This is an accepted manuscript of an article published by Taylor & Francis in the Journal of Quantitative Linguistics on 05 Feb 2019, available online at https://www.tandfonline.com/doi/full/10.1080/09296174.2019.1566975

Comparing lexical bundles across corpora of different sizes: the Zipfian problem

Yves Bestgen
Centre for English Corpus Linguistics, Université catholique de Louvain, Louvain-la-Neuve, Belgium

ABSTRACT

Formulaic sequences in language use are often studied by means of the automatic identification of frequently recurring series of words, often referred to as 'lexical bundles', in corpora that contrast different registers, academic disciplines etc. As corpora often differ in size, a critically important assumption in this field states that the use of a normalized frequency threshold, such as 20 occurrences per million words, allows for an accurate comparison of corpora of different sizes. Yet, several researchers have argued that normalization may be unreliable when applied to frequency threshold. The study investigates this issue by comparing the number of lexical bundles identified in corpora that differ only in size. Using two complementary random sampling procedures, subcorpora of 100,000 to two million words were extracted from five corpora, with lexical bundles identified in them using two normalized frequency thresholds and two dispersion thresholds. The results show that many more lexical bundles are identified in smaller subcorpora than in larger ones. This size effect can be related to the Zipfian nature of the distribution of words and word sequences in corpora. The conclusion discusses several solutions to avoid the unfairness of comparing lexical bundles identified in corpora of different sizes.

Yves Bestgen (2019) Comparing Lexical Bundles across Corpora of Different Sizes: The Zipfian Problem, Journal of Quantitative Linguistics, DOI: 10.1080/09296174.2019.1566975

INTRODUCTION

One of the most frequently used approaches for studying formulaic sequences in language is based on the automatic identification of frequently recurring series of words in a corpus, often referred to as 'lexical bundles' (Biber, Johansson, Leech, Conrad, & Finegan, 1999), but also 'recurrent word combinations' (Altenberg, 1998) and 'clusters' (Scott &Tribble, 2006). These are sequences such as 'I don't know what', 'as can be seen', 'in the case of' or 'there was no significant', which are 'recurrent expressions, regardless of the idiomaticity, and regardless of their structural status' (Biber *et al.,* 1999, p. 990). The majority of the studies that use this approach compare lexical bundles present in two or more corpora, which contrast different registers, varieties of English, academic disciplines, historical periods, levels of language proficiency of the authors, among other factors. These studies have greatly increased our knowledge of formulaicity in language use, resulting in important implications for domains such as translation studies (e.g., Xiao, 2011; Lee, 2013), language for academic purposes and for specific purposes (e.g., Cortes, 2004; Hyland, 2008), and, more generally, language learning and teaching (e.g., Biber, Conrad, & Cortes. 2004; Chen & Baker, 2016).

Lexical bundles are identified in a corpus on the basis of two criteria: a minimum frequency threshold whose goal 'is to identify bundles that recur often enough to be regarded as typical', and a minimal number of documents in which a sequence must be present to ensure 'that the bundles are typical of the entire corpus, not just a few texts' (Pan, Reppen, & Biber, 2016, p. 63). The minimum frequency threshold is expressed in the form of a normalized frequency threshold, usually in a number of occurrences per million words, and then converted into a raw frequency, which must be reached in a corpus of a given size, by multiplying this normalized threshold by the ratio between the size of the corpus and one million (i.e., a threshold of 20 occurrences per million words is converted for a corpus of 250,000 words into a raw frequency threshold of five by multiplying 20 by 250,000 divided by one million). This normalization is supposed to allow for an accurate comparison of corpora of different sizes (Allan, 2016; Biber & Barbieri, 2007; Pickering & Bird, 2008; Reppen, 2009). This is a critically important assumption in this field of research, in which comparing corpora of different sizes is very common (e.g., Ädel and Erman, 2012; Allan, 2016; Berglund, 2000; Bychkovska & Lee, 2017; Dutra, Orfano, & Berber Sardinha*,* 2014; Huang, 2015; Hyland, 2008; Juknevičienė, 2009; Lee, 2013; Reppen, 2009). For example, Chen and Baker (2016) compared the lexical bundles in learner corpora ranging in size from 26,000 words to 88,000 words, a ratio of one to three. Biber *et al.* (2004) analysed them in the case of two university registers on the basis of corpora of 760,000 and 1.25 million words, as well as comparing them to those found in previous research on conversation and academic prose, based on corpora of 5.3 million words and seven million words, a maximum ratio of one to nine. Biber and Barbieri (2007) analysed the use of bundles in nine spoken and written university registers on the basis of corpora of which the smallest was 39,000 words long and the largest was 5.3 million words long, a ratio of one to 135.

However, several researchers have argued that using the same normalized threshold to compare bundles in corpora of different sizes is potentially problematic (Bestgen, 2018; Chen & Baker, 2016; Cortes, 2008, 2015; Gray, 2016; Hyland, 2012; Oakey, 2009; Schnur, 2014). According to Hyland (2012, p. 151): 'Such normalization methods, which are widely used to compare individual words across different sized corpora, may, however, be unreliable when working with lexical bundles, and more research is needed to establish their validity.' Chen and Baker (2016) also highlight the need for further research on this topic. Cortes (2015, p. 205) is a lot more pessimistic, stressing 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'. The empirical arguments are nevertheless scarce. In a study aimed at comparing three-word bundles in corpora, which contained either the same number of tokens or the same number of documents, Oakey (2009) observed that the most frequent bundles in a corpus of five million words were also the most frequent in a subcorpus

whose size was approximately 1.8 million words. However, the two compared corpora were relatively large for the field and the analysis was focused on the most common bundles, rather than on the number of word sequences that reach a normalized frequency threshold in corpora of different sizes. Using a statistical inferential test for determining whether the usual frequency thresholds for identifying lexical bundles in corpora were high enough to avoid selecting sequences that could result from chance, Bestgen (2018) observed a strong effect of corpus size on the efficiency of a normalized thresholds. The smaller the corpus, the higher the threshold must be to select only sequences that pass the inferential test. This result suggests that the use of normalized thresholds could be problematic. However, one might wonder if it means more than the well-known relation in statistics between sample size and significance.

