Corpora, 13(2), 2018, 205-228.
https://www.euppublishing.com/doi/abs/10.3366/cor.2018.0144

Evaluating the frequency threshold for selecting lexical bundles by means of an extension of the Fisher's exact test

Yves Bestgen,
Centre for English Corpus Linguistics, Université catholique de Louvain
Place du Cardinal Mercier, 10 1348-Louvain-la-Neuve Belgium

Abstract
This methodological study uses an extension of the Fisher's exact test to sequences longer than two words in a bid to evaluate whether the frequency thresholds conventionally used for identifying lexical bundles in corpora are high enough to ensure that the selected sequences are unlikely to result from chance. Sequences of four, three and two words were analysed in corpora, the sizes of which ranged from 50 000 to 4 200 000 words. Results suggested that the usual frequency cut-offs were appropriate for four-word sequences and that, as expected, it was questionable to extract two-word sequences on the sole basis of a frequency criterion. For three-word sequences, greater cut-offs than 40 times per million words should be favoured for corpora with a size of 250 000 words or less. This study also highlights the effect of corpus size on the efficiency of the frequency thresholds when they are expressed in normalized frequency.

Keywords: lexical bundles; n-grams; corpus-driven approach; normalized frequency cut-off; Fisher's exact test; corpus size

> ‘How many examples of a three-, four-, or five-word sequence are necessary for it to be considered a phrase? As this is not an answerable question [...]’ (Hunston, 2002: 147)[1]

1. Introduction

One of the frequently used approaches for studying formulaic language is based on the automatic identification of recurrent continuous sequences of words in a corpus (Cortes, 2015). These sequences are often called ‘lexical bundles’ (Biber et al., 1999: Chapter 13), but are also referred to as ‘recurrent word combinations’ (Altenberg, 1998), ‘chains’ (Stubbs, 2002) or ‘chunks’ (O'Keeffe et al., 2007). Studying these sequences has highlighted phraseological differences between registers, genres, academic disciplines and geographical dialects, among others (e.g., Aijmer, 2009; Biber, 2009; Biber et al., 1999; Durrant, 2017; Cortes, 2004, 2008; Hyland, 2008; Scott & Tribble, 2006). Additionally, they have also been used to distinguish texts by novice and expert writers, as well as by native speakers and learners of foreign languages (e.g., Ädel & Erman, 2012; Cortes, 2004; Chen & Baker, 2010, 2016; De Cock, 1998; Groom, 2009; Salazar, 2014; Vidakovic & Barker, 2010). The majority of studies have focused on four-word sequences, considered not too frequent for a qualitative analysis but frequent enough for good diversity (Chen & Baker, 2010); however, shorter sequences composed of three words or even two were also analysed (Altenberg, 1998; Carter, 2006; Crossley & Salisbury, 2011; De Cock, 1998; Groom, 2009; O'Keeffe et al., 2007).

Two criteria are used to identify lexical bundles in a corpus: a frequency threshold, which is supposed to guarantee that the bundles show a statistical tendency to co-occur and the number of documents in which a sequence occurs, which is used to eliminate bundles specific to a few speakers or writers (Biber et al., 1999: 989-993). Regarding the latter criterion, broad consensus has been reached for setting the threshold between three and five texts, although some studies prefer to use a threshold expressed as a percentage of the total number of texts included in the corpus (Hyland, 2008).

For the first criterion, however, large variations of the cut-off have been observed. While this is usually set between ten and forty occurrences per million words[2], cut-offs as high as eighty-eight

---

[1] The context of this excerpt is as follows: 'In the "furious scribbling" example above, the phrase *After a few moments of* is a candidate 'phrase' in English. The Bank of English has 642 instances of *after a moment*, 99 instances of *after a few moments* and 12 instances of *after a few moments of*. How many examples... '

[2] The usual convention for normalizing raw frequencies to numbers of occurrence per million words is used here even if, when the corpus is much smaller than this number, it requires a potentially problematic extrapolation (Gray, 2016).

occurrences per million words (De Cock, 1998) and as low as four occurrences per million words (O'Keeffe et al., 2007) have also been used. This variability can be linked to Hunston's question highlighted above and quoted by Gries (2008: 423) to emphasize that 'it seems as if there is as yet no rigorous operationalization of when something is frequent enough to be considered a unit in the above sense of the term.' The arbitrariness of the threshold is a consequence of the way in which co-occurrence frequency is conceptualized in these studies, i.e., as a measure of effect size and, more precisely, as a measure of the strength of association between the words that compose a bundle (Biber, 2009: 291; Evert, 2009: 1224). The larger the frequency, the stronger the association and the more interesting the bundle. The threshold is set for practical reason at a level that allows for selecting enough lexical bundles for sufficient diversity, but not too much for allowing a close look at their properties.

Taking co-occurrence frequency to be a reliable measure of effect size poses a difficulty. The frequency of a sequence, though itself an important feature, is not sufficient for identifying bundles of words that show a statistical tendency to co-occur where this is what frequency is used for according to the core of the definition of lexical bundles. In their seminal work, Biber et al. (1999: 988-989; see also Biber and Conrad (1999: 183) and Hyland (2008: 5)) defined lexical bundles as 'extended collocations: bundles of words that show a statistical tendency to co-occur' and collocations as 'associations between lexical words, so that the words co-occur more often than expected by chance' (1999: 988). Frequency is how this property is operationalized (Biber et al. 1999: 992). Many scholars warn that a sequence can achieve an high frequency because it is made up of very frequent words (see, e.g., Evert, 2005: 20; Gries, 2010: 275, 2015: 94; Hunston, 2002: 70; McCarthy & Carter, 2006: 17; Stubbs, 2002: 235-236). For example, Biber and Jones (2009: 1296) stresses that 'Simple frequency information can present a biased measure of the strength of a collocation, because very frequent words are likely to occur together simply by random chance'. This problem can easily be illustrated in the case of two-word sequences, or bigrams[3]. Table 1 provides counts for the two bigrams '*so many'* and '*it that'* in the conversation section of the British National Corpus[4] (BNC-CONV: roughly 4 200 000 words; see section 3) in the form of 2x2 contingency tables. The rows in these tables represent the presence or absence of the first word of the bigram and columns those of the second word. Cell *a* indicates the number of times the bigram

[3] All lexical bundles are n-grams, which are uninterrupted sequences of words, but only n-grams that fulfil the selection criteria are lexical bundles (Cortes 2015: 200). In this paper, *n-grams* (bigrams, trigrams, etc.) and *word sequences* are used to designate all contiguous sequences of words while *lexical bundles* are used to describe those that meet the selection criteria.

[4] http://www.natcorp.ox.ac.uk/corpus/

was observed in the corpus; cell *b* indicates the number of times a bigram that begins with the first word but does not end with the second was observed and so on.

Table 1. Frequency counts for two bigrams in the BNC-CONV

	Second word				Second word			
First word	many		¬many	Total	First word	that	¬that	Total
so	***a*** 348	***b***	24 054	24 402	it	348	127 656	128 004
¬so	***c*** 1573	***d***	4 206 284	4 207 857	¬it	84 334	4 019 921	4 104 255
Total	1921		4 230 338	4 232 259	Total	84 682	4 147 577	4 232 259

Both bigrams were observed the same number of times in the corpus, although '*so many'* is more clearly phraseological than '*it that'*. Compared to the frequency of *so* and especially of *many*, 348 is a large number; compared to those of *it* and *that*, 348 is small. Considering the simple frequency of co-occurrence as a measure of effect size and selecting the sequences that exceed a threshold leads to neglecting the frequency of the words that compose these sequences and thus to treating '*so many'* and '*it that'* in the same way even though the high frequency of '*it that*' may simply be due to chance. As emphasized by Ellis et al. (2015), there is no reason to think that the same phenomenon will not be observed with sequences of more than two words.

