manuscript, to appear in *ZAA* 54.2 (2006): 177-190 ©

STEFAN EVERT

How Random is a Corpus? The Library Metaphor

Abstract: There is a stark contrast between the random sample model underlying the statistical analysis of corpus frequency data and our intuitive knowledge that sentences are more than random bags of words. The 'library metaphor' illustrates how randomness results from the selection of a corpus as the basis for a linguistic study. At the same time it reveals two reasons why corpus data do not fully meet the assumptions of the random sample model. Finally, practicable methods for identifying and quantifying non-randomness are introduced and demonstrated on the example of passive verb forms.

1. Introduction

Any quantitative study of corpus data requires a statistical analysis of the observed frequencies in order to generalise from the finite sample at hand to a language as a whole. A wide range of statistical methods are available for this purpose, including frequency estimates for lexical items, phrases and syntactic constructions (Kucera and Francis 1967), frequency comparisons (Kilgarriff 2001), association measures for word co-occurrences (Evert 2004), models of language change in syntax and the lexicon (Zuraw 2003), models of vocabulary richness and vocabulary growth (Baayen 2001, 2003), as well as factorial analysis of style and register variation (Biber 1995). All these methods – like most of statistics – are based on a random sample model, which assumes that the observed data (i.e. the corpus) were selected randomly from the language or sublanguage of interest.

Obviously, the random sample model is very unrealistic when applied to natural language. Taken at face value, it seems to imply that sentences are just random 'bags of words'. This simplistic view of language can easily be reduced to absurdity with arguments like the following: since the definite article *the* accounts for roughly one in 17 words in English, the ungrammatical sequence[1] *the the* should occur about once in every 300 words ($1/17 \times 1/17 = 1/289$; cf. Baayen 2001, 163). By contrast, our intuition tells us that hardly anything is left to chance in our choice of words and the way they are combined into sentences. Words are chosen to convey a specific meaning or intention, and their arrangement follows the intricate rules of syntax. Therefore, we may ask: what has randomness to do with linguistics in the first place? And if there is nothing random about language, why should we apply statistical methods (based on the random sample model) at all?

The explanation for this apparent paradox lies in a widespread misapprehension: the source of randomness is not to be found in language production (which would make it an intrinsic property of the utterances themselves), but rather in the choice of a corpus as the basis for a linguistic study. This can be illustrated with the 'library metaphor': imagine a gigantic library that represents the entirety of a language or sublanguage as the object of study. Each book in this library corresponds to a fragment of the language – some large, some small – that could be used as a linguistic corpus. Selecting or compiling a corpus, thus, amounts to picking a

[1] Indeed, the grammar checker of Microsoft Word suggests to "delete the duplicated word".

book at random from one of the shelves. In this way, randomness enters quantitative corpus studies, even if it is not inherent in the object of study itself, viz. the language under investigation.

Section 2 of this paper explores the role of statistical inference in corpus linguistics and identifies the quantities for which statistical estimates can be obtained. Section 3 introduces the library metaphor as a plausible explanation for randomness in corpus data and explains its connection with the random sample model. Section 4 discusses the possible sources of non-randomness and their effects on corpus frequency data. Central notions such as 'representativeness', 'balanced corpora' and 'term clustering' are introduced and linked to the imagery of the library metaphor. Section 5 presents two practicable methods for an empirical validation of the randomness assumption. The focus is here on methodological aspects rather than the presentation of experimental data (some data from the Brown corpus are used for illustration, though). Section 6 summarises the contributions of this paper and suggests avenues of future research.

2. The role of statistical inference in corpus linguistics

First of all, it is important to understand the role that statistical methods play in a corpus-linguistic study. The true goal, of course, is not the estimation of statistical parameters, but to learn something about a linguistic phenomenon, e.g. whether a given word is a technical term, which words or constructions are most characteristic of British English, at what point in time a syntactic construction was established, whether a word combination is compositional or semantically opaque, whether a text was written by a particular author (e.g. Shakespeare), etc. What all these research questions have in common is that they represent an *intensional* view of language, focusing on the competence of human speakers (or the properties of a language as a formal system). Consequently, the answers to these questions are not directly observable from corpus data, although they are bound to be reflected in the speakers' output in some way.

A quantitative, corpus-based study depends on something that can be observed and counted. Therefore, it has to adopt an *extensional* view of language as an infinite body of text, comprising all the utterances that have ever been made or will ever be made by the relevant group of speakers (under suitable conditions). The research question then has to be rephrased in terms of the frequency of use of an observable phenomenon.[2] For instance, a corpus cannot tell us whether a particular genre is more formal than another. It can only show that texts from one genre contain a higher proportion of passives, nominalisations, etc. than texts from another genre. It is up to the linguist to draw meaningful conclusions from these observable quantities.