Cortes reported in her unpublished PhD dissertation (2002, pp. 72-75) an analysis directly addressing this issue. In this exploratory study, she extracted four-word bundles from a four-million-word corpus using a normalized frequency threshold of 20 occurrences per million words and a dispersion criterion of five texts. She then divided the corpus into subcorpora of different sizes and extracted the bundles using the same thresholds. She observed that the smaller the corpus, the more bundles it contains. Ninety-one bundles were identified in the 500,000-word subcorpus, while 84 were identified in the two-million-word subcorpus and only 75 in the four-million-word corpus. These differences may not seem huge; but they correspond to an increase of 121% in the case of the 500,000-word corpus, and it is unknown how many bundles would have been identified in smaller subcorpora. Since the study was exploratory, Cortes analysed only one of the subcorpora of each size, leaving open the possibility that a different number of bundles would have been identified in the other subcorpora, especially since the divisions were carried out by respecting the original order of the documents in the full corpus. It is also unknown what would have been observed with a higher normalized frequency threshold and whether the results can be generalized to other corpora. It is therefore difficult to reach a conclusion from this study alone. It nevertheless suggests that the size of the corpora could matter. If this is truly the case, it would affect not only the comparison of the total number of lexical bundles in corpora of different sizes as discussed above, but also the comparison of the proportions of bundles after their categorization according to their structural characteristics or the qualitative analyses of specific bundles. All these analyses are carried out on word sequences that pass a normalized frequency threshold and can thus be affected by any size difference between the compared corpora.

Of course, size is not the only dimension by which corpora differ in these studies and researchers are well aware of the many problems that can arise. Why then does the difference in size deserve special attention? Three reasons can be put forward: it is quite ubiquitous; its impact could be very strong since some authors consider that it makes the analysis incorrect; and there is at least one relatively simple solution if the problem is shown to be real, that is, comparing corpora of sizes as similar as possible.

The aim of the study is to contribute to the scientific discussion on this issue by comparing the number of lexical bundles identified in corpora that differ only in size. As explained in the next section, the analyses were conducted on five well-known corpora of which 100,000- to two-million-word subcorpora were extracted by two complementary random sampling procedures. These analyses show that the size of a corpus has a significant impact on the number of bundles identified. As observed by Cortes (2002), the smaller the corpus, the larger the number of bundles identified for a given normalized frequency threshold. An explanation is proposed in the discussion section. It relies on the Zipfian distribution of the frequency of words and sequences of words in natural language, in corpora and in texts: there are always many more rare words and rare sequences of words than frequent ones. The conclusion considers several solutions to avoid the unfairness of comparing corpora of different sizes.

DATA AND METHODOLOGY

Corpora and lexical bundle identification

The first two corpora analysed in this study are the FLOB (Freiburg LOB Corpus of British English) and the FROWN (Freiburg Brown Corpus of American English) corpora, which are available on the ICAME CD-ROM (Hofland, Lindebjerg, & Thunestvedt, 1999). Each corpus contains a million words, corresponding to 500 extracts of approximately 2,000 words from texts published in the early 1990s. These corpora were often used in linguistics, including in studies based on the lexical bundle approach. As they have been compiled to be as similar as possible, except in terms of the variety of English, comparing the number of lexical bundles identified in a small subcorpus of the FLOB corpus to the number of lexical bundles identified in the full FROWN corpus, and vice versa, will help to get an idea about the potential impact of the corpus size on the findings of a study.

The third corpus is the International Corpus of Learner English (ICLE, Granger, Dagneaux, & Meunier, 2002), a corpus of essays written by learners of English as a foreign language. It has been used in a series of studies aimed at describing the typical bundles of non-native speakers. It is significantly larger than the first two corpora, since it contains over 2.4 million words from 3,583 learners of 11 different mother-tongue backgrounds. It is composed mainly of texts of similar size, the average length being 680 words, with half of the texts between 504 and 787 words long, even if 1% of them are less than 220 words long (minimum: 107 words) and 1% are more than 1,780 words long (maximum: 4,139 words).

The last two corpora are composed of texts of very variable sizes ranging from just 1,000 words to more than 160,000 words. They are extracted from the British National Corpus (BNC[1]) and correspond to two registers, which were the topic of the first analyses labelled 'lexical bundles': conversation and academic prose (Biber *et al.,* 1999). Since it was preferable in the present study to compare initial corpora of the same size, they were constructed using the following procedure. First, texts for a total number of words as close as possible to four million were randomly selected from the conversation section of the BNC, whose size is approximately 4.2 million words, on the condition that these texts contain at least 1,000 words. This step resulted in a corpus of 142 texts and 4,000,008 words. Then, the same number of texts was randomly selected from the academic section of the BNC, whose size is approximately 15.7 million words, so as to obtain the same number of words.

In all the analyses reported below, potential bundles were identified on the basis of the sequences of lower-cased word forms, uninterrupted by any punctuation mark.

Procedures used to build the subcorpora

The procedure used to build the subcorpora must allow for a comparison of the number of lexical bundles identified in subcorpora as similar as possible to the initial corpus, except for their size. Among the possible procedures for constructing them, two seem particularly adequate because they respect the usual unity of construction for a corpus: documents that are texts or extracts of texts. The simplest procedure consists of randomly selecting a sample of documents in the full corpus so that the total number of words in the sample corresponds as closely as possible to the desired number. The second procedure aims at constructing a subcorpus, which constitutes a fraction of the full corpus, whose numerator is one and whose denominator is an integer greater than one, for example, a half or a third. This involves dividing each document in the original corpus into a number of (almost) equally sized segments corresponding to the denominator of the fraction. These segments always start immediately after and ends just before a punctuation mark. Then, a segment is randomly selected from each document. In this way, it is possible to build many different subcorpora of

[1] http://www.natcorp.ox.ac.uk.

a given size, which are made of an extract of each document whose length is proportional to the total number of words in this document.