One may therefore wonder whether the usual frequency thresholds are high enough to ensure that the selected sequences are unlikely to result from chance. Attempting to bring a new perspective on this issue is the primary objective of the present study. One way to answer this question is to estimate the probability that chance alone has to produce at least as many instances of the sequences as the number actually observed in the corpus. In the case of sequences of two words, this issue has received significant attention and several inferential tests have been proposed, including the Chi-square test, the log-likelihood test, a variant of the Student t test or the Fisher's exact test (Evert, 2005; Pecina, 2010). While almost all these tests have been criticized for their inadequacy in the context of studying bigrams and similar units (Pedersen et al., 1996; Evert, 2009; Stubbs, 1995), the Fisher's test, already proposed by Jones and Sinclair in 1974, has received much support and is used as a benchmark for assessing other association indices (Evert, 2009; Moore, 2004; Pedersen, 1996; Stefanowitsch & Gries, 2003). Recently, Bestgen (2014) proposed an extension of this test for sequences longer than two words. The present study uses this extension in an attempt to evaluate whether the frequency thresholds conventionally used for identifying bundles are not likely to select sequences that chance alone may have produced as many times as observed in the corpus. In these analyses, special attention is given to the size of the corpus from which the lexical bundles are extracted as it greatly differs from one study to another, ranging from 40,000 to more than five

million words (Chen & Baker, 2010). This will allow us to evaluate whether, according to the statistical approach used here, a given standardized threshold is as effective in a small as in a large corpus, a question on which opinions diverge (Biber & Barbieri, 2007; Cortes, 2015; Hyland, 2012).

It is important to note that this study does not aim to challenge the classical operationalization of lexical bundles as the most frequently recurring sequences of words. It is obviously more useful to distinguish registers or texts produced by native and non-native speakers by means of very frequent sequences of words than by means of infrequent sequences. However, a (relatively) high lexical bundle frequency can be misleading as it can result from the high frequency of the words that compose it. The objective of the study is to propose a technique by which to evaluate whether the thresholds conventionally used to decide that an n-gram is a lexical bundle are high enough to avoid such problems.

To complement the frequency criterion, several studies have recently proposed the use of mutual information (MI), a well-established association index for bigrams that can be extended to longer sequences (Groom, 2009; Salazar, 2014; Simpson-Vlach & Ellis, 2010). In the present study, however, this index is inappropriate as it is a measure of effect size rather than an inferential test (Evert, 2009). Moreover, Biber (2009) analysed in depth the use of MI for identifying lexical bundles and concluded that this approach has the major disadvantage of penalizing word sequences comprising frequent words while lexical bundles usually incorporate very frequently occurring function words. It should be noted that this property is typical of MI and that other association indices, such as the t-score, the log-likelihood measure or Fisher's exact test, do not disfavour sequences composed of frequent words (Evert, 2009; Stefanowitsch & Gries, 2003).

The following section describes the extension of the Fisher's test to sequences longer than two words. Then, two studies that use this test to evaluate the frequency thresholds generally employed for selecting lexical bundles are described: the first focuses on a series of corpora similar to those already used in this type of research; the second is based on subcorpora of variable size, extracted from the same corpus of reference. The conclusion summarizes the recommendations that can be drawn from this study and points out some of its primary limitations.

2. The Fisher's exact test for bigrams and its extension to longer sequences

Fisher's exact test can be used to analyse contingency tables like those presented in Table 1, by calculating the probability that chance alone produces at least as many instances of the bigram (Jones & Sinclair, 1974; Pedersen et al., 1996). To compute this probability, Fisher's proposition is to examine all the contingency tables that can be constructed in accordance with the marginal totals and determine the proportion of those which are at least as extreme as the observed table. A table at

least as extreme is a table in which the bigram frequency is at least as high as is observed in the corpus (cell *a* in Table 1). The formula is as follows (Evert, 2005: 80):

$$p = \sum_{i=a}^{\min(a+b,a+c)} \frac{\binom{(a+c)}{i}\binom{(b+d)}{(a+b-i)}}{\binom{(a+b+c+d)}{(a+b)}}$$

the letters *a*, *b*, *c* and *d* correspond to the counts in the cells as shown in Table 1 and *i* represents the possible values for the cell *a*, which produce a table at least as extreme as the original one.

For the '*so many'* bigram in Table 1, Fisher's test returned a probability of less than 1^{-320}. In other words, this bigram had virtually no chance of occurring at least 348 times in the corpus if the words were ordered at random. The reverse was true for the '*it that'* bigram: chance alone would produce much more than 348 occurrences of it.

If the Fisher's test has become the benchmark for testing bigram frequencies, its use for analysing longer sequences is problematic, because these sequences require the construction of contingency tables of three or more dimensions. Exact tests (or approximations) proposed for such tables (Agresti, 1992; Zelterman et al., 1995) are not suited to the study of word sequences, because of the sample size and the large number of tests that must be carried out and because a very specific and directional hypothesis must be tested: the probability of having at least as many instances of a sequence by chance alone.

Bestgen (2014) proposed a generalization of the Fisher's exact test for the analysis of three-words and longer sequences by means of a Monte Carlo estimation procedure. The starting point of this generalization is an alternative method for calculating the probability of the Fisher's exact test. Typically, this probability is obtained by using the formula given above, but another approach is also possible: using a Monte Carlo permutation procedure to generate a random sample of the possible contingency tables, given the marginal totals and determining the proportion of these tables that give rise to a value at least as extreme as that observed (Agresti, 1992). This is indeed the procedure proposed by Pedersen et al. (1996) in their article recommending the use of the Fisher's test for the study of bigrams.

A corpus being a long sequence of graphic forms, its permutation is simply randomly mixing all these forms and counting the number of times a given bigram is observed within this permutation. Thus, any permutation of the order of the tokens in the corpus generates a random contingency table for each bigram. The permutations mix all the words in the corpus without taking into account the text boundaries because otherwise the resulting probabilities would no longer correspond to those produced by Fisher's exact test. This estimation procedure can easily be generalized to sequences of

more than two words by counting in each random permutation not only the bigrams, but also the trigrams, quadrigrams, etc.

In summary, the purpose of this inferential test is to estimate the probability that chance alone has of generating at least as many instances of an n-gram as the actual number observed. To do this, a Monte Carlo test was used. It was composed of two steps that were repeated many times:

- Randomly swap all the tokens in the corpus,
- For each n-gram present in the original corpus, determine if its frequency in the permutation is at least equal to its frequency in the original corpus. If this is the case, add 1 to the counter corresponding to this n-gram.

When the desired number of iterations is reached, the counter value for each n-gram is divided by the number of iterations performed. The resulting number is the estimation of probability.

Bestgen (2014) showed that this procedure permits an almost perfect estimation of Fisher's exact probability for all the 2-grams in a corpus, with the caveat that the number of permutations carried out limits the precision of the probability. The procedures major weakness is thus its computational cost. Therefore, it does not lead to a viable criterion by which to identify the lexical bundles in a corpus, but it might help in the assessment of the frequency criterion.

The following two sections are based on this extension of the Fisher's test for evaluating the frequency cut-offs used for selecting lexical bundles in corpus linguistics. The analyses focus on sequences composed of two to four words. The four-word sequences are undoubtedly the most often studied. The three-word sequences have also been the subject of numerous studies and Simpson-Vlach and Ellis (2010: 509) found that many important recurrent word combinations are indeed three-word bundles. The inclusion of two-word sequences may seem surprising (e.g., 'The purposes of the lexical bundle approach require that multi-word sequences be identified with priority given to frequency, fixedness, and sequences longer than two words' – Conrad & Biber, 2004: 58). Nonetheless, two-word sequences have been analysed in several studies (e.g., Crossley & Salisbury, 2011; Groom, 2009; O'Keeffe et al., 2007). Moreover, it is the inclusion of these two-word sequences in his seminal analyses that is responsible for Altenberg's often-quoted assertion that 'A rough estimation indicates that over 80 per cent of the words in the corpus form part of a recurrent word-combination in one way or another' (Altenberg, 1998: 102; see Wray (2002: 28) for an in-depth discussion of this type of estimation).

