Know thy corpus!
Exploring frequency distributions in large corpora

Serge Sharoff

Abstract One of the first widely cited papers by Adam Kilgarriff was "Putting Frequencies in the Dictionary" (1997) with his BNC frequency list being widely used by researchers working in lexicography, computational linguistics and language education. However, word frequency lists coming from different corpora differ considerably in spite of relatively small changes in their composition, because some words can become too frequent in a relatively small number of texts specific to a corpus. The present chapter aims at challenging the current practice based on unreliable frequency counts. In the proposed framework, the frequency lists are produced by using document-level measures to filter out frequency bursts via robust statistics. The method which this study found to be most useful is based on *huberM* and S_n estimators of expected values and the percentile bootstrap for the confidence intervals. This helps in describing the frequency distributions from different corpora, in making more reliable estimates of how common the words and their constructions are, and in inferring the significant differences in the lexicon of different text collections, e.g., detecting problems in a given corpus, how a Web crawl is different from the BNC, etc.

Key words: Frequency lists, robust statistics, core vocabulary

Serge Sharoff
Centre for Translation Studies, University of Leeds, e-mail: s.sharoff@leeds.ac.uk

Nothing in Nature is random... A thing appears random only through the incompleteness of our knowledge. Spinoza, Ethics I

Language is never, ever, ever, random [17]

1 Introduction

Frequency of linguistic phenomena, e.g., how common a word or construction is overall or is expected to be in a new text, has been of interest to researchers even before the invention of the computers. In the beginning of the 1900s, Andrei Markov investigated the frequencies of n-grams in poetry [11], the first proper frequency dictionary has been produced for German at the end of the 19^{th} century [14], which was followed in the middle of the 20th century by the General Service List for English [29] and frequency studies on the Brown Corpus [21]. Knowledge of word frequencies is important to determine which words to teach at what stage. Adam Kilgarriff's BNC list (1997) has been used as the basis for defining the English language curriculum in Japanese schools.[1] Knowing the core lexicon of a language is also helpful in specifying the limits of a dictionary word list.

In addition to language teaching and lexicographic applications, frequency lists are needed to produce the probability estimates in many NLP applications, such as Machine Translation, Information retrieval, Speech recognition, Text classification. A lot of attention in language modelling has been paid to estimating the frequency of *unseen* n-grams. However, a reliable estimate of how frequent *known* n-grams are involves more than a straightforward count of the occurrences in a corpus, since raw frequency counts are prone to bursts.

Adam Kilgarriff referred to this as a "whelk" problem [16]. If you have a text about whelks, no matter how infrequent this word is in the rest of your corpus, it's likely to be in nearly every sentence in this text. My personal experience of whelks is from an early version of the Russian National Corpus [26], which contained a Russian sequel to Tolkien's The Lord of the Rings. Even though that novel was less than one percent of the whole corpus at the time, the word *hobbit* made it to the first thousand of most frequent Russian words. The whelks of the British National Corpus (BNC) are medical terms, as seen in the following two extracts from the BNC frequency list:[2]

moon, assert, crown, ***gastric***$_{3763}$*, correct, lock, mutual, thoroughly*
planner, evil, cage, ***pylorus***$_{5955}$*, disguise, sunlight, repay*

While texts taken from the Journal of Gastroentorology and Hepatology contribute less than 0.8% (713 thousand tokens) to the BNC, the frequency bursts of words from this domain propel them into the core lexicon, assuming that the top 10,000 words can be definitely considered as the core of a language. No corpus

[1] Personal communication with Adam Kilgarriff.

[2] Here and below the subscripts indicate the rank of the respective word in the frequency lists. The lists have been produced for the lemma+PoS combinations as output by TreeTagger.

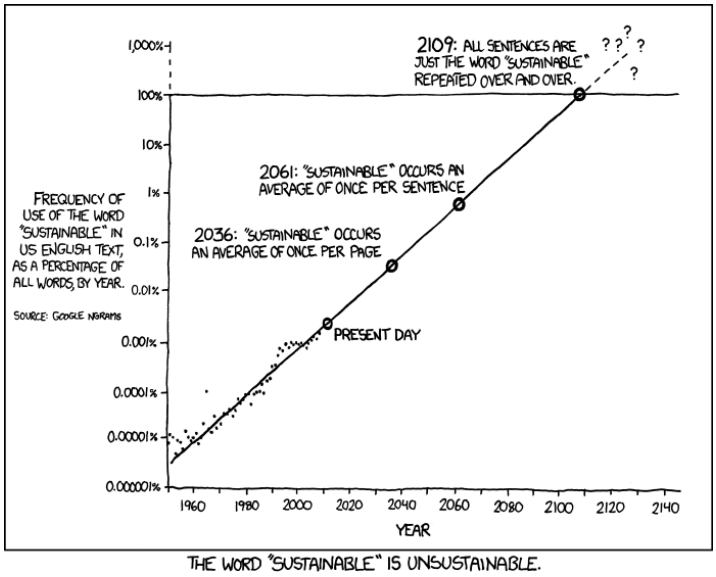

Fig. 1 The word *sustainable* is unsustainable.

is immune to whelks. For example, the frequency of *Texas* in the LDC Gigaword corpus is greater than that of such common words as *long* or *car*.