The first procedure has the advantage of respecting the integrity of the original documents, but it may face difficulties in constructing small subcorpora if the original corpus is composed in part of large documents. The second procedure can be applied to corpora containing very long documents as well as short documents, but the length of an extract can be very small if the original corpus already contains very small documents. Since each procedure has advantages and disadvantages, and since both situations could occur in actual studies (comparing two corpora containing documents of similar sizes, but one containing significantly fewer documents than the other, or two corpora containing a similar number of documents, but one being significantly smaller), they were both employed. There is an additional reason for using the second procedure: only this one is applicable to the two corpora extracted from the BNC because of the very large differences in size between the documents that comprise them[2].

In both of these procedures, the sample is drawn without replacement since corpora normally never contain the same document more than once. This operation is repeated 100,000 times in order to limit the impact of random variability and therefore to estimate with sufficient precision the number of bundles identified in a subcorpus on the basis of a given normalized frequency threshold. In sampling theory terms, the samples are drawn using a 'simple random sampling' procedure, which consists of extracting a random sample without replacement, in which each observation in the population has the same probability of being selected (Cochran, 1977; Thompson, 2012). In the present study, the observation is a word sequence, the sample is a subcorpus and the population is a full corpus. This type of sampling, the most classical in statistics, allows for an unbiased estimate of a parameter of a population, which means that the expected value calculated on the basis of all the possible samples will be equal to the true population value, and that each sample is as likely to produce a higher estimate than a lower estimate of this population value. For a given population, the larger the sample, the more precise the estimation of the parameter will be (Thompson, 2012, p. 17). This is particularly evident when the sample covers a sizeable proportion of the population, at least 5 to 10% (Cochran, 1977, p. 25), which is the case in the majority of the samples analysed in this study. As a result, the estimates based on the largest samples (50% of the population in this study) are less likely to deviate from the population value compared to those based on the smaller samples (2.5%). But it is important to note that this only affects the variability of the estimate and not its unbiasedness. By extracting a large number of independent samples, the accuracy of the estimate is further increased and therefore the average value obtained should be very close to the value in the population. The most important consequence of this sampling theory-based analysis is that the procedure used to answer the research question does not disadvantage a response in favour of the use of a normalized frequency threshold for identifying lexical bundles in corpora of different sizes.

Compared conditions

On the basis of the most commonly used values in the literature (Chen & Baker, 2010), the following conditions were compared: subcorpus sizes ranging from 100,000 words to half the total size of the full corpus, sequences of three and four words, normalized frequency thresholds of 20 and 40 occurrences per million words, and dispersion thresholds of three and five texts[3].

---

[2] If the procedure based on the selection of documents had been used, those of 160,000 words would have had no chance of being selected for a subcorpus of 100,000 words, and their selection for a subcorpus of 250,000 words would have imposed major constraints on the length of the other documents that could be associated with them in the subcorpus.

[3] The dispersion threshold is sometimes expressed as a percentage of the number of texts that makes up the corpus (Hyland, 2008). Such a threshold is more demanding for corpora that

ANALYSES AND RESULTS

The 100,000 samples were extracted by each sampling procedure for each combination of a corpus, a subcorpus size and a word sequence length. Each of these samples was used to determine the number of lexical bundles identified according to the two normalized frequency thresholds and the two dispersion thresholds. The dispersion threshold introduces a difficulty because, in certain combinations of conditions, it may be greater than the raw frequency threshold obtained on the basis of the corpus size and the desired normalized frequency threshold (e.g., a subcorpus size of 100,000 words and a normalized threshold of 20 leads to a raw frequency threshold of two occurrences, which is below both dispersion thresholds). In these cases, the analyses were not performed.

Since the two sampling procedures yielded very similar outcomes, the results presented in this section are based on the subcorpora composed of a random sample of the documents in the full corpus, except in the case of the BNC, for which only the procedure based on sampling extracts is possible. The results that are not presented in this section can be seen in the Supplemental material.

The following tables give the average numbers of lexical bundles identified according to the frequency and dispersion thresholds in the 100,000 samples for each subcorpus size, as well as the number of bundles identified in the full corpora. They also give, for each subcorpus size, the ratio in percentage terms between the number of lexical bundles identified in the subcorpus and the number in the full corpus since a difference of 50 bundles is more important when there are 75 bundles in the full corpus than 1,000. An index of variability or a confidence interval for the mean is not given because, as explained above, this variability is strongly affected by the subcorpus size.

It is important to stress that the averages reported below are based on randomizations that do not all produce exactly the same values. It is therefore questionable whether the results are reliable, even though a relatively large number of randomizations have been performed (i.e., 100,000 for each of the 132 independent analyses). Of course, one can always perform more randomization, but a replication experiment in the strictest sense, which consists of starting over the same experiment with the same corpora under the same conditions, while only modifying the random drawings, confirmed that the estimate of the average number of bundles for each condition is extremely accurate. For all the results reported below, the maximum difference in absolute value between the data in the tables and the values in the replication is less than 0.4.

As shown in Tables 1 to 3, there are always more lexical bundles selected from the subcorpora than in the corresponding full corpus, and the difference may be large since it can exceed a 200% increase[4]. The tables also show that, the smaller the subcorpus, the greater the number of bundles selected when the same normalized frequency threshold is used. This conclusion is valid for the five corpora, the two sequence lengths, the two normalized thresholds and the two dispersion thresholds. The comparison of the FLOB and FROWN corpora for the three-word sequences (Table 1) highlights the potential impact of this relationship between the corpus size and the number of bundles identified on the basis of a normalized threshold. The two tested normalized frequency thresholds lead to the

---

contain the larger number of texts, which are often the biggest. In the present study, this is not of much interest, since the concern in the literature is that small corpora favour the identification of a larger number of bundles (Cortes, 2015).