3. Study 1

3.1 Material and methods

3.1.1 Corpora

To ensure sufficient generality to the conclusion of this study, four corpora, similar to those used in lexical bundle research, were selected by varying the size, mode and native characteristics of the authors of the texts. The details of these corpora are provided in Table 2.

The largest corpus (BNC-CONV) contains approximately 4 200 000 words, a number near the upper limit of the corpus size used in such studies. It includes all the documents (spontaneous conversations) of at least 1000 words of the Demographic Spoken section of the British National Corpus. It is similar in size and content to the corpora used by Biber et al. (1999) for instance.

The next two corpora by size are composed of roughly 150 000 words, a fairly common size in studies that aim at describing the formulaic differences between native and non-native writers (e.g., Ädel & Erman, 2012; Chen & Baker, 2010; Juknevičienė, 2009). The first corpus (ICLE-223) is a subpart of the International Corpus of Learner English (Granger et al., 2009). It is composed of 223 argumentative texts written by learners of English from three different native languages: French, German and Spanish (Thewissen, 2013). The second corpus (LOCNESS-US) is the American component of the Louvain Corpus of Native English Essays, a corpus of argumentative essays written by American university students.

The final corpus (LONGDALE-FR1) consists of only 50 000 words, a value close to the size of the smallest corpora used in this type of research (Biber & Barbieri, 2007; Chen & Baker, 2016; De Cock, 1998; Lin, 2013). It was extracted from the Longitudinal Database of Learner English and contains texts written by French-speaking students during their first year in an English language and literature curriculum.

Table 2. Details of the four corpora analysed

Corpus	Number of texts	Number of words
BNC-CONV	149	4 232 259
ICLE-223	223	153 481
LOCNESS-US	175	151 362
LONGDALE-FR1	89	52 965

3.1.2 Corpora processing

All corpora were tokenized using CLAWS[5]. All orthographic forms (words, but also numbers, symbols and punctuation) detected by CLAWS were considered as tokens to be randomly mixed. All these tokens were lowercased. The C code used in Bestgen (2014) was run on several Intel Xeon E5-2650v2 workstations. Four million permutations were performed on the BNC-CONV corpus and 20 million on the three other corpora, which were much smaller. To obtain the sequences and their frequency in the original corpus and in each permutation, only strings of

[5] http://ucrel.lancs.ac.uk/claws/

uninterrupted word-forms were taken into account. Thus, any punctuation mark or sequence of characters that did not correspond to a word interrupted the sequence extraction.

3.3 Analyses and Results

The main research question this study attempts to answer is whether the thresholds conventionally employed to extract bundles from a corpus allows for selecting only sequences whose frequencies are unlikely to result from chance. The rejection level used to decide whether chance was responsible for the high frequency of a sequence was set to the classic value of 0.05. Any probability equal to or smaller indicated a sequence considered as unproblematic. To take into account the large number of tests, which increases the probability of wrongly deciding that at least one test is statistically significant, Holm's sequential procedure was used (Holm, 1979; Gries, 2005). This procedure ensures a family-wise error rate (i.e., the probability of wrongly rejecting the null hypothesis in at least one test) of 0.05, regardless of the number of sequences extracted from each corpus.

As the results of this study are presented graphically, the first part of this section explains how these graphics should be read using the four-word sequences in the largest of the four corpora as an illustration. Next, the results of the 4-grams, 3-grams and 2-grams in the four corpora are presented. The last part reports a complementary analysis conducted to evaluate the potential impact on the results of the second criterion used to select the lexical bundles: the minimal number of different texts in which a lexical bundle must occurs.

3.3.1 Graphical representation of the results

To answer the research question, finding the minimum frequency threshold that ensures that all selected sequences successfully pass the inferential test appears sufficient. However, given the large variability of the frequency cut-offs used in the literature (from four to more than eighty occurrences per million words), it is interesting to determine the percentage of sequences that successfully passes this test for many different values of the frequency criterion. These two ways of presenting the permutation results are given in the following figures. To plot them, the percentage of sequences selected by the frequency cut-off that were statistically significant was determined for all possible values of the cut-off from the highest (the frequency of the most frequent sequence in the corpus) to the lowest (the frequency per million words corresponding to a single occurrence of the sequence).

To render this process more concrete, we may consider the results for the four-word sequences in the BNC-CONV corpus, as illustrated in Figure 1. The most frequent sequence is '*i do n't know*' with a raw frequency of 4653, corresponding to a frequency of 1099 per million words. According to the Monte Carlo test, the probability that such a high frequency will occur by chance alone is

much lower than 0.05. The next most frequent sequence, '*i do n't think*', occurred 1936 times in the corpus, another highly unlikely frequency. This is the same for the next 2066 sequences up to the raw frequency of twenty-two, corresponding to a normalized frequency of 5.19. If in this corpus the cut-off is set to this value (or to a higher one), 100% of the selected four-grams will be too frequent to result by chance, according to the test used. In Figure 1, this value corresponds to the rightmost point.

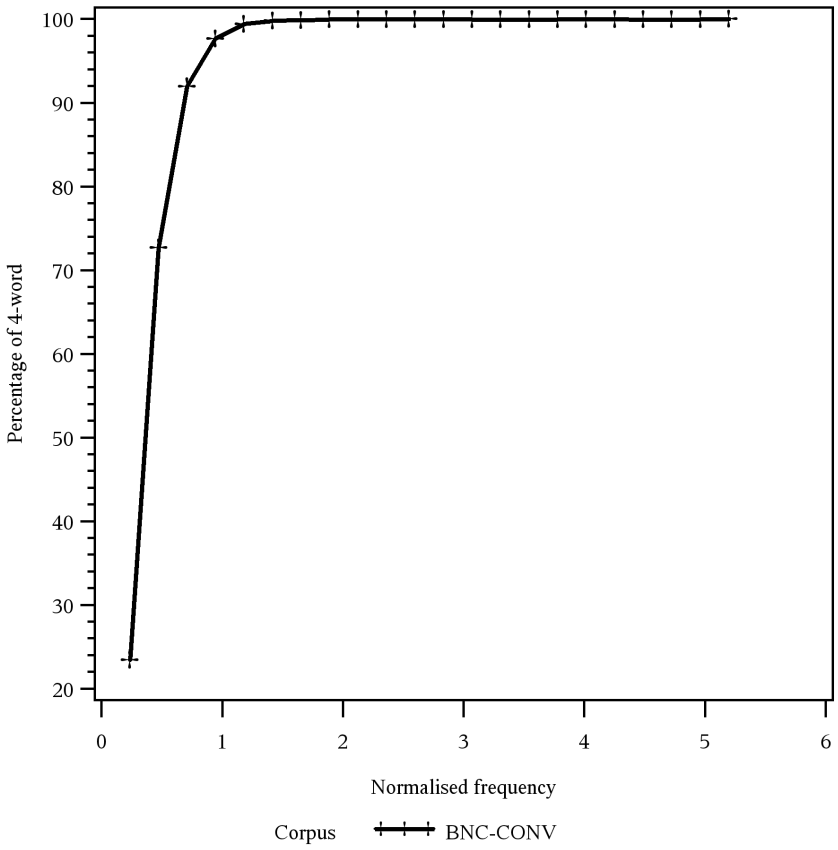

Figure 1. Percentage of four-word sequences that pass the inferential test in the BNC-CONV corpus

The next threshold is twenty-one or 4.96 occurrences per million words; 148 sequences appeared this many times in the corpus and one, '*i i i i*', was not significant for the test. Statistically speaking, this does not mean that this sequence resulted from chance, but rather that given the frequency of '*i*' in a corpus of this size, chance alone could have produced at least twenty-one occurrences of this sequence. Linguistically speaking, '*i i i i*' is a marker of hesitation, regardless of the number of times '*i*' is repeated and therefore, it cannot be considered specifically as a four-word bundle. If the frequency cut-off is set at this value, 2215 four-word sequences out of 2216 (or 99.955%) will be accepted by the statistical test. This value corresponds to the second rightmost point in Figure 1.