A related issue is illustrated by an XKCD comic shown in Figure 1.[3] We know that the argument is false, as the nature of language posits reasonable limits to possible frequencies of words. However, in documents which do not contain running text, it is possible to have frequencies well above the expected threshold, for example, in spam webpages or in catalogues, where a word is repeated many times in respective cells. Frequency estimation needs to deal with such cases too.

On the other hand, some words are *less likely* to experience frequency bursts, which puts them in inferior positions in the frequency lists in comparison to those which do. Kilgarriff referred to this as a "banana" problem [18]: even if we do talk about everyday objects we only mention them in passing and do not repeat their names in the same way we use topical words. In the BNC, the word *banana* is in a reasonable position in the BNC list (also because of the *banana skin* metaphor commonly used by journalists), but this is not true for many other everyday objects:
anchor, instrumental, sodium, ***banana***$_{6965}$*, tilt, hunter, armour*
leer, enthrall, sheaf, ***toothbrush***$_{19676}$*, dungeon, stocky, lawsuit*

One way of addressing such problems is by collecting a better corpus [19, 27] or by cross-checking corpus resources across languages [18]). However, each corpus has its own examples of *gastric* or *Moroccan oil* [20], so we need a procedure for

[3] `http://xkcd.com/1007/` XKCD - A webcomic of romance, sarcasm, math, and language. Last accessed 22 March 2019.

discovering such anomalies and for mitigating their impact on the resources produced from this corpus. This chapter investigates the problem from the statistical viewpoint and proposes a language- and corpus-independent procedure which is based on robust statistics.

In the rest of the chapter, the following notation will be used:

c_i	the number of occurrences of a word in a text i
n_i	the size of a text i in tokens
$C = \sum c_i$	the total number of occurrences of a word in a corpus
$R = \|\|c_i > 0\|\|$	the number of texts a word occurs in
$N = \sum n_i$	the number of tokens in a corpus
$T = \|\|n_i\|\|$	the number of texts in a corpus
$p_i = \frac{c_i}{n_i}$	the probability of a word in a text i
$f_i = p_i \times 10^6$	by-text normalised frequency per million words (ipm)
$\mu = \frac{\sum f_i}{T}$	macro-average of normalised by-text frequencies
$\sigma = \sqrt{\frac{\sum (f_i - \mu)^2}{T}}$	standard deviation of normalised by-text frequencies

The problem studied in this chapter concerns reliable frequency estimates together with reliable confidence intervals in the presence of frequency bursts. More specifically, the goal is to determine:

- the expected frequency $E[f]$ for a word on the basis of an existing collection;
- methods for detecting frequency bursts and for mitigating their influence;
- the confidence intervals for $E[f]$ assuming that any new corpus is drawn from the same infinite library [8]

The study will be illustrated by the frequencies of a number of words from the BNC and ukWac. The BNC is a corpus manually constructed in the 1990s to represent the then current state of British English [1], while ukWac is a snapshot of HTML webpages taken in 2005 from the `.uk` domain [4]. Even though the two corpora are spaced diachronically and the BNC texts tend to be more formal in comparison to a snapshot of texts from the Web [27], the two corpora are broadly comparable as they represent a variety of texts in British English, so we can expect that their core lexicons are sufficiently similar. Comparisons will be also made to the corpus of English Wikipedia and the English Gigaword corpus, which primarily consists of newswires [7]. In comparison to ukWac and the BNC, these corpora are more specialised, also not exclusively British, but they are also widely used in NLP studies.

The unit for analysis is the frequency of lemmas taken with their generalised POS tags, e.g., *correct.J* and *correct.V* are separate units for the adjectival and verbal readings of *correct*. The latter one covers the individual word forms *corrects, corrected, correcting*. The lemmas and POS tags for all corpora have been produced by TreeTagger [25]. The frequencies are counted within the boundaries of individual texts, since texts are normally written by a specific author in a specific genre on a specific topic, so they offer natural units of analysis for studying variations of word use. There are some statistical complexities introduced by focusing on whole texts in comparison to splitting corpora into equally sized parts, but this is the preferred form of analysis for this chapter, because texts provide natural boundaries between

Table 1 Corpora in this study

	Ktexts	Mwords	References
BNC	4	100	[1]
Giga-EN	2853	1357	[7]
ukWac	2542	1875	[4]
Wikipedia	2524	1242	Dump from November 2011