[4] However, in a few rare cases, the same number of lexical bundles was identified in the full corpus and in the largest subcorpus when the average number in this subcorpus is rounded to the nearest integer.

identification of more bundles in the British English FLOB corpus than in the American English FROWN corpus. This is also true when subcorpora of the same size are compared. However, if a small subcorpus derived from the FROWN corpus is compared to the full FLOB corpus, the difference is reversed. Table 3 indicates, as already observed by Biber *et al.* (1999), that the conversational register allows for the identification of a significantly larger number of bundles than the academic register. The size of this difference is large enough to be observed even when comparing the smallest academic subcorpus to the full conversational corpus. However, the importance of the difference varies markedly according to the sizes of the compared (sub)corpora, since it can go from 200% to more than 800%.

Table 1: Number and percentage (second line) of lexical bundles for each set of conditions in the FLOB and FROWN corpora.

Disp.	Norm. freq.	100,000	150,000	200,000	250,000	500,000	Full
Three-word bundles in FLOB							
3	20		1302	1174	1056	822	743
			175	*158*	*142*	*111*	*100*
	40	375	298	260	239	196	180
		208	*166*	*144*	*133*	*109*	*100*
5	20				854	812	742
					115	*109*	*100*
	40		260	249	233	195	180
			144	*138*	*129*	*108*	*100*
Three-word bundles in FROWN							
3	20		1073	979	872	655	570
			188	*172*	*153*	*115*	*100*
	40	294	234	203	184	149	129
		228	*181*	*157*	*143*	*116*	*100*
5	20				689	646	569
					121	*114*	*100*
	40		203	195	181	149	129
			157	*151*	*140*	*116*	*100*
Four-word bundles in FLOB							
3	20		130	107	91	64	53
			245	*202*	*172*	*121*	*100*
	40	29	21	18	16	11	9
		322	*233*	*200*	*178*	*122*	*100*
5	20				69	63	53
					130	*119*	*100*
	40		18	17	15	11	9
			200	*189*	*167*	*122*	*100*
Four-word bundles in FROWN							
3	20		95	78	66	45	38
			250	*205*	*174*	*118*	*100*
	40	21	16	13	12	10	9
		233	*178*	*144*	*133*	*111*	*100*
5	20				51	45	38
					134	*118*	*100*
	40		14	13	12	10	9
			156	*144*	*133*	*111*	*100*

Regarding the impact of the different conditions, the effect of the subcorpus size on the number of bundles is more important for the three-word sequences than for the four-word sequences, while the opposite difference is observed for the percentages based on the number of bundles identified in the full corpus. It is not easy to draw any conclusion, given the very large difference in the number of bundles identified for these two sequence lengths. The same difference contaminates the comparison of the five corpora. The comparison of the two normalized frequency thresholds shows that, almost systematically, the higher the normalized threshold, the lower the difference in both raw number and percentage. The rare exceptions

are observed when the number of bundles is small or when the raw frequency threshold, to which the normalized threshold corresponds, is equal or close to the dispersion threshold (for example, a dispersion threshold of five and a normalized frequency threshold of 20 in a subcorpus of 250,000 words). However, the cost of using a normalized threshold of 40 occurrences per million words is not negligible: only a small number of four-word bundles are identified in the FLOB and FROWN corpora and in the academic prose corpus. Moreover, if the use of a more conservative frequency threshold reduces the differences, it does not cancel them. Increasing the dispersion threshold reduces the size of the effect, but only for the smaller subcorpora, where it is close to the raw frequency threshold corresponding to the normalized threshold.

Table 2: Number and percentage (second line) of lexical bundles for each set of conditions in the ICLE.

Disp.	Norm. freq.	100,000	150,000	200,000	250,000	500,000	1 million	Full
			Three-word bundles					
3	20		2838	2590	2373	1960	1793	1723
			165	*150*	*138*	*114*	*104*	*100*
	40	965	814	738	694	612	574	555
		174	*147*	*133*	*125*	*110*	*103*	*100*
5	20				2090	1952	1792	1723
					121	*113*	*104*	*100*
	40		761	724	689	612	574	555
			137	*130*	*124*	*110*	*103*	*100*
			Four-word bundles					
3	20		582	500	439	329	285	266
			219	*188*	*165*	*124*	*107*	*100*
	40	152	117	102	93	78	73	69
		220	*170*	*148*	*135*	*113*	*106*	*100*
5	20				371	327	285	266
					139	*123*	*107*	*100*
	40		106	99	92	79	73	69
			154	*143*	*133*	*114*	*106*	*100*

As indicated above, the results are very similar when using the second sampling procedure, which is based on the random selection of an extract from each document in the original corpus. Table 4 presents the results for the three-word sequences in the FLOB corpus by means of this second sampling procedure. The numbers of bundles are very close to those obtained by the text-based approach (see the corresponding columns in Table 1), except when the raw frequency threshold is equal to the dispersion threshold (e.g., a normalized threshold of 20 occurrences per million words and a dispersion threshold of five in a subcorpus of 250,000 words). This result can be explained by the fact that the subcorpora built by this second procedure contain significantly more documents (even if the extracts are shorter) than those produced by the text-based approach, such that it is easier to reach a dispersion threshold equal to the raw frequency threshold.

DISCUSSION

The analyses reported above clearly answer the research question. When comparing corpora that differ only in size, the smaller the corpus, the larger the number of lexical bundles selected. However, these analyses focused on only five corpora. Even if more corpora had been analysed and produced the same conclusion, the question would remain as to whether this conclusion also applies to any other corpus. The situation would be less uncertain if the origin of the size effect on the number of bundles identified could be traced and if it could be shown that any corpus would be affected. This is the purpose of this section.

Table 3: Number and percentage (second line) of lexical bundles for each set of conditions in the conversation and academic prose corpora.