All of the following sequences by descending frequency are statistically too frequent for the test up to the raw frequency of nine, or 2.13 times per million words, where '*to you and i*' is tagged as

non-significant. At this threshold, 8246 out of 8248 sequences were validated by the test. Following on, the test rejected three more sequences at the eight cut-off, seven more at the seven cut-off and 18 more at six. Even at the six cut-off, 99.79% of the selected sequences are validated by the inferential test. This percentage does not really begin to decrease until the raw frequency cut-off of four, shown in Figure 1 by the fourth point from the left.

These results therefore indicate that, for the four-word sequences in this corpus, frequency cut-offs ranging from ten to forty occurrences per million words guarantee the selection of bundles for which the observed frequency should not be the result of chance only. The benefit of this graphical representation is that it indicates the lowest frequency per million words for which 100% of the sequences are statistically significant for the test, as well as the percentage of problematic sequences that are selected if the frequency threshold is set below this value. In addition, as recommended by Chen and Baker (2010), it simultaneously provides the normalized frequency (per million words, using the x-axis) and the raw cut-off frequency (by counting the points on the graph from left to right, the first point corresponding to a raw frequency of 1).

3.3.2 Comparison of the results for the four corpora

Figure 2 shows the percentage of four-word sequences in the four corpora that pass the inferential test for all possible values of the frequency criterion. As already noted above (see Figure 1 for a more detailed representation), using a normalized frequency cut-off (per million words) as low as six in the BNC-CONV corpus of 4 200 000 words allows for selecting only sequences that are too frequent to result by chance, according to the test used. Results for the two corpora of 150 000 words were almost identical to one another and showed that a raw frequency threshold of four selected only non-problematic sequences. For the corpus of 50 000 words, a raw frequency threshold of three had the same effect. Note, however, that the smaller the corpus, the higher the threshold must be when it is expressed in normalized values to only select statistically significant sequences. As explained in the conclusion, this observation makes sense from a statistical point of view.

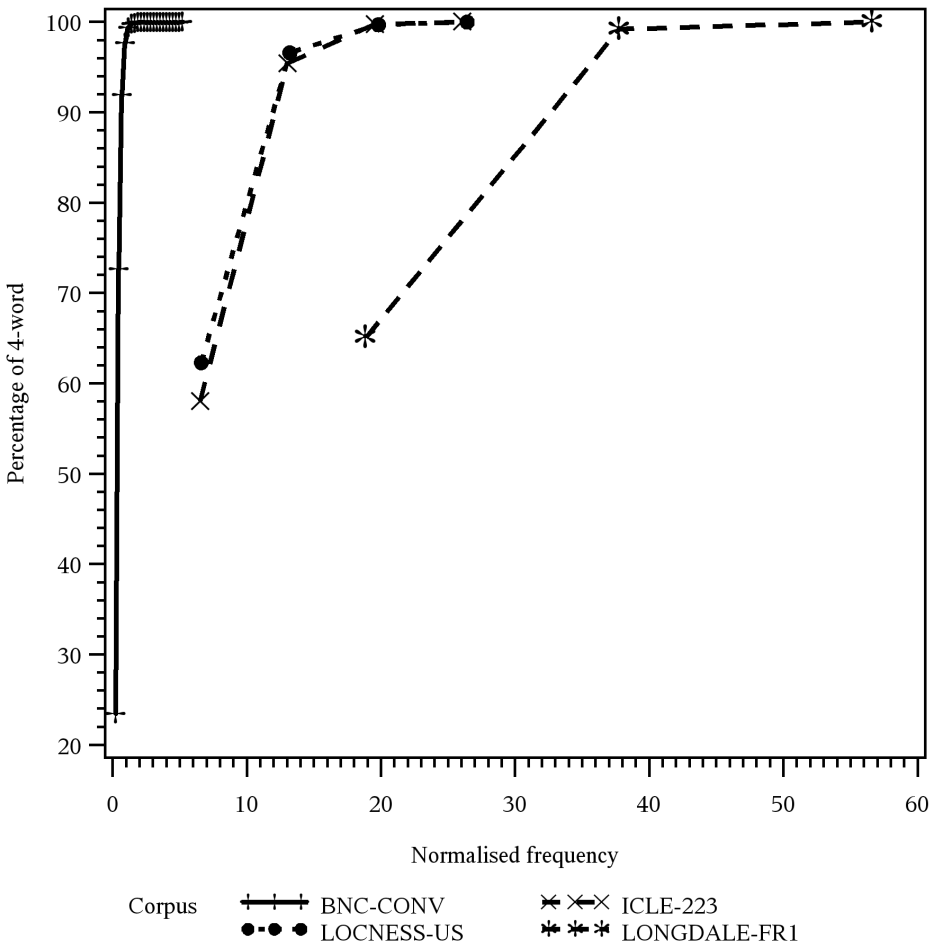

Figure 2. Percentage of four-word sequences that pass the inferential test

Figure 3 shows these percentages for the three-word sequences. As may be expected, much higher thresholds were needed to ensure that all selected bundles were genuine, according to the inferential test used: a normalized cut-off of about twenty-five for a 4 200 000-word corpus, of eighty for the 150 000-word corpora and of 140 for the 50 000-word corpus. In raw frequency, the threshold was twelve for the 150 000-word corpora and seven for the 50 000-word corpus. For the latter corpus, the percentage of non-problematic sequences decreased strongly when the criterion was set at a lower level, while in the LOCNESS-US corpus, a normalized cut-off around forty results in 98.3% of the identified sequences being classified as non-problematic.