Table 2 Raw frequencies in the BNC

	anxiously.R	correct.V	gastric.J	moon.N	moon.V	thoroughly.R	toothbrush.N
Count	603	2053	2057	2065	38	2042	183
Range	338	1038	65	701	32	1100	123
Frequency distributions in ipm per text:							
IPM	$3.6^{\pm 0.6}$	$21.3^{\pm 3.3}$	$3.6^{\pm 2.6}$	$17.4^{\pm 3.8}$	$0.2^{\pm 0.08}$	$18.0^{\pm 2.0}$	$1.8^{\pm 0.7}$
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.000
1st Q	0.00	0.00	0.00	0.00	0.00	0.00	0.000
Med.	0.00	0.00	0.00	0.00	0.00	0.00	0.000
3rd Q	0.00	12.50	0.00	0.00	0.00	18.73	0.000
Max	479.16	3897.79	2957.45	4143.39	79.37	2028.40	1046.025

topics. Some of the documents in the BNC have been produced by merging several texts, e.g., Text K5D, 277,145 words is a collection of articles from an issue of *The Scotsman*, or Text HJ0, 191,524 words is a collection of ESRC grant abstracts. Other corpora used in this study, ukWac, Wikipedia and the English Gigaword corpus are less affected by such artefacts of corpus construction. The corpora are listed in Table 1.

The words chosen for a closer investigation in Table 2 are determined by the close neighbours of the misbehaving word *gastric* in the raw frequency list of the BNC. The task is to investigate how the frequency bursts affect words with the same frequency, but of different POS classes, i.e., nouns, verbs and adverbs. However, in the case of *correct* and *moon* there are two POS classes possible for the same headword, which also adds to the possibility of studying the frequency bursts of words in different frequency classes.

2 Frequency counts

Let's start with the raw frequency counts in the BNC for several words (the first line in Table 2). Some of them have roughly the same count frequency (*correct.V, gastric.J, moon.N, thoroughly.R*), while they differ in their range, the number of texts they occur in. A standard way to estimate the variation of their frequencies is by using the 95% confidence interval for the binomial distribution [2, p. 50ff]:[4]

[4] The superscript with $^{\pm}$ refers to the confidence intervals.

$correct.V, gastric.J, moon.N, thoroughly.R = 2050^{\pm 89}$
$anxiously.R = 603^{\pm 49}, toothbrush.N = 183^{\pm 27}$

This does not look adequate since we can expect that *gastric* is a specialised word, which in the BNC occurs most often in the texts from the Journal of Gastroentorology and Hepatology, while *correct.V* and *thoroughly* are more evenly distributed across the BNC texts. The upper limits of the confidence interval of *anxiously* and *toothbrush* are well below the lower limit of *gastric*.

Another way of estimating the frequencies is by taking into account the text boundaries, measuring the probabilities (or ipm frequencies for presentational purposes) in each individual text and using its mean. The standard error of the mean can be used to estimate the 95% confidence interval: $\mu \pm 1.96\frac{\sigma}{\sqrt{N}}$. The results are also presented in Table 2.

The words in question do not occur in more than half of the texts in the BNC, so their median per-text frequencies in Table 2 are 0. Since *gastric* occurs in a small number of texts, its mean IPM frequency becomes smaller than the means of *correct.V* or *moon.N*. However, it is still quite high, so that the upper limit of its 95% confidence interval (6.2 ipm) is above the upper limit of the respective confidence intervals of such common words as *collaborate, fury* or *downward*, which have higher μ, but lower σ.

The reason for this effect comes from the discrepancy between the distribution of word frequencies and the basic assumptions of statistical testing:

- Independence of observations
  One occurrence of a word is independent from another occurrence
- Normal (Gaussian) distribution
  Frequencies vary following a bell shape as in Figure 2
- Homoscedasticity, i.e., equal variance of data:
  Word frequencies across documents vary in similar ways
- Linearity (for linear models)

The standard confidence intervals can be estimated based on the assumption that for data coming from the normal distribution 95% of the values are within $\mu^{\pm 2\sigma}$, see the upper part of Figure 2.

In reality, the distributions of document frequencies are very far from being normal. Even for a relatively well-behaved word like *correct.V* the distribution of its frequencies peaks at zero (it occurs in just 1038 out of 4054 documents in the BNC) and has a long tail of varying frequencies, see the bottom part of Figure 2, where *emp* stands for 'empirical', the actual density of word frequency values, the left part shows the entire range of densities, while the right part focuses on the non-zero counts. Both the number of zeros and the fat tail of non-zeros in the empirical distribution are much higher than what is predicted by the normal distribution model, while the exponential distribution (suggested as the best fit by the `fitdistrplus` package in R) is only good in predicting the zeros, leaving almost nothing for the non-zero values.