Disp.	Norm. freq.	100,000	200,000	250,000	500,000	1 million	2 million	Full
Three-word bundles in conversation								
3	20		3106	2919	2507	2318	2229	2198
			141	*133*	*114*	*105*	*101*	*100*
	40	1338	1103	1056	969	927	907	889
		151	*124*	*119*	*109*	*104*	*102*	*100*
5	20			2471	2493	2316	2229	2198
				112	*113*	*105*	*101*	*100*
	40		1086	1050	968	927	907	889
			122	*118*	*109*	*104*	*102*	*100*
Three-word bundles in academic prose								
3	20		1664	1564	1327	1208	1147	1121
			148	*140*	*118*	*108*	*102*	*100*
	40	590	457	430	377	349	332	329
		179	*139*	*131*	*115*	*106*	*101*	*100*
5	20			1154	1245	1170	1123	1106
				104	*113*	*106*	*102*	*100*
	40		420	406	366	342	328	326
			129	*125*	*112*	*105*	*101*	*100*
Four-word bundles in conversation								
3	20		751	677	536	479	451	434
			173	*156*	*124*	*110*	*104*	*100*
	40	263	192	180	158	147	141	141
		187	*136*	*128*	*112*	*104*	*100*	*100*
5	20			549	532	478	451	434
				126	*123*	*110*	*104*	*100*
	40		189	178	157	147	141	141
			133	*126*	*111*	*104*	*100*	*100*
Four-word bundles in academic prose								
3	20		239	214	163	142	128	121
			198	*179*	*135*	*117*	*106*	*100*
	40	70	48	45	38	34	32	31
		226	*155*	*145*	*123*	*110*	*103*	*100*
5	20			147	150	135	125	121
				121	*124*	*112*	*103*	*100*
	40		43	41	36	34	32	31
			139	*132*	*116*	*106*	*103*	*100*

Table 4: Number and percentage (second line) of three-word lexical bundles in the FLOB corpus by means of the second sampling procedure.

Disp.	Norm. freq.	1/10 100,000	1/5 200,000	1/4 250,000	1/2 500,000	Full
3	20		1207	1060	824	743
			162	*143*	*111*	*100*
	40	380	254	234	196	180
		211	*141*	*130*	*109*	*100*
5	20			954	819	742
				129	*110*	*100*
	40		252	233	196	180
			140	*129*	*109*	*100*

The use of a normalized frequency threshold is based on the commonly recommended procedure to compare the frequency of a word or a sequence of words in two corpora of different sizes: the normalization procedure, which consists of norming the frequency of the term in each corpus to a common base, for example, in occurrence per million words (Biber *et al.,* 1998; McEnery, Xiao, & Tono*,* 2006). If a word or a sequence of words is observed 80 times in a corpus of two million words, it can be expected that it will be observed

(approximately) 20 times in a 500,000-word corpus and four times in a 100,000-word corpus, everything else being equal. The normalization procedure will convert these three numbers to a value of 40 occurrences per million words. It seems logical to modify a frequency threshold accordingly: if the chosen threshold is 40 occurrences per million words, a raw frequency threshold of 80 occurrences will be used for a two-million-word corpus and of four occurrences for a 100,000-word corpus.

The problem with this conjecture is that applying normalization to a frequency threshold means applying it to all the words or sequences of words in a corpus (Gray, 2016). If there are 50 four-word sequences that occur at least 40 times in a one-million-word corpus, it is claimed that there will also be 50 that will occur at least eight times in a 200,000-word corpus identical to the first, except in terms of size, and also 50 that will occur at least four times in a 100,000-word corpus. This assertion neglects a fundamental property of the frequency distribution of words (Baayen, 2001; Baroni, 2008; Zanette & Montemurro, 2005; Zipf, 1935/1965), but also of word sequences (Bannard & Lieven, 2009; Baroni, 2008; Ha, Sicilia-Garcia, Ming, & Smith, 2002), in human languages: its Zipfian nature, which has been observed in each analysed natural language and for all the lengths of texts and corpora from a few thousand words up to several tens of millions. In any text or corpus, 'a few words occur with very high frequency while many words occur but rarely' (Zipf, 1935/1965, pp. 40-41), and this overrepresentation of rare items is larger for smaller texts and corpora (Baayen, 2001; McEnery & Gabrielatos, 2006; Zeldes, 2013; Zipf, 1935/1965). However, when the same normalized frequency threshold is used in corpora of different sizes, this overrepresentation of rare words and rare sequences in the smaller corpora is not taken into account and a disproportionately large number of word sequences is selected from them.

The Zipfian problem, resulting from the use of the same normalized frequency threshold in corpora of different sizes, can be illustrated graphically by means of the well-known Zipf curve, the rank/frequency plot on which the horizontal axis plots the word sequence frequency using a logarithmic scale, and the vertical axis plots the rank of the word (also using a logarithmic scale) when the words are ordered from the most frequent to the less frequent. Figure 1 shows this Zipf curves for the four-word sequences in the full ICLE corpus and in the six subcorpora analysed above using a dispersion threshold of three. For the subcorpora, the plotted lines correspond to the average curves based on the 100,000 random samples. As can be seen, they are broadly in the shape of a straight line, suggesting that the rank of a four-word sequence is approximately inversely proportional to its frequency as stated by Zipf's law. These curves also illustrate a weaker version of Zipf's law, stated in the following way by Manning and Schütze (1999, p. 24): 'there are a few very common words, a middling number of medium frequency words, and many low frequency words'. The stepped form, especially towards the small frequencies, confirms the overrepresentation of the low frequency word sequences. In this figure, each horizontal line indicates, for each corpus size, the raw frequency threshold corresponding to a normalized threshold of 40 occurrences per million words. The vertical lines indicate the number of lexical bundles identified in each of these subcorpus sizes. They correspond (on a logarithmic scale) to the values given in Table 2.