Formally, the extensional definition required for a quantitative analysis consists of a unit of measurement (typically words, phrases or sentences) and an observable phenomenon (such as instances of passive verb forms). The quantity of interest is the relative frequency of this phenomenon, i.e. its number of occurrences per thousand or million tokens of the chosen unit of measurement (in the text produced by the relevant group of speakers under suitable conditions). Continuing the example of passives, we could define the unit of measurement as word tokens and the observable phenomenon as instances of passive verb forms.[3] The relative

---

2 Frequency of use is to be understood in the widest sense here, often involving the comparison or correlation of relative frequencies in different sublanguages.

3 For the purposes of this paper, we approximate passive verb forms by the following pattern: an inflected form of the verb *be* (*be*, *am*, *are*, *is*, *was*, *were*, *been*, *being*, *'m*, *'re*, *'s*), followed by optional adverbs and a past participle.

frequency of passives in written English is then found to be approx. 12 occurrences per 1,000 words. When the unit of measurement is defined as verbal groups (which occur with a frequency of 140 per 1,000 words), on the other hand, the relative frequency is 86 passives per 1,000 verbal groups.[4]

The goal of a statistical analysis is to make accurate statements about such relative frequencies, i.e. to generalise from the number of occurrences observed in a finite text sample (in other words, a corpus) to the relative frequency of the same phenomenon in the entire output of the relevant group of speakers, which encompasses a (theoretically) infinite amount of text. In doing so, the statistical procedures correct for the randomness that is inherent in any finite sample and that causes the observed frequencies to vary from one sample to the next. Once this generalisation has been made, the extensional quantities can be used by a linguistic researcher to draw conclusions about the intensional phenomenon which is the true object of study. For instance, the fact that scientific texts contain more passives than non-scientific discourse (17.3 per 1,000 words, or 143.2 per 1,000 verbal groups, based on the 'natural and pure sciences' domain in the BNC) might be taken as evidence that scientific writing is more formal than other genres (of course, such a conclusion would only be valid in combination with additional independent evidence).

Statistical methods, however sophisticated they may be, only produce numbers and can never answer linguistic questions directly. They must be accompanied by a linguistic interpretation that invests the numbers with meaning.

3. Random sampling and the library metaphor

While it is obvious that classical applications of statistical methods (such as an opinion poll or spot checks of manufactured goods) operate on random samples, our intuition about language tells us that very little is left to chance in the output of a speaker. Therefore, we may ask: is there any inherent randomness in a corpus? And if there is no such randomness, why should statistical methods be applied?

The source of randomness in a corpus study can be illustrated and made plausible using the following 'library metaphor'. As pointed out in Section 2, any quantitative approach to corpus data must be founded on extensional concepts. In particular, the sublanguage under study has to be defined extensionally as the (past, present and future) output produced by the relevant group of speakers under suitable conditions. Now picture this theoretically unlimited amount of text as a gigantic library containing an infinite number of books. The quantity of interest, i.e. the extensional concept targeted by the statistical analysis, is the relative frequency of a certain phenomenon in the entire library. Note that there is nothing random about the text in the library: every sentence was produced for some specific purpose.

A corpus – as a finite fragment of text from the relevant sublanguage – corresponds to one

---

This pattern can be identified and counted in electronic corpora – with a certain margin of error – using standard corpus annotation and query software.

[4] This value is based on the Brown corpus (Francis and Kucera 1964) and the written part of the British National Corpus (BNC, see Aston and Burnard 1998). Brown corpus frequencies of passive verb forms and verbal groups, respectively, were determined from the tagged version of the Brown corpus distributed as part of the Penn Treebank (Marcus et al. 1993), using the CQP query processor (Christ 1994) and the queries shown below. For the calculation of relative frequencies, a sample size of 1,000,000 word tokens was assumed.