Kenneth Church has shown that splitting a text into two parts ('history' and 'test') leads to much greater probability of seeing a word in the test part once it occurred

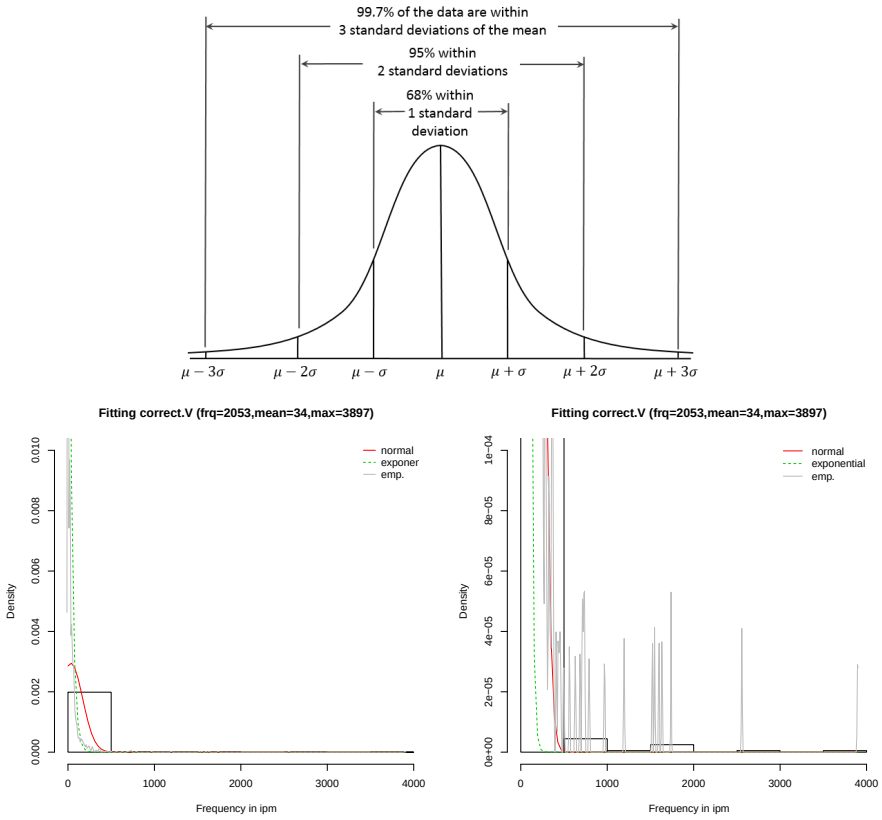

Fig. 2 Approximation of confidence intervals and real data

in the history part [6]. In the end, if the probability of seeing a topical word in a text once is $p(k = 1)$, then the probability of seeing it twice $p(k = 2) \approx p/2$ rather than p^2 as expected in the binomial distribution. Cf. also a more in-depth discussion by Harald Baayen in Chapter 5 of [3].

In his Spanish frequency dictionary Juilland introduced a measure of dispersion of word frequencies, which is essentially based on the standard error of the mean normalised by the mean [13]:

$$D = 1 - \frac{\sigma}{\mu\sqrt{||n_i|| - 1}} \tag{1}$$

This proposal was followed by several other measures aimed at identification and mitigation of such bursts, e.g., Carroll's, Rosengren's, Engvall's measures, see an overview by Stefan Gries in [9]. Because of the inadequacies of these measures, Gries has also suggested his own measure, Deviation of Proportions (DP), which is

defined as:[5]

$$DP = 1 - \frac{\sum |\frac{c_i}{C} - \frac{n_i}{N}|}{2} \quad (2)$$

More burstiness measures have been suggested by Katz with the aim of using them in speech recognition, information retrieval and terminology detection [15]:

$p(k=0) = p_0$	probability of no occurrences of the term in a text
$p(k=1) = p_1$	probability of a single occurrence of the term in a text
$p(k \geq 2) = \sum_{r \geq 2} p_r$	probability of multiple occurrences of the term in a text
$\alpha = 1 - p_0$	proportion of texts containing the term
$\gamma = 1 - \frac{p_1}{1-p_0}$	proportion of 'topical' texts for the term
$B = \frac{\sum r p_r}{\sum p_r} (r \geq 2)$	topical burstiness parameter

α means how likely the word is to occur in a text irrespectively of the number of times it occurs there (a normalised Range); γ means how likely it is to be used 'topically' (i.e. more than once); and B means how intensely, on average, the word is used when it is used topically.