It is important to remember that these Zipf curves have been drawn by considering the dispersion threshold, as is the rule in the lexical bundle approach to formulaicity. The situation would be even more disproportionate if this second criterion had not been used, since it is much more difficult for a word sequence occurring four times in a 100,000-word corpus, thus passing a normalized threshold of 40 occurrences per million words, to occur in three different texts, compared to a sequence occurring 20 times in a 500,000-word corpus. Another consequence of using a dispersion criterion is that it makes infeasible a probabilistic analysis based on the exact distribution or its approximation by a parametric model (Baayen, 2001; Baroni, 2008), at least according to the current state of knowledge. Without this criterion, it is possible, in the case of words, to use a binomial interpolation based on the

larger of the two corpora (Baayen, 2001, pp. 63-69) in order to estimate the raw frequency threshold that should be used in the smaller corpus, such that, all things being equal, the same number of words would be selected. The fact that the distribution of the word sequences in a corpus is Zipfian (Bannard & Lieven, 2009; Baroni, 2008; Ha *et al.,* 2002) suggests that this procedure could be extended to word sequences. But this remains to be verified, and, as noted above, this is not applicable, together with a dispersion threshold, since the frequency distributions are truncated below this threshold.

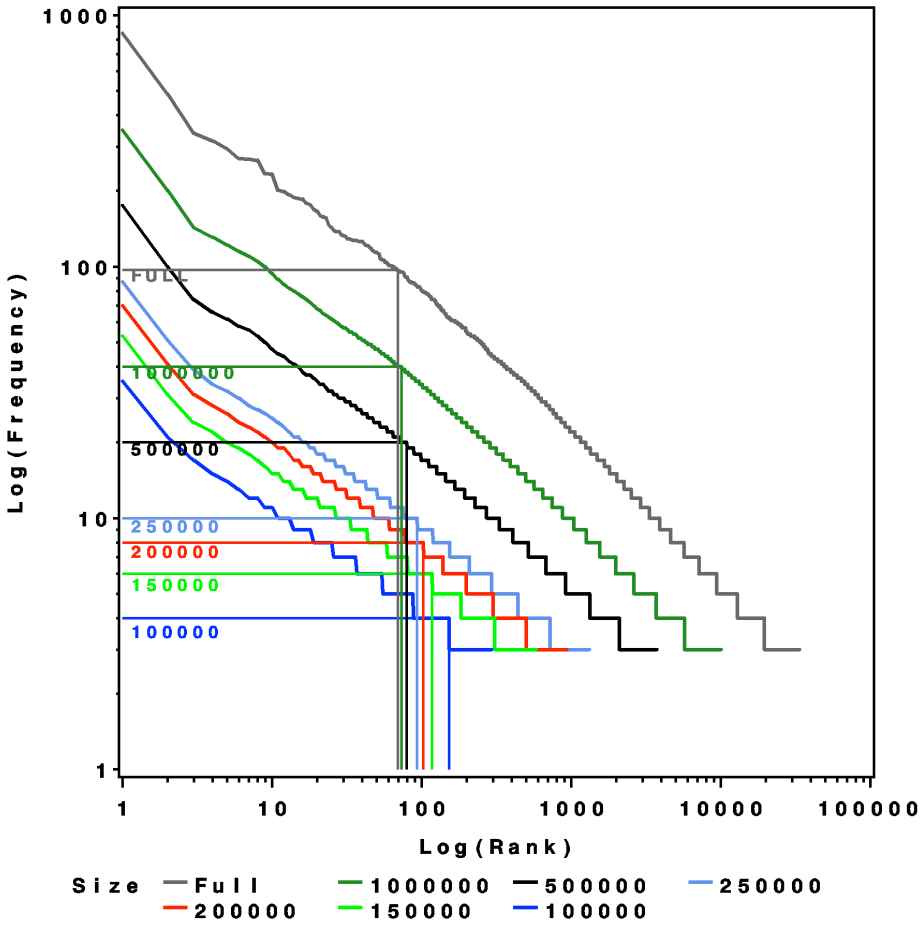

Figure 1: Zipf curves for the four-word sequences in the ICLE.

CONCLUSION

The present study aimed at determining whether the use of a normalized frequency threshold allows for extracting lexical bundles from corpora of different sizes in an unbiased way, in order that they can be compared. The performed analyses indicated that this is not the case: the smaller the corpus, the greater the number of bundles selected by a given normalized threshold. These results were obtained through the analysis of five well-known corpora of different genres and sizes. The generality of this conclusion is further reinforced by the possibility to relate these observations to the Zipfian nature of the distribution of vocabulary in natural language corpora. Since, in such corpora, there are many more rare word sequences than frequent ones, and since this disproportionality increases in relation to the smallness of the corpus, using a lower raw frequency threshold in a smaller corpus involves selecting more bundles than in a larger corpus, unless the raw frequency threshold becomes so high that only high frequency sequences are selected. This is a first potential solution to avoid the size issue highlighted in this study. This condition can be achieved in two ways: by increasing the normalized threshold or by analysing corpora of relatively large sizes. These two options can, of course, be used simultaneously. The results presented in the previous section support this

conclusion, since the comparison of subcorpora of at least one million words does not seem to pose serious problems, and the problems are less important for the largest of the two normalized thresholds evaluated.

The simplest and most correct way to avoid unfairness in the comparison is to analyse corpora of sizes as similar as possible, as, for instance, in Cortes (2008) and Xiao (2011). If the corpora are of relatively different sizes, the analyses carried out in this study would indicate that these size differences can have a significant impact on the number of bundles identified. It may therefore be desirable to reduce the size of the larger corpora by eliminating the least relevant documents for the study purposes. In the past, this type of reduction could have been seen as useless because of the relatively generalized confidence in the normalization procedure of the frequency thresholds. The present study suggests that the disadvantages of comparing corpora of different sizes may outweigh the impression of objectivity given by analysing off-the-shelf corpora. If it is difficult to identify less relevant documents, using either of the two sampling procedures presented above may be recommended. Repeating the random sampling of texts or extracts several times will limit the impact of the specific sample drawn. In this case, the final list of lexical bundles should take account of as many bundles as the average number of bundles identified in the different random draws, picked up by starting with the bundle that is on average the most frequent in the samples and going down the list.