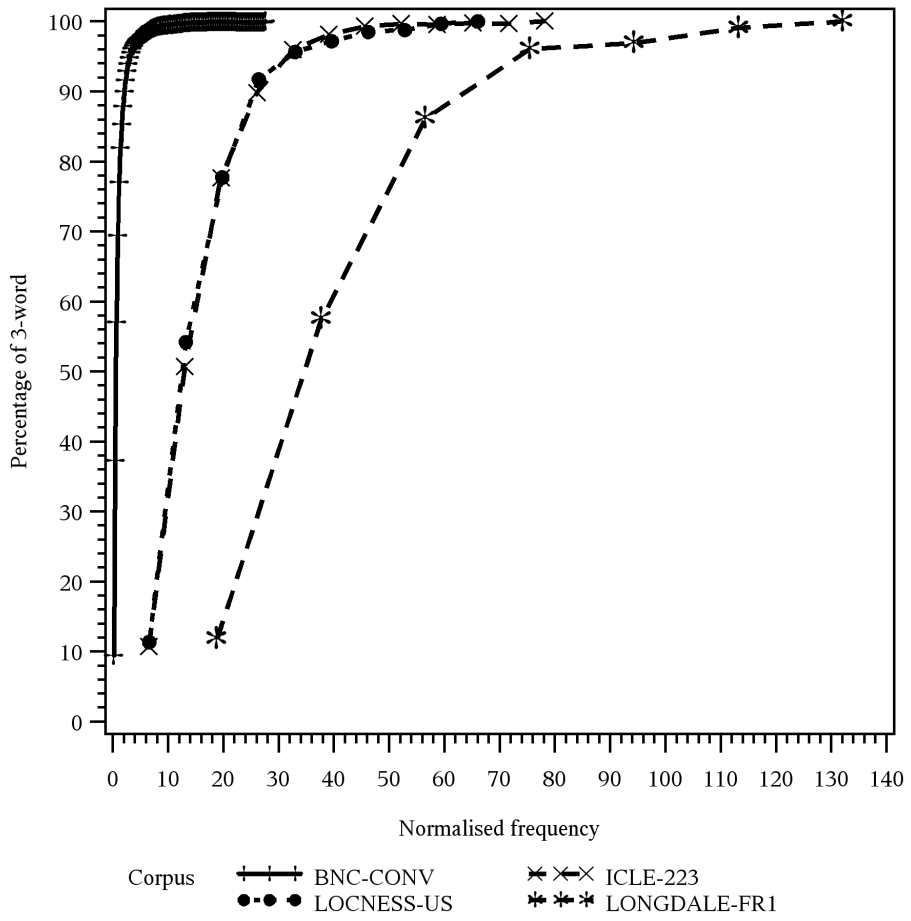

Figure 3. Percentage of three-word sequences that pass the inferential test

The results for the two-word sequences, shown in Figure 4, are clear-cut. For all corpus sizes, a normalized frequency above 400 was necessary to ensure that the selected sequences are unlikely to result from chance. The sharp drop in the percentage from the second rightmost point curves for both 150 000-words corpora is explained by the fact that an extremely common bigram, '*and the*', did not pass the inferential test. It was the tenth most common bigram in LOCNESS-US and the eleventh most common in ICLE-223. At that point the percentage of non-problematic sequences falls from 100% to 91% (or 90%) since one bigram amongst the eleven (or ten) most frequent bigrams is rejected by the inferential test.

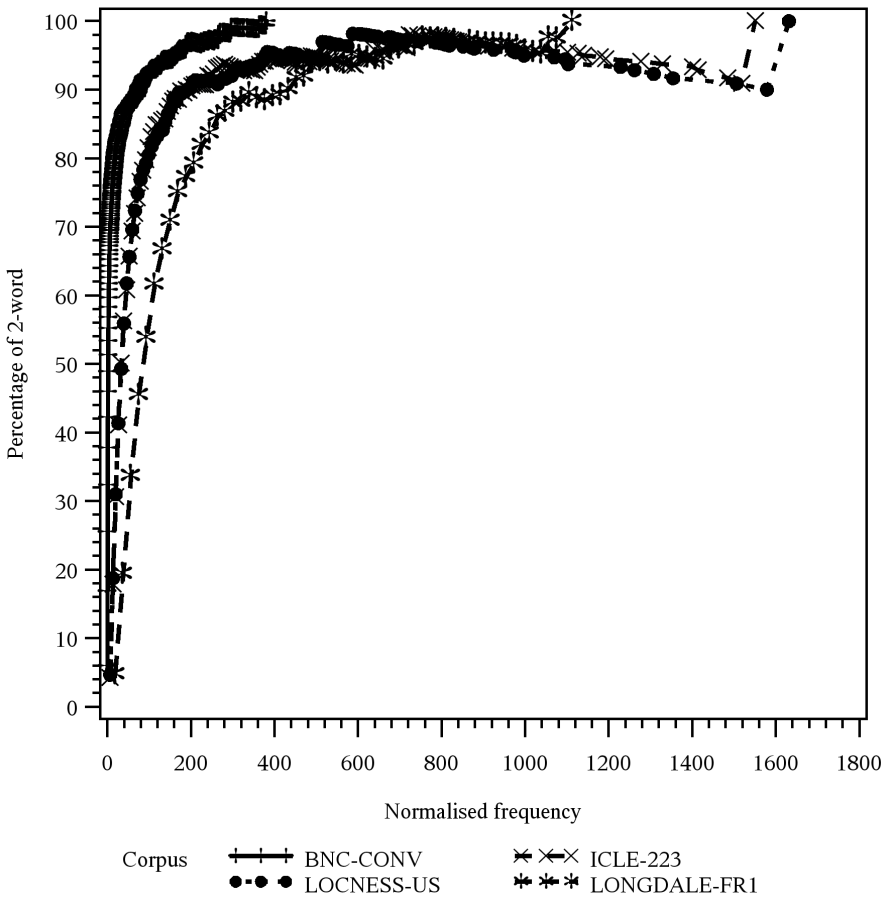

Figure 4. Percentage of two-word sequences that pass the inferential test

3.4. Discussion

The analyses suggest that the usual frequency cut-offs are appropriate for four-word sequences and at least questionable for two-word sequences. For three-word sequences, larger cut-offs than 40 times per million words should be favoured for corpora with a size of 150 000 words or less. There were no differences between the corpora of texts written by native or non-native speakers. The size of the corpora appeared to have a fairly strong effect: the smaller the corpus, the higher the normalized threshold must be. However, it is not easy to gain an accurate idea of this issue, because the corpora analysed differed in more ways than size only and because there is no corpus of intermediate size between 150 000 and 4 200 000 words. The next study attempts to provide a better idea regarding the impact of corpus size by comparing more corpora of different sizes that have been extracted from the same reference corpus.

Before reporting this study, however, it is necessary to consider whether the second criterion used for selecting lexical bundles, the number of texts in which they appear, modifies the conclusions reported above. This criterion has the effect of excluding word sequences that pass the frequency cut-off, but that are not distributed in a large enough range of texts. The analysis carried

out by setting this criterion to three or five, the commonly used thresholds, showed that this criterion did not make a difference for the conclusions in any of the four corpora. As an example, Figure 5 compares the results for the three-word sequences in the LOCNESS-US corpus when this criterion was not used and when it was set to three. The two curves were almost identical; the only difference was that the curve resulting from taking into account the criterion could only start when the sequences reached at least a frequency equal to this value. It is, however, very important to note that this analysis does not indicate this second criterion as irrelevant, but rather that its use does not alter the findings presented above regarding the primary selection criterion, i.e., the sequence frequency in the corpus.

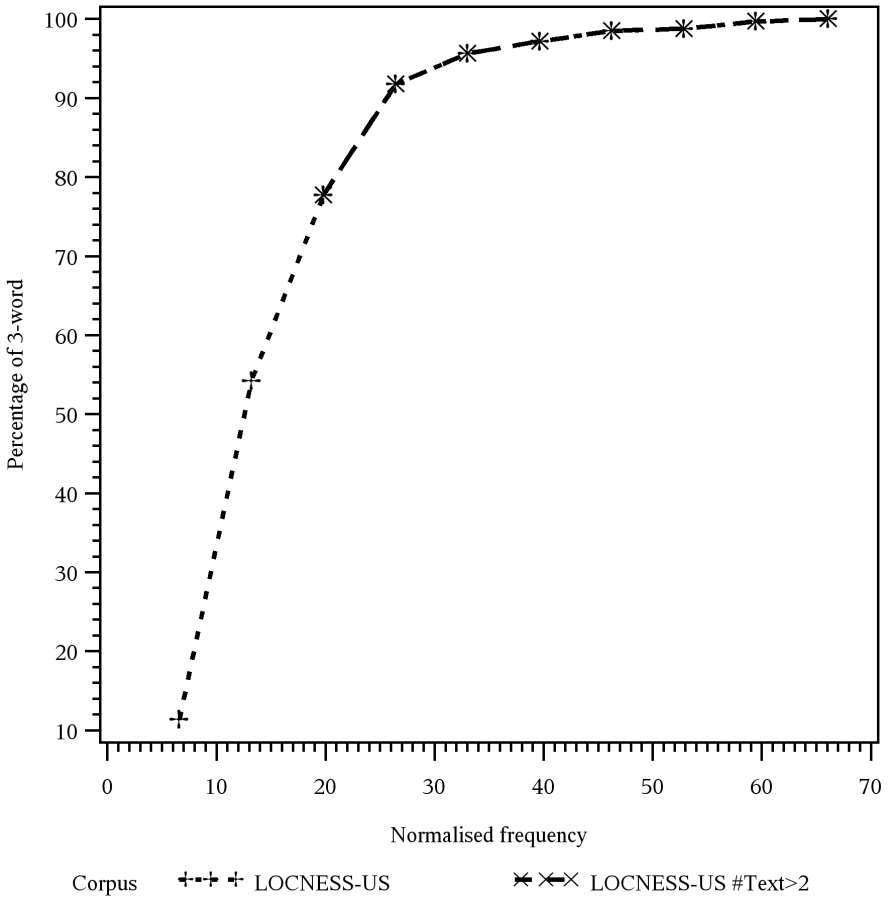

Figure 5. Percentage of three-word sequences that pass the inferential test in the LOCNESS-US corpus when the minimum number of texts in which they appear is set to one or to three