Table 3 Illustration of dispersion measures

	anxiously.J	correct.V	gastric.J	moon.N	moon.V	thoroughly.R	toothbrush.N
St Dev (σ)	18.01	108.57	83.57	123.02	3.13	64.01	23.56
Juilland's D	0.95	0.89	0.60	0.87	0.72	0.93	0.78
Gries' DP_{norm}	0.47	0.38	0.03	0.27	0.01	0.41	0.06
Katz' α	0.08	0.26	0.02	0.17	0.008	0.27	0.03
Katz' γ	0.37	0.39	0.37	0.49	0.13	0.41	0.28
Katz' B	3.12	3.48	84.00	4.98	2.50	3.11	2.71

Word	α	γ	B
not	0.98	0.98	114.31
be	1.00	1.00	885.61
have	0.99	0.99	293.40
pylorus	0.002	0.78	160.86
do	0.98	0.98	136.25

The values of these measures for our words are given in Table 3. All measures indicate that *gastric* is bursty, there is some disagreement to the degree of burstiness of *moon.V* and *toothbrush*, but with the exception of Katz' B the other measures put them uncomfortably close to *gastric*. Katz' B measure does separate *gastric* from other words in this illustration, but at the same time it is not discriminative with respect to fairly common words, as many of them have even higher burstiness index than *gastric*, see an illustration of their values in the bottom part of Table 3. Also Katz' B does not have a predictable range of acceptable values.

α is effectively the range frequency normalised by the corpus size in texts ($\alpha = R/T$), also it is the same as Engvall's measure [9]. This is also the inverse of the IDF (Inverse Document Frequency) measure. By this count *toothbrush* becomes more

[5] This can be followed by normalisation to make sure its value is within $[0,1]$. This formulation also aligns the direction of DP with Juilland's D.

common than *gastric*. However, range-based frequency lists are also limited by the fact that they do not distinguish evidence coming from short and long texts, so that their values can vary radically between otherwise reasonably similar corpora. For example, 748 words in the BNC occur in more than half of documents, with 2430 words occurring in more than 25% of documents. Given that the texts in ukWac are considerably shorter than those in the BNC, only seven words occur in more than half of the ukWac texts, with only 47 words occurring in more than 25% of its texts.

As mentioned above, language modelling pays a lot of attention to smoothing, i.e., estimating the frequency of 'unseen' n-grams, while the frequency of observed n-grams is measured as it is without using information from the document frequencies, only the sentence frequencies are sometimes taken into account [12]. Therefore, LM does not distinguish between the probabilities of *gastric mucosa* vs *thoroughly enjoy*. Using the BNC data, KenLM estimates the log probability of *gastric* as -4.79 vs the log probability of *thoroughly* as -5.18.

Another problem which concerns all of these measures is that we do not have any estimation of what reliable counts are likely to be: we can detect the lack of a well-behaving distribution across a number of documents, but this does not help in detecting the *expected* frequency value. A common practice in frequency dictionaries is to multiply the raw counts by a dispersion measure (by Juilland's D in the dictionaries mentioned above), but this applies a uniform correction measure to the overall count, while the frequency bursts are specific to individual texts. Also this does not provide us with estimates of the confidence intervals.

3 Robust statistics for outliers

3.1 Robust estimators of location and scale

This study estimates reliable word frequency counts by introducing robust statistics, which restricts contributions from outlying observations, such as the Journal of Gastroentorology texts in the BNC, which boost the frequency of *gastric*. It is known that traditional frequency measures, such as the mean (an estimator of location) and standard deviation (an estimator of scale) are not robust to outliers: a single frequency burst can move them out of bounds. Therefore, the field of robust statistic has introduced several robust estimators [24].

A commonly used robust estimator of scale is Median Absolute Deviation (*MAD*):

$$MAD = b \times median\,|x_i - median(x)| \tag{3}$$

i.e., taking the median of the absolute differences from the median of x. Rousseeuw & Croux introduced another scale estimator S_n with more attractive gross-error sensitivity properties in comparison to MAD:

$$S_n = c \times median_i \left(median_j(\left| x_i - x_j \right|) \right) \tag{4}$$

i.e. taking the medians of pairwise differences in word frequencies across texts. The values of the constants $b = 1.48$ for *MAD* and $c = 1.19$ for S_n are used to match the standard deviation value σ for normally distributed data [24].

As for robust estimators of location, the median is commonly used. However, it completely ignores variation of values around the median item, i.e., for skewed distributions it does not reflect the difference between the two tails. Research in robust statistics proposed other robust measures of location, such as Huber's M-estimator, which is based on the idea of taking the values of non-outlying items at their face value and discounting the effect of the items outside this range [30]. The procedure is iterative: it starts with $\mu_0 = Median$ and updates μ_{k+1} by discounting the contribution of the items which satisfy the condition:

$$|x_i - \mu_k| > 1.28 \times MAD \tag{5}$$

One problem in direct application of robust measures to word frequency lists consists in the prevalence of zero frequencies. This leads to the zero values of *median*, *MAD* and *huberM* for word frequency distributions. One way of dealing with this issue is by using robust methods to detect outliers within *non-zero* frequency documents and to use the traditional mean of the discounted frequencies (Winsorisation).