- "be|am|are|is|was|were|been|being|'m|'re|'s" %c [pos = "RB"]* [pos = "VBN"] ;
- [pos = "VB.*|MD"] [pos = "(VB|MD|RB).*"]* [: pos != "(VB|MD|RB).*" :] ;

BNC frequencies for passive verb forms and verbal groups were determined with the BNCWeb interface, CQP edition (Hoffmann and Evert forthcoming), using the following queries:

- [pos = "VB.*"] [pos = "AV0.*"]* [pos = "V.N.*"] ;
- [pos = "V.*"] [pos = "V.*|AV0.*"]* [: pos != "V.*|AV0.*" :] ;

of the books in the library, and the quantitative evidence it provides is the frequency of an observable phenomenon in this individual book. The selection of a particular corpus as the basis for a study – among all the other language fragments that could also have been used – is like picking an arbitrary book from one of the shelves in the library. It is this choice which introduces an element of randomness into corpus frequency data; and it is this element of randomness, in turn, that needs to be accounted for by the methods of statistical inference.[5]

book chosen as corpus	frequency of passives
A Christmas Carol	7.5 per 1,000 words
A Tale of Two Cities	9.3 per 1,000 words
David Copperfield	8.0 per 1,000 words
Dombey and Son	7.8 per 1,000 words
Great Expectations	8.3 per 1,000 words
Hard Times	9.0 per 1,000 words
Master Humphrey's Clock	9.5 per 1,000 words
Nicholas Nickleby	8.2 per 1,000 words
Oliver Twist	8.5 per 1,000 words
Our Mutual Friend	8.0 per 1,000 words
Sketches by BOZ	10.8 per 1,000 words
The Old Curiosity Shop	7.7 per 1,000 words
The Pickwick Papers	8.3 per 1,000 words
Three Ghost Stories	11.0 per 1,000 words

Table 1: Relative frequency of passives in several novels by Charles Dickens

For example, assume that we first walk to a random bookshelf, which happens to contain novels by Charles Dickens. When we now pick a book from this shelf and count the number of passives, the relative frequency obtained would be different for each of the possible choices, as shown by the list of novels in Table 1. The relative frequencies are spread quite evenly across the range from 7.5 to 9.5 passive verb forms per 1,000 words, exceptions being *Sketches by BOZ* (10.8 per 1,000 words) and *Three Ghost Stories* (11 per 1,000 words).[6]

The goal of statistical inference is to account for this variation of observed frequencies from one corpus to another. In general, the observed relative frequency in a corpus is used as an estimate for the overall relative frequency in the entire library. Hence, this immediate result of a quantitative study will be different for each of the possible choices: sometimes it will be accurate, sometimes it will happen to be too high or too low. Since we cannot know which of these classes our corpus belongs to (without knowing the contents of the entire library), statistical inference determines a range of plausible values for the overall frequency in the

[5] It may seem that the selection of a corpus is often not as arbitrary as this comparison suggests: many corpus-linguistic studies are based on a subset of the BNC or another existing corpus. However, this corpus itself represents a random choice among all the material that could have been used to compile the BNC. In the metaphor, studies based on the BNC always pick the same book (and will therefore obtain the same numbers for the same phenomena), but the initial choice of the book was random and has to be taken into account when the results are generalised to the full language that the BNC represents.

[6] The relative frequencies listed in Table 1 are based on electronic editions of the novels from Project Gutenberg (http://www.gutenberg.net/), collected in a demonstration corpus that is distributed together with the IMS Corpus Workbench (Christ 1994). This corpus can be accessed online at http://www.ims.uni-stuttgart.de/projekte/CQPDemos/cqpdemo.html. The same corpus query as for the Brown corpus was used to identify passive verbs. Note that the relative frequencies shown by the online version are approximately 15% lower than those given in this paper because they are computed relative to all tokens (including punctuation) rather than just the word tokens. The frequencies are also considerably lower than the average for modern written English (cf. Section 2). We will return to this point in Section 4 when discussing non-randomness in the corpus data.

library from the observed data, which is called a 'confidence interval' (DeGroot and Schervish 2002, 409-415). Taking *Oliver Twist* (1,350 passives among 159,391 word tokens) as an example, a conservative confidence interval would range from 7.7 to 9.3 passives per 1,000 words, encompassing the values found in most of Dickens' novels (except for the 'outliers' *Sketches by BOZ* and *Three Ghost Stories*.[7]

A major problem for the application of such confidence intervals and similar methods of statistical inference is that they require a good understanding of how much variation there is between the different books in the library. Unfortunately, it is impossible to determine this variation without access to the entire library. We may seem to be caught in a vicious circle, but there is a somewhat vandalistic solution. Imagine that someone went through the library and cut every book into small paper slips, each one carrying a single token (word, phrase or sentence, depending on the unit of measurement). This would leave a big heap of paper slips, containing exactly the same words with exactly the same relative frequencies as the original library. Instead of picking a book from one of the shelves, we can now take a handful of paper slips from this heap, giving us a *random sample* of tokens from the library.