4. Study 2

4.1 Material and methods

In order to more accurately analyse the impact of the size of a corpus on the effectiveness of the frequency thresholds for selecting lexical bundles, six corpora were successively extracted from the

BNC-CONV corpus (referred to here as BNCC-4M) in such a way that their size was in each case approximately equal to half the size of the corpus just larger: BNCC-2M contains two million words, BNCC-1M contains 1 million words and so forth, up to BNCC-62m, which contains 62 000 words. These subcorpora were extracted by selecting the shortest texts in the BNCC-4M corpus, so that they contained the largest possible number of texts. The details of these corpora are provided in Table 3.

Table 3. Details of the seven corpora analysed

Corpus	Number of texts	Number of words
BNCC-4M	149	4 232 259
BNCC-2M	119	2 017 298
BNCC-1M	90	999 750
BNCC-500m	66	503 801
BNCC-250m	46	247 704
BNCC-125m	31	126 672
BNCC-62m	20	61 741

The procedure for processing these corpora is identical to that used in Study 1. Four million permutations were performed on each of the six new corpora, while the four million permutations performed for Study 1 were used for the BNCC-4M corpus.

4.2 Results

Figures 6 and 7 show the results for the sequences containing four and three words. The results for the two-word sequences are not presented, because the first study showed that it was at least questionable to extract sequences of this length on the sole basis of a frequency criterion. Both figures confirm the effect of the corpus size obtained in Study 1: the smaller the corpus, the higher the threshold expressed in normalized frequency should be to ensure that the selected sequences are statistically significant.

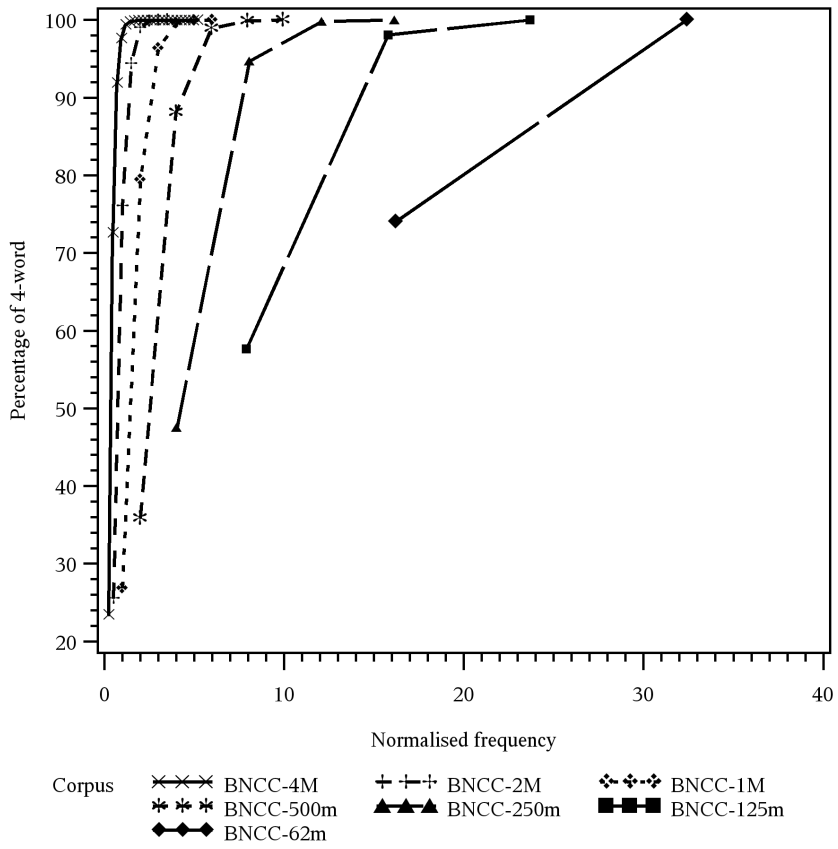

Figure 6. Percentage of four-word sequences that pass the inferential test

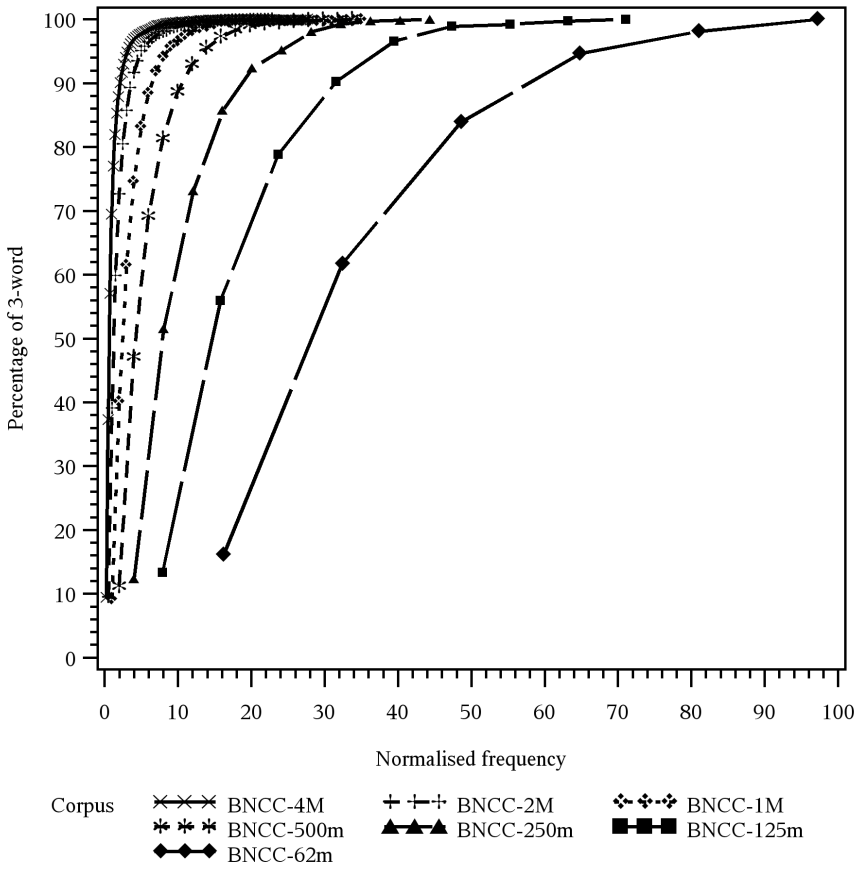

Figure 7. Percentage of three-word sequences that pass the inferential test

It is interesting to look at how the normalized frequency thresholds of ten and forty perform, since the usual thresholds often fall between these two values. As soon as a corpus reaches at least 500 000 words, the threshold of ten selects only four-word sequences that successfully pass the inferential test and the threshold of forty guarantees the same result for all tested corpus sizes. For three-word sequences, thresholds above ten are necessary for all corpus sizes and thresholds higher than forty for corpora of 250 000 words or less.