More specifically, *huberM* and S_n are computed for the normalised frequencies from all documents in which a word occurs. After that, the frequencies in these documents are capped by the value of $huberM + 2.24S_n$.[6] In principle the lower limit $huberM - 2.24S_n$ is possible, but only the following nine words in the BNC list *be.V, have,V, not.R, make.V, take.V* have their *huberM* slightly larger than $2.24S_n$, and for each of these words the condition of $f_i < huberM - 2.24S_n$ occurs in a small number of documents. Anyway, since the natural lower bound of word frequency is zero, any deviations below the reliable frequency limit will not have considerable effect on frequency estimation.

The final step in determining robust frequency estimates is to calculate their robust confidence intervals (CIs). Even after Winsorisation of outlying frequencies the distribution is far from normal. Winsorisation also tends to produce two modes: the zero frequency as well as the upper limit, which was used for clipping the outliers. In the end it is difficult to approximate the word frequencies by a theoretical distribution with known CIs.

A robust (even if computationally expensive) procedure for determining the 95% CIs involves "percentile bootstrap" [30]. It starts with random sampling of the observations (with replacements) for B iterations (usually B ranges from 400 to 1000). Then it determines robust location estimates μ_k^* for each iteration (as implemented above to take into account the prevalence of zero frequencies). The procedure finishes by taking the 5% and 95% quantiles of the sequence $\mu_1^*, \ldots, \mu_B^*$ as the limits of the confidence interval.

[6] Wilcox argues in favour of having $\sqrt{\chi^2_{0.975,1}} \approx 2.24$, i.e. the square root of the 0.975 quantile of a chi-squared distribution with one degree of freedom, as a threshold for detecting the outliers [30]. However, the thresholds can be set in other ways depending on the application.

Table 4 Robust frequency estimates

BNC	anxiously.R	correct.V	gastric.J	moon.N	moon.V	thoroughly	toothbrush
Raw rank	9157	3769	3763	3746	47820	3783	19676
New rank	7341	2641	16727	3275	27278	2647	12017
Raw IPM	$3.6^{\pm 0.6}$	$21^{\pm 3.3}$	$3.6^{\pm 2.6}$	$17^{\pm 3.8}$	$0.20^{\pm 0.08}$	$18^{\pm 2.0}$	$1.8^{\pm 0.7}$
Estimate	$2.8^{\pm 0.3}$	$13^{\pm 0.9}$	$0.7^{\pm 0.2}$	$10^{\pm 0.8}$	$0.17^{\pm 0.06}$	$13^{\pm 0.8}$	$1.2^{\pm 0.2}$
ukWac							
Raw IPM	$0.6^{\pm 0.05}$	$16^{\pm 0.3}$	$1.7^{\pm 0.15}$	$13^{\pm 0.4}$	$0.08^{\pm 0.02}$	$20^{\pm 0.3}$	$1.3^{\pm 0.13}$
Estimate	$0.5^{\pm 0.03}$	$12^{\pm 0.2}$	$1.1^{\pm 0.07}$	$9^{\pm 0.2}$	$0.05^{\pm 0.01}$	$16^{\pm 0.2}$	$0.8^{\pm 0.05}$
Wikipedia							
Raw IPM	$0.19^{\pm 0.03}$	$7.8^{\pm 0.2}$	$1.4^{\pm 0.15}$	$14^{\pm 0.4}$	$0.07^{\pm 0.02}$	$3.5^{\pm 0.1}$	$0.36^{\pm 0.06}$
Estimate	$0.16^{\pm 0.09}$	$6.1^{\pm 0.2}$	$1.0^{\pm 0.08}$	$10^{\pm 0.2}$	$0.05^{\pm 0.01}$	$2.9^{\pm 0.1}$	$0.25^{\pm 0.03}$
Giga-EN							
Raw IPM	$1.3^{\pm 0.07}$	$61^{\pm 1.6}$	$0.3^{\pm 0.04}$	$13^{\pm 0.3}$	$0.09^{\pm 0.02}$	$6.4^{\pm 0.15}$	$0.7^{\pm 0.06}$
Estimate	$1.1^{\pm 0.05}$	$22^{\pm 0.3}$	$0.2^{\pm 0.02}$	$11^{\pm 0.2}$	$0.08^{\pm 0.01}$	$5.5^{\pm 0.1}$	$0.5^{\pm 0.03}$

The resulting frequencies for our focus words are shown in Table 4. Since Winsorisation only caps frequency bursts in individual documents, the resulting robust counts are always lower than their raw versions. In the end, for words like *correct.V, gastric.J, moon.N*, the estimated counts are closer across the BNC and ukWac than their raw counts, which indicates that we are more successful in predicting their frequency in general language. However, for *anxiously* and *moon.V* the ukWac counts are consistently lower than in the BNC, which can be explained by differences in their genre composition: there is much less fiction in ukWac. The counts for *moon.V* remain stable in all other corpora without fiction. There is a surprising burst in the frequency of *correct.V* in the Giga-EN corpus, which is mostly caused by the presence of '*CORRECTED:*' lines (sometimes several of them) in very short articles.