Other approaches are possible, such as extracting lexical bundles from a very large reference corpus or from previous published studies, and then searching for them in the corpora on which the study is based, as undertaken, for instance, by Cortes (2004). However, this approach does not provide an answer to some of the questions at the heart of the classical lexical bundle procedure. In particular, it is impossible to compare the most frequent bundles (according to a given normalized frequency threshold) in each corpus or to list the bundles of a corpus that is not in the reference corpus (Cortes, 2004). Another solution is to extract the bundles from the combination of all the corpora studied and to search for them in each of the specific corpora (Nesi & Basturkmen, 2006). The major difficulty with this approach is that, if some of these corpora are significantly smaller than others, the bundles they contain are less likely to be identified on the basis of the aggregate corpus. Yet another potential solution consists of using a dynamic threshold for frequency as in the study of Chen and Baker (2016), who considered that four-word bundles must occur three times in a 26,000-word corpus (114 occurrences per million words) and four times in an 88,000-word corpus (45 occurrences per million words). For the same purpose, Biber and Barbieri (2007) used a stricter normalized threshold and considered a dispersion threshold for corpora of 50,000 words or less. The major issue with this solution is the arbitrariness of the chosen thresholds.

To conclude, it is worth restating that the performed analyses showed that comparing corpora of different sizes does not necessarily lead to erroneous conclusions. For example, even when comparing a 100,000-word corpus of academic prose to a four-million-word corpus of conversation, it is the latter that contains the most lexical bundles. If the disparities in the number of bundles are sufficiently large between the corpora, a difference in size should not reverse the conclusion. Similarly, the identification of a larger number of bundles in the biggest of two corpora cannot be challenged by any size disparity. Nevertheless, in both cases, the size effect will distort the difference. More generally, this study tries to highlight, like others before it (e.g., Bestgen, 2014, 2017; Evert, 2017; Gries, 2015; McEnery & Hardie, 2012; Wallis, 2013), how much the point of view of quantitative linguistics deserves to be taken into account when new methodologies are developed in corpus linguistics.

Acknowledgement
This work was supported by the Fonds de la Recherche Scientifique 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/UCLouvain) and the Consortium des Equipements de Calcul Intensif en Fédération Wallonie Bruxelles (CECI) funded by the F.R.S-FNRS.

Supplemental material

Supplemental material is provided at the end of the manuscript

REFERENCES

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

Allan, R. (2016). Lexical bundles in graded readers: To what extent does language restriction affect lexical patterning? *System, 59*, 61–72.

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.

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht: Kluwer.

Bannard, C., & Lieven. E. (2009). Repetition and reuse in child language learning. In R. Corrigan, E. A. Moravcsik, H. Ouali & K. M. Wheatley (Eds), *Formulaic language: Volume 2 Acquisition, loss, psychological reality, and functional explanations* (pp. 299–321), Amsterdam: John Benjamins.

Baroni, M. (2008). Distributions in text. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 803–822). Berlin: Mouton de Gruyter.

Berglund, Y. (2000). Utilising Present-day English corpora: A case study concerning expressions of future. *ICAME Journal, 24*, 25–63.

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

Bestgen, Y. (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.

Bestgen, Y. (2018). Evaluating the frequency threshold for selecting lexical bundles by means of an extension of the Fisher's exact test. *Corpora, 13*, 205–228.

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

Biber, D., Conrad, S., & Cortes. V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*, 371–405.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge, UK: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English*. London: Longman.

Bychkovska, T., & Lee, J. J. (2017). At the same time: Lexical bundles in L1 and L2 university student argumentative writing. *Journal of English for Academic Purposes, 30*, 38–52.

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

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

Cochran, W. G. (1977). *Sampling techniques* (3th ed.). New York, NY: Wiley.

Cortes, V. (2002). *Lexical bundles in published and student academic writing in history and biology* (Unpuhlished doctoral dissertation). Northern Arizona University, Flagstaff.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: examples from history and biology. *English for Specific Purposes, 23*, 397–423.
Cortes, V. (2008). A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora, 3*, 43–57.
Cortes, V. (2015). Situating lexical bundles in the formulaic language spectrum: Origins and functional analysis developments. In V. Cortes & E. Csomay (Eds), *Corpus-based research in applied linguistic* (pp. 197–216). Amsterdam: John Benjamins.
Dutra, D. P., Orfano, B. M., & Berber Sardinha, T. (2014). Stance bundles in learner corpora. In S. M. Aluisio & S. E. O. Tagnin (Eds), *New language technologies and linguistic research: A two-way road* (pp. 2–15). Newcastle: Cambridge Scholars Publishing.
Evert, S. (2017, July). Making sense of multivariate analyses of linguistic variation. Poster presented at Corpus Linguistics 2017, University of Birmingham.
Granger S., Dagneaux, E. & Meunier, F. (2002). *International corpus of learner English*. Handbook and CD-ROM. Louvain-la-Neuve: Presses universitaires de Louvain.
Gray, B. (2016). Lexical Bundles. In P. Baker & J. Egbert (Eds.), *Triangulating methodological approaches in corpus linguistic research* (pp. 33–55). New York, NY: Routledge.
Gries, S. T. (2015). Some current quantitative problems in corpus linguistics and a sketch of some solutions. *Language and Linguistics, 16*, 93–117. 2015
Ha, L. Q., Sicilia-Garcia, E. I., Ming, J., & Smith, F. J. (2002). Extension of Zipf's Law to Words and Phrases. *Proceedings of COLING 2002* (pp. 315–320). Taipei, Taiwan.
Hofland, K., Lindebjerg, A., &Thunestvedt, J. (1999). *ICAME collection of English language corpora.* [CD-ROM]. Bergen: The HIT Centre.
Huang, K. (2015). More does not mean better: frequency and accuracy analysis of lexical bundles in Chinese EFL learners' essay writing. *System, 53*, 13–23.
Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*, 4–21.
Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics, 32*, 150–169.
Juknevičienė, R. (2009). Lexical bundles in learner language: Lithuanian learners vs. native speakers. *KaLBOTYRa, 61*, 61–72.
Lee, C. (2013). Using lexical bundle analysis as discovery tool for corpus-based translation research. *Perspectives, 21*, 378–395.
Manning, C. D., & Schütze, H. (2009). *Foundations of statistical natural language Processing*. Cambridge, MA: MIT Press.
McEnery, T., & Gabrielatos, C. (2006). English Corpus Linguistics. In B. Aarts & A. McMahon (Eds), *The handbook of English linguistics* (pp. 33–71). Oxford: Blackwell.
McEnery, T. & Hardie, A. (2012). *Corpus Linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.
McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Taylor & Francis.
Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics, 11*, 283–304.
Oakey, D. (2009). Fixed collocational patterns in isolexical and isotextual versions of a corpus. In P. Baker (Ed.), *Contemporary corpus linguistic*s (pp. 140–158). London: Continuum.
Pan, F., Reppen, R., & Biber, D. (2016). Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in Telecommunications research journals. *Journal of English for Academic Purposes, 21*, 60–71.
Pickering, L., & Byrd, P. (2008). An investigation of relationships between spoken and written academic English:Lexical bundles in the AWL and in MICASE. In D. Belcher &