5. Conclusion

This methodological study aimed to provide an initial response to the question of whether the frequency thresholds conventionally used for identifying lexical bundles were high enough to avoid selecting sequences that could have been produced by chance. To do this, the probability of ending up by chance with at least as many instances of the sequence as the number actually observed in the corpus was estimated by means of an extension of Fisher's exact test, applied to sequences of more than two words. A priori, such an outcome may seem unlikely, due to the implausibility of the null hypothesis underlying the inferential tests in the case of language: ‘Language is never, ever, ever, random’ (Kilgarriff, 2005: 263) and ‘Words are never combined at random in natural language’ (Evert, 2009:1244). The linguistic constraints that determine word ordering render unacceptable a number of sequences generated at random. These sequences are nevertheless considered by the null hypothesis of a randomly ordered corpus as possible, reducing the probability of observing acceptable sequences (Stubbs, 1995). Many sequences found in a corpus are therefore considered as extremely unlikely even if their formulaic character is far from evident (Ellis et al., 2015). Nevertheless, the analyses reported here showed that a non-negligible percentage of the three-word sequences, selected on the basis of usual frequency cut-offs, did not successfully pass the inferential test. Regarding the four-word sequences, which are by far the most studied using this methodology, the conventional thresholds were high enough to select only statistically significant sequences. These positive findings also apply to sequences of more than four words such as the extremely long lexical bundles studied by Cortes (2013), because they are necessarily less probable than the sequences of four words that compose them. One must nevertheless keep in mind that the method used in this study is designed for analysing frequent sequences and not for very long, but much more rare, fixed sequences such as proverbs.

The analyses highlighted the marked effect of corpus size on the efficiency of the frequency thresholds when they are expressed in normalized frequency. The smaller the corpus, the higher the threshold must be to ensure that all selected sequences pass the inferential test. This observation was not unexpected and can be rephrased as follows: the less data there are, the larger an effect

must be to be declared statistically significant. This relationship between significance and sample size is well-known in the field of statistics. Contrarily, the normalized frequency threshold is constant for all corpus sizes, except for the rounding problems discussed in Chen and Baker (2010). Applying the same normalized frequency threshold on corpora of various sizes leads to a larger number of sequences that does not pass the inferential test in the smaller corpora because the sample size determines, in part, the level of confidence one can have in the observed effect. Smaller sample means more uncertainty. This relationship underscores an issue for which opinions are divergent: can one use an identical normalized threshold to identify lexical bundles in corpora of different size? According to Biber and Barbieri (2007: 267), the answer is positive: 'By using a normalized rate of occurrence, we are able to compare the bundles across sub-corpora of different sizes'. Contrariwise Cortes (2015: 205) points out that 'Comparison of bundles yielded by small corpora and large corpora has been shown to be problematic because applying the usual normalization formula results in unreliable figures' while Hyland (2012: 151) stresses the need of more research to establish their validity. Since the current study adopts only one point of view – that of inferential tests – it cannot settle this debate. It nevertheless suggests that it is desirable to be cautious when presenting the conclusions of such an analysis.

An important limitation of this study lies in the very narrow perspective taken regarding the question of the reliability of the frequency thresholds. The approach is exclusively based on quantitative analyses and statistical inference. It must be remembered that each lexical bundle study systematically includes a qualitative analysis of the selected sequences and that this step allows researchers to control and refine the automatically extracted lists. The conclusions reported here apply only to the preliminary step of automatically selecting lexical bundles, the importance of which cannot be overestimated, however, since it is this automatic step that justifies classifying the approach amongst the strict corpus-driven approaches to formulaic language (Biber, 2009; Cortes, 2015).

A second limitation is that the analyses were focused on only one aspect of the problem that may arise when using a frequency threshold to select lexical bundles, i.e., will it select sequences that could have been produced by chance? In inferential statistics, this question corresponds to the risk of committing a Type I error. There is also a second risk – the risk of committing a Type II error by disregarding sequences that are not the result of chance. This is a risk accepted by researchers who employ a frequency cut-off, because their aim is to only extract sequences that show a statistical tendency to co-occur (Biber et al., 1999: 988-989; Hyland, 2008: 5). As such, this perspective was applied in the current study. However, it is not the only possible perspective. As argued by Sinclair and by Tognini-Bonelli (cited in Biber, 2009: 280), simple recurrence (i.e., two independent occurrences) may be sufficient for warranting a thorough linguistic description.

The procedure used to calculate probabilities also had a weakness that cannot be ignored, i.e., its computational cost, which limits the possible size of the corpus that can be analysed. It is therefore not applicable to the 425 million Corpus of Contemporary American English[6] (COCA), from which Lenko-Szymanska (2014) extracted the most frequent trigrams with a cut-off of 7.6 occurrences per million words to serve as a reference list for analysing a much smaller corpus of texts. More generally, the use of a Monte Carlo procedure severely limits the accuracy of estimated probabilities, since they cannot be smaller than one divided by the number of iterations. In contrast, the Fisher exact test, available in statistical software for the analysis of two-word sequences, can calculate probabilities as small as 1^{-320}. This lack of precision in the estimation of the smallest probabilities explains why the extension of the Fisher test to the sequences of more than two words is not a viable association index that is usable for ordering sequences from the more outstanding to the least one. It therefore does not address the need emphasized by Evert (2009: 1244), i.e., 'to develop suitable measures for word triples and larger n-tuples'; nonetheless, it may help in the assessment of proposed indices.

Other development paths can also be pursued. Although the use of diverse corpora in the first study provides at least some generality to the recommendations, it is regrettable that only one large corpus has been analysed (due to the extended computational time needed for doing so). It may be interesting to analyse in future work a relatively large corpus of academic writing, as several studies have pointed out that conversation and academic writing strongly differ by their number of lexical bundles (Conrad & Biber, 2004). It will also be interesting to apply the extension of the Fisher's exact test to sequences that include variable slots (Renouf & Sinclair, 1991).

Acknowledgement

This work was supported by the Fonds de la Recherche Scientifique (FRS-FNRS) under Grant J.0025.16. The author is a Research Associate of this institution. Computational ressources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Equipements de Calcul Intensif en Fédération Wallonie Bruxelles (CECI) funded by the FRS-FNRS.

References

Ädel, A., and B. Erman. 2012. 'Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach', *English for Specific Purposes 31* (2), pp. 81-92.

[6] http://corpus.byu.edu/coca/

Agresti, A. 1992. 'A survey of exact inference for contingency tables', *Statistical Science 7* (1), pp. 131-153.

Aijmer, K. 2009. 'The pragmatics of adverbs' in G. Rohdenburg and J. Schlüter (eds.) *One Language, Two Grammars? Differences between British and American English,* pp. 324-340. Cambridge: Cambridge University Press.

Altenberg, B. 1998. 'On the phraseology of spoken English: The evidence of recurrent word-combinations' in A. Cowie (ed.) *Phraseology: Theory, Analysis, and Applications,* pp. 101-122. Oxford: Oxford University Press.

Bestgen, Y. 2014. 'Extraction automatique de collocations : Peut-on étendre le test exact de Fisher à des séquences de plus de 2 mots?', *Actes de JADT 2014, pp.* 79-90.

Biber, D. 2009. 'A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing', *International Journal of Corpus Linguistics 14* (3), pp. 275-311.

Biber, D., and F. Barbieri. 2007. 'Lexical bundles in university spoken and written registers', *English for Specific Purposes 26* (3), pp. 263-286.

Biber, D., and S. Conrad. 1999. 'Lexical bundles in conversations and academic prose' in H. Hasselgard and S. Oksefjell (eds.), *Out of corpora: studies in honour of Stig Johansson*, pp. 181–190. Amsterdam: Rodopi.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Biber, D., and J. Jones. 2009. 'Quantitative methods in corpus linguistics' in A. Ludeling and M. Kytö (eds.) *Corpus Linguistics. An International Handbook*, pp. 1286-1304. Berlin: Mouton de Gruyter.

Carter, R. 2006. 'What Is advanced-level vocabulary? The case of chunks and clusters' in C. Coombe (Ed.) *Words Matter: The importance of Vocabulary in English Language Teaching and Learning*, pp. 23-41. Alexandria: TESOL.