In the case of the "gastric" words, they have been demoted in the Winsorised frequency list for the BNC:
vindictive, bitumen, cleave, ***gastric***$_{16727}$*, minke, railwayman*
verger, rigorist, ***pylorus***$_{37868}$*, moonbeam, correlative, gallivant*

For ukWac the effect of Winsorisation was also measured by using the log-likelihood (LL) score [23], i.e., by comparing the original raw frequency counts vs the Winsorised frequency list from the same corpus. The words most different between the two versions in ukWac according to the LL score are:
insurance, loan, puzzle, HMS, wedding, RAF, course, campus, God, mortgage, dog, child, Select, pension, credit, Sale, Estate, nigritude

The word *nigritude* is a remainder of a Search Engine Optimisation contest run in 2004, in which the aim was to win by having a contestant's page at the top of Google searches for a non-sensical phrase *nigritude ultramarine*. Many of these pages remained in 2005 when ukWac was crawled, while they do not contain meaningful text and they should not contribute to the frequency count for *nigritude*. Some of the demoted words are related to insufficient webpage cleaning (e.g., *Select*), some to commercial promotion (*loan, Sale*), some to text extracted from tabular formats (*course, HMS*, as a part of a unit name).

Unlike overall frequency correction measures such as Juilland's D, Winsorisation reduces frequencies only in selected documents, those, which violate the conditions of being a representative sample from a random library. For example, a list of admission criteria to local schools can make frequent references to school names and children. A small number of pages of this sort in ukWac can lead to over-estimation of the frequencies of such words if no correction is done. However, such words do not present a problem in the vast majority of their use in other pages, so the frequency drop caused by Winsorisation is less significant in proportion to all other uses, but still it is important for the LL measure. For example, the raw count of the word *school* dropped from 978,962 to 772,913, while for *mortgage* the drop was from 103,805 to 62,749 on the ukWac data, reflecting the possibility of its use in spam pages.

Since the frequency estimates and the corresponding ranks of bursty words go down, words less prone to bursts can increase their ranks. This also helps in determining the core lexicon, even though some words from what can be considered to be a general lexicon remain underrepresented because of corpus composition, for example, *toothbrush*. In the version of the BNC list after robust Winsorisation, the words *banana* and *toothbrush* get the following neighbours:
overview, floating, group, ***banana***$_{5148}$*, wounded, catch, philosopher*
balancing, unhappily, suicidal, ***toothbrush***$_{12017}$*, cuisine, retaliate, take-off*

4 Conclusions

The main message of this study is that frequency estimation for known words cannot rely on raw counts. It is not safe to estimate the probability of seeing a word as $p = \frac{C}{N}$. Otherwise, it is easy to infer that $p(gastric) = p(correct)$ in British English. Even the mean over all normalised document frequencies provides a better estimate of p. However, it is even better to reduce frequency bursts by Winsorising the document frequencies. The mechanism which is based on *huberM* and S_n is effective and computationally efficient. It requires constructing a vector of document frequencies for each word in a corpus, which is not prohibitively expensive computationally, even for a corpus of a few million documents (such as ukWac). The frequency estimation implementation used in this chapter is based on the `robustbase` package in R [22] and is available under the GPL license.[7] The message about using the Robust Confidence Intervals is a bit less clear. They definitely reflect the range of values expected in a given text collection. However, the text collections themselves differ too much to make the confidence intervals easily interpretable across corpora. It seems that the metaphor of sampling from the same infinite library does not work with real corpora, because sampling is done from fairly different parts of that library.

The amount of Winsorisation applied to a word can be an indicator that a collection is biased, see the *nigritude* example. However, it is important to note that some

[7] `https://github.com/ssharoff/robust` Last accessed 25 March 2019.

words, such as *gastric* or *moon.N* (or *hobbit, Noriega, whelk* from other studies) are inherently bursty, while others, like *anxiously* or *toothbrush*, are not. This can be linked to the Hallidayan notion of lexical cohesion [10]. A text is cohesive partly because of the lexical chains, in which the words are repeated, expressed via synonyms, abbreviations or pronouns. The bursty words contribute to lexical cohesion more than their non-bursty counterparts. They remain vital for any text processing tasks. They become a nuisance only when their repeated use is converted from tokens to types for frequency estimation purposes.

The current study only addressed the frequency bursts of selected examples. Another study is needed for the overall assessment of the ways to mitigate the frequency bursts, possibly via extrinsic evaluation, for example, in the context of language modelling. Further experiments should include expanding the mechanism to n-grams, which also exhibit frequency bursts, as well as to other counts, such as the frequency of POS tags or the distribution of domains and genres in crawled text collections [28]. It is also interesting to perform extrinsic evaluation concerning the influence of the robust frequency estimates on the performance of such applications as Sentiment Analysis or Machine Translation. Another practical application of Winsorising the frequency bursts is related to word embeddings. Many neural models need to limit their vocabulary to a certain number of words, so that the search space remains manageable [5]. Instead of limiting the word list by raw counts, a robust frequency estimation procedure will produce a cleaner list which reflects the core lexicon for the task.