A. Hirvela (Eds.), *The oral/literate connection: Perspectives on L2 speaking, writing and other media interactions* (pp. 110–132). Ann-Arbor, MI: University of Michigan Press.
Reppen, R. (2009). Exploring L1 and L2 Writing Development through Collocations: A Corpus-based Look. In A. Barfield & H. Gyllstad (Eds), *Researching collocations in another language* (pp. 49–59). London: Palgrave Macmillan.
Schnur, E. (2014). Phraseological signaling of discourse organization in academic lectures: A comparison of lexical bundles in authentic lectures and EAP listening materials. *Yearbook of Phraseology, 5*, 95–122.
Scott, M., & Tribble, C. (2006). *Textual patterns. Key words and corpus analysis in language education*. Amsterdam: John Benjamins.
Thompson, S. K. (2012). *Sampling*. New York, NY: Wiley.
Wallis, S. (2013). z-squared: the origin and application of χ^2. *Journal of Quantitative Linguistics*, *20*, 350–378.
Xiao, R. (2011). Word clusters and reformulation markers in Chinese and English. Implications for translation universal hypotheses. *Languages in Contrast, 11*, 145–171.
Zanette, D., & Montemurro, M. (2005). Dynamics of text generation with realistic Zipf's distribution. J*ournal of Quantitative Linguistics, 12*, 29–40
Zeldes, A. (2013). *Productivity in argument selection: From morphology to syntax*. Berlin: Walter de Gruyter.
Zipf, G. K. 1935/1965. *The psycho-biology of language*. Cambridge, MA: MIT Press.

Supplemental online material for
Comparing lexical bundles across corpora of different sizes: the Zipfian problem

The following tables give the average numbers of lexical bundles identified according to the frequency and dispersion thresholds in the 100,000 samples for each subcorpus size, as well as the number of bundles identified in the full corpora. They also give, for each subcorpus size, the ratio in percentage terms between the number of lexical bundles identified in the subcorpus and the number in the full corpus. They were obtained using the second sampling procedure, which is based on the random selection of an extract from each document in the original corpus.

Table 5: Number and percentage (second line) of four-word lexical bundles in the FLOB corpus by means of the second sampling procedure.

Disp.	Norm. freq.	1/10 100,000	1/5 200,000	1/4 250,000	1/2 500,000	Full
3	20		109	90	63	53
			206	*170*	*119*	*100*
	40	29	16	15	11	9
		322	*178*	*167*	*122*	*100*
5	20			79	63	53
				149	*119*	*100*
	40		16	15	11	9
			178	*167*	*122*	*100*

Table 6: Number and percentage (second line) of three-word lexical bundles in the FROWN corpus by means of the second sampling procedure.

Disp.	Norm. freq.	1/10 100,000	1/5 200,000	1/4 250,000	1/2 500,000	Full
3	20		975	845	642	570
			171	*148*	*113*	*100*
	40	292	189	173	146	129
		226	*147*	*134*	*113*	*100*
5	20			781	639	569
				137	*112*	*100*
	40		188	173	146	129
			146	*134*	*113*	*100*

Table 7: Number and percentage (second line) of four-word lexical bundles in the FROWN corpus by means of the second sampling procedure.

Disp.	Norm. freq.	1/10 100,000	1/5 200,000	1/4 250,000	1/2 500,000	Full
3	20		76	62	45	38
			200	*163*	*118*	*100*
	40	20	12	11	9	9
		222	*133*	*122*	*100*	*100*
5	20			57	45	38
				150	*118*	*100*
	40		12	11	9	9
			133	*122*	*100*	*100*

Table 8: Number and percentage (second line) of three-word lexical bundles in the ICLE corpus by means of the second sampling procedure.

Disp.	Norm. freq.	1/24 100770	1/12 201540	1/10 242849	1/5 483697	1/3 806162	1/2 1209243	Full
3	20		2621	2238	1854	1842	1784	1723
			152	*130*	*108*	*107*	*104*	*100*
	40	977	729	645	630	589	575	555
		176	*131*	*116*	*114*	*106*	*104*	*100*
5	20			2176	1854	1842	1784	1723
				126	*108*	*107*	*104*	*100*
	40		729	645	630	589	575	555
			131	*116*	*114*	*106*	*104*	*100*

Table 9: Number and percentage (second line) of four-word lexical bundles in the ICLE corpus by means of the second sampling procedure.

Disp.	Norm. freq.	1/24 100770	1/12 201540	1/10 242849	1/5 483697	1/3 806162	1/2 1209243	Full
3	20		502	405	305	296	280	266
			189	*152*	*115*	*111*	*105*	*100*
	40	151	98	84	81	74	72	69
		219	*142*	*122*	*117*	*107*	*104*	*100*
5	20			392	305	296	280	266
				147	*115*	*111*	*105*	*100*
	40		98	84	81	74	72	69
			142	*122*	*117*	*107*	*104*	*100*