Chen, Y., and P. Baker. 2010. 'Lexical bundles in L1 and L2 academic writing', *Language Learning & Technology 14* (2), pp. 30-49.

Chen, Y., and P. Baker. 2016. 'Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1', *Applied Linguistics, 37* (6), pp. 849–880.

Conrad, S., and D. Biber. 2004. 'The frequency and use of lexical bundles in conversation and academic prose', *Lexicographica: International Annual for Lexicography 20* (1), pp. 56-71.

Cortes, V. 2004. 'Lexical bundles in published and student disciplinary writing: Examples from history and biology', *English for Specific Purposes 23* (4), pp. 397-423.

Cortes, V. 2008. 'A comparative analysis of lexical bundles in academic history writing in English and Spanish', *Corpora 3* (1), pp. 43-57.

Cortes, V. 2013. 'The purpose of this study is to: Connecting lexical bundles and moves in research article introductions', *Journal of English for Academic Purposes* 12 (1), pp. 33-43.

Cortes, V. 2015. 'Situating lexical bundles in the formulaic language spectrum: Origins and functional analysis developments', in V. Cortes and E. Csomay *Corpus-based Research in Applied Linguistic*, pp. 197-216. Amsterdam: John Benjamins.

Crossley, S., and T. Salisbury. 2011. 'The development of lexical bundle accuracy and production in English second language speakers', *IRAL: International Review of Applied Linguistics in Teaching 49* (1), pp. 1-26.

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

Durrant, P. 2017. 'Lexical bundles and disciplinary variation in university students' writing: Mapping the territories', *Applied Linguistics 38* (2), pp. 165–193.

Ellis, N., R. Simpson-Vlach, U. Römer, M. O'Donnell, and S. Wulff. 2015. 'Learner corpora and formulaic language in SLA' in S. Granger, G. Gilquin and F. Meunier (eds.) *Cambridge Handbook of Learner Corpus Research*, pp. 357-378. Cambridge: Cambridge University Press.

Evert, S. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Unpublished PhD thesis. Stuttgart, DE: Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Evert, S. 2009. 'Corpora and collocations' in A. Ludeling and M. Kytö (eds.) *Corpus Linguistics. An International Handbook*, pp. 1211-1248. Berlin: Mouton de Gruyter.

Granger, S., E. Dagneaux, F. Meunier, and M. Paquot. 2009. *The International Corpus of Learner English*. Version 2. Handbook and CD-Rom. Louvain-la-Neuve: Presses Universitaires de Louvain.

Gray, B. 2016. 'Lexical Bundles' in P. Baker and J. Egbert (eds.), *Triangulating methodological approaches in corpus linguistic research*, pp. 33-55. New York: Routledge.

Gries, S. 2005. 'Null hypothesis significance testing of word frequencies: a follow-up on Kilgarriff', *Corpus Linguistics and Linguistic Theory 1* (2), pp. 277-294.

Gries, S. 2008. 'Corpus-based methods in analyses of SLA data' in P. Robinson and N. Ellis (eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition*, pp. 406-431. New York: Routledge.

Gries, S. 2010. ‘Useful statistics for corpus linguistics’ in A. Sánchez and M. Almela (eds.) *A Mosaic of Corpus Linguistics: Sselected Approaches*, pp. 269-291. Frankfurt am Main: Peter Lang.

Gries, S. 2015. ‘Some current quantitative problems in corpus linguistics and a sketch of some solutions’, *Language and Linguistics 16* (1), pp. 93–117.

Groom, N. 2009. ‘Effects of second language immersion on second language collocational development’ in A. Barfield and H. Gyllstad (eds.) *Researching Collocations in Another Language*, pp. 21-33. London: Palgrave Macmillan.

Holm, S. 1979. ‘A simple sequentially rejective multiple test procedure’, *Scandinavian Journal of Statistics 6* (2), pp. 65-70.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hyland, K. 2008. ‘As can be seen: Lexical bundles and disciplinary variation’, *English for Specific Purposes 27* (1), pp. 4-21.

Hyland, K. 2012. ‘Bundles in academic discourse’, *Annual Review of Applied Linguistics 32*, pp. 150-169.

Jones, S., and J. Sinclair. 1974. ‘English lexical collocations. A study in computational linguistics’, *Cahiers de Lexicologie 24* (1), pp. 15-61.

Juknevičienė, R. 2009. ‘Lexical bundles in learner language: Lithuanian learners vs. native speakers’, *KaLBOTYRa 61* (3), pp. 61-72.

Kilgarriff, A. 2005. ‘Language is never, ever, ever random’, *Corpus Linguistics and Linguistic Theory 1* (2), pp. 263-275.

Leńko-Szymańska, A. 2014. ‘The acquisition of formulaic language by EFL learners: A cross-sectional and cross-linguistic perspective’, *International Journal of Corpus Linguistics 19* (2), pp. 225-251.

Lin, Y. 2013. ‘Discourse functions of recurrent multi-word sequences in online and face-to-face intercultural communication’ in J. Romero-Trillo (ed.) *Yearbook of Corpus Linguistics and Pragmatics*, pp. 105-129. Dordrecht: Springer.

McCarthy, M., and R. Carter. 2006. ‘This that and the other: Multi-word clusters in spoken English as visible patterns of interaction’ in M. McCarthy, *Explorations in Corpus Linguistics*, pp. 7-26. Cambridge, Cambridge University Press.

Moore, R. 2004. ‘On log-likelihood-ratios and the significance of rare events’ in *Proceedings of EMNLP 2004*, pp. 333-340. Barcelona, Spain. July 25-26. Association for Computational Linguistics.

O’Keeffe, A., M. McCarthy, and R. Carter. 2007. *From Corpus To Classroom*. Cambridge: Cambridge University Press.

Pecina, P. 2010. 'Lexical association measures and collocation extraction', *Language Resources & Evaluation 44* (1), pp. 137-158.

Pedersen, T. 1996. 'Fishing for exactness' in *Proceedings of the South Central SAS Users Group*, pp. 188-200. Austin, TX, October 27–29.

Pedersen, T., M. Kayaalp, and R. Bruce. 1996. 'Significant lexical relationships' in *Proceedings of the 13th National Conference on Artificial Intelligence*, pp. 455-460.

Renouf, A., and J. Sinclair. 1991. 'Collocational frameworks in English' in K. Aijmer and B. Altenberg (eds) *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, pp. 128-143. London: Longman.

Salazar, D. 2014. *Lexical Bundles in Native and Non-native Scientific Writing: Applying a Corpus-based Study to Language Teaching*. Amsterdam: John Benjamins.

Scott, M., and C. Tribble. 2006. *Textual Patterns. Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.

Simpson-Vlach, R., and N. Ellis. 2010. 'An academic formulas list: New methods in phraseology research', *Applied Linguistics 31* (4), pp. 487-512.

Stefanowitsch, A., and S. Gries. 2003. 'Collostructions: Investigating the interaction of words and constructions', *International Journal of Corpus Linguistics 8* (2), pp. 209-243.

Stubbs, M. 1995. 'Collocations and semantic profiles: On the cause of the trouble with quantitative studies', *Functions of Language 2* (1), pp. 23-55.

Stubbs, M. 2002. 'Two quantitative methods of studying phraseology in English', *International Journal of Corpus Linguistics 7* (2), pp. 215-244.

Thewissen, J. 2013. 'Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus', *Modern Language Journal 97* (S1), pp. 77-101.

Vidakovic, I., and F. Barker. 2010. 'Use of words and multi-word units in Skills for Life Writing Examinations', *Cambridge ESOL: Research Notes 41*, pp. 7-14.

Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Zelterman, D., I. Chan, and P. Mielke. 1995. 'Exact tests of significance in higher dimensional tables', *The American Statistician 49* (4), pp. 357-361.