References

1. Aston, G., Burnard, L.: The BNC Handbook: Exploring the British National Corpus with SARA. Edinburgh University Press, Edinburgh (1998)
2. Baayen, H.: Analyzing linguistic data. Cambridge University Press, Cambridge (2008)
3. Baayen, R.H.: Word Frequency Distributions. Kluwer Academic Publishers, Dordrecht (2001)
4. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation **43**(3), 209–226 (2009)
5. Chen, W., Grangier, D., Auli, M.: Strategies for training large vocabulary neural language models. In: Proc 54^{th} ACL, pp. 1975–1985. Berlin, Germany (2016). URL `http://www.aclweb.org/anthology/P16-1186`
6. Church, K.: Empirical estimates of adaptation: the chance of two Noriegas is closer to $p/2$ than p^2. In: Proc COLING, pp. 180–186. Saarbrücken, Germany (2000)
7. Cieri, C., Liberman, M.: Language resources creation and distribution at the Linguistic Data Consortium. In: Proc LREC, pp. 1327–1333 (2002). Las Palmas, Spain
8. Evert, S.: How random is a corpus? The library metaphor. Zeitschrift für Anglistik und Amerikanistik **54**(2), 177–190 (2006)
9. Gries, S.T.: Dispersions and adjusted frequencies in corpora. International Journal of Corpus Linguistics **13**(4), 403–437 (2008)
10. Halliday, M.A.K., Matthiessen, C.M.I.M.: Introduction to Functional Grammar. Arnold, London (2004)
11. Hayes, B., et al.: First links in the Markov chain. American Scientist **101**(2), 92 (2013)
12. Heafield, K., Pouzyrevsky, I., Clark, J.H., Koehn, P.: Scalable modified Kneser-Ney language model estimation. In: Proc 51st ACL. Sofia (2013)

13. Juilland, A.: Frequency dictionary of Spanish words. Mouton (1964)
14. Kaeding, F.W.: Häufigkeitswörterbuch der deutschen Sprache: Festgestellt durch einen Arbeitsausschuss der Deutschen Stenographiesysteme. Berlin (1898)
15. Katz, S.M.: Distribution of content words and phrases in text and language modelling. Natural Language Engineering **2**, 15–59 (1996)
16. Kilgarriff, A.: Putting frequencies in the dictionary. International Journal of Lexicography **10**(2), 135–155 (1997)
17. Kilgarriff, A.: Language is never, ever, ever, random. Corpus Linguistics and Linguistic Theory **1**(2), 263–276 (2005)
18. Kilgarriff, A., Charalabopoulou, F., Gavrilidou, M., Johannessen, J.B., Khalil, S., Kokkinakis, S.J., Lew, R., Sharoff, S., Vadlapudi, R., Volodina, E.: Corpus-based vocabulary lists for language learners for nine languages. Language Resources and Evaluation **48**(1), 121–163 (2014)
19. Kilgarriff, A., Reddy, S., Pomikálek, J., PVS, A.: A corpus factory for many languages. In: Proc LREC. Valletta, Malta (2010)
20. Kilgarriff, A., Suchomel, V.: Web spam. In: Proc Web as Corpus workshop (WAC8) at Corpus Linguistics Conference. Lancaster (2013)
21. Kučera, H., Francis, W.N.: Computational analysis of present-day American English. Brown University Press, Providence (1967)
22. R Development Core Team: R: A Language and Environment for Statistical Computing. Vienna, Austria (2011). URL `http://www.R-project.org`
23. Rayson, P., Garside, R.: Comparing corpora using frequency profiling. In: Proc Comparing Corpora Workshop at ACL 2000, pp. 1–6. Hong Kong (2000)
24. Rousseeuw, P.J., Croux, C.: Alternatives to the median absolute deviation. Journal of the American Statistical Association **88**(424), 1273–1283 (1993)
25. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proc International Conference on New Methods in Language Processing. Manchester (1994)
26. Sharoff, S.: Methods and tools for development of the Russian Reference Corpus. In: D. Archer, A. Wilson, P. Rayson (eds.) Corpus Linguistics Around the World, pp. 167–180. Rodopi, Amsterdam (2005)
27. Sharoff, S.: Open-source corpora: using the net to fish for linguistic data. International Journal of Corpus Linguistics **11**(4), 435–462 (2006)
28. Sharoff, S.: Functional text dimensions for the annotation of Web corpora. Corpora **13**(1), 65–95 (2018)
29. West, M.: A General Service List of English Words. Longman, Green and Co., London (1953)
30. Wilcox, R.R.: Introduction to robust estimation and hypothesis testing. Academic Press (2012)