1))/niter2
 }
 else pfpa<-(pfpa*(niter1-1)+1)/niter1
 if ((statistic(wt)>0 &  whichcor==2) | (statistic(wt)<0 &  whichcor==1))
      { #For discordances between ranks and frequencies
 cat(sprintf("*%-24s  %10.0f %10.0f  %10.2f %13.4e %10.2f %13.4e
      %10.0f  %10.8f %10.8f\n",
 da[myidx-
      2,1],oriXsq$observed[1],oriXsq$observed[3],oriXsq$statistic,oriXsq
      $p.value,LL,LL_pval,range,pwta,pfpa))
 }
 else { #For the normal case
```

```
      cat(sprintf("%-25s  %10.0f %10.0f  %10.2f %13.4e %10.2f %13.4e
          %10.0f  %10.8f %10.8f\n",
      da[myidx-2,1],oriXsq$observed[1],oriXsq$observed[3],oriXsq$statistic,o
          riXsq$p.value,LL,LL_pval,range,pwta,pfpa))
     }
    }
    }#End of the range and the maxpll conditions
  } #Loop end
  sink() #End of output in a file
}
```