The variation of observed frequencies for the different possible random samples can be predicted mathematically. This leads to a binomial distribution (DeGroot and Schervish 2002, 247-250) for low-frequency data (roughly, less than 10 occurrences) and a normal distribution (DeGroot and Schervish 2002, 268-281) for higher frequencies. All the statistical models and applications listed at the beginning of Section 1 are based on these two distributions. In order to see that there is no difference in principle between picking a book and scooping up a handful of paper slips, imagine further that our conscience-stricken vandal, in an attempt to make up for his misdeed, fills the paper slips into small boxes, one handful at a time, which he then puts back onto the shelves. Now we have a library full of boxes of paper slips instead of the original books, but still containing the same words in the same quantities. Picking a box from one of the shelves – in the same way as we picked a book from the original library – amounts to taking a random sample of tokens, so the statistical methods are appropriate in this case.

In other words, the random sample model predicts the distribution of observed frequencies across boxes of paper slips, while it is unable to predict the same distribution for the original books. The key question is thus as follows: what is the difference between a book and a handful of paper slips in a box?

4. Sources of non-randomness

4.1. Representativeness and balanced samples

What non-randomness means precisely is that the theoretical statistical distributions derived from the random sample model do not match the actual variation of observed frequencies between different corpora. This is just another way of saying that a box of paper slips (i.e. a random sample) differs from a book (i.e. a corpus). One problem for the random sample model that is relatively easy to solve lies in our choice of a single book from one of the shelves, which of course cannot be representative for the entire library. *Oliver Twist* may be typical of the texts found on the Dickens shelf or in the section with 19th century English literature, but it cannot tell us very much about the contents of the entire library.

---

[7] This binomial confidence interval was computed in Gnu R (R Development Core Team 2003) with the following command: binom.test(1350, 159391, conf.level=.999). The confidence level of .999 indicates that the statistical model is 99.9% certain that the true value of the relative frequency in the entire library falls somewhere within the computed range.

This situation becomes obvious in Table 1. *Oliver Twist* gives a good idea of the frequency of passives in Dickens' writing, but all values listed in the table are considerably lower than what is found for modern written English (in the BNC and the Brown corpus).[8] They are much closer to the frequencies in BNC genres that might be found in neighbouring sections of the library, such as biographies (10.3 per 1,000 words), personal letters (7.9 per 1,000 words) or drama (6.3 per 1,000 words).

Of course, a corpus study will rarely be based on a single book by a single author, for the reason we have just seen, namely that a book is not representative of the full sublanguage being studied (except when this sublanguage happens to be the idiolect of a single author, as in many studies on Shakespeare's English). Instead, a *balanced* corpus is compiled by taking books or fragments of books (henceforth referred to as 'documents') from all sections of the library. Both the Brown corpus and the BNC are such balanced corpora. Brown consists of 500 samples of 2,000 words each, taken from different books, newspapers, etc. The BNC contains more than 4,000 documents of widely different sizes: one can imagine it as a collection of 4,000 short books from the library.

In a sense, a balanced corpus is representative of the relevant sublanguage because it contains material from all the different sections of the library. However, one problem remains: in order to give an accurate picture of relative frequencies in the entire library, books must be selected in proportional numbers according to the relative sizes of the different sections. Without access to the full library, it is impossible to know the sizes of the sections, though. Hence, this step involves assumptions about how much material each section contributes to the library, i.e. assumptions that are necessarily subjective and often disputable. For instance, the BNC contains slightly more than 10% of spoken material. If BNC frequencies are taken to be representative of modern British English, there is an implicit assumption that only 10% of the output of British speakers consists of speech, while the remaining 90% are produced in writing. Based on this assumption, the frequency of passives in modern British English would be estimated to be 11.2 per 1,000 words (from relative frequencies of 12.1 in the written part and 4.2 in the spoken part of the BNC). It is quite likely that the true proportions are just the other way round, in which case the overall frequency would only be 5 passives per 1,000 words.

However, the composition of a balanced corpus is a decision that the linguistic researcher has to make when adopting an extensional view of the sublanguage under investigation. The validity of statistical inference from the balanced corpus to the entire library is not affected by the researcher's decisions, as long as the corpus is a random sample from the library. It may just be that this particular library, i.e. the extensionally defined sublanguage, does not accurately reflect the group of speakers and production conditions that the researcher had in mind (defined by intensional criteria such as age, gender and genre). Once again, we are dealing with a problem of linguistic interpretation rather than statistical inference.

4.2. The unit of sampling

In the previous section an 'external' type of non-randomness was addressed, which is connected to the extensional definition of the goals and settings of a quantitative study, and which can largely be circumvented by compiling a suitably balanced corpus. However, the most serious problem for the random sample model is inherent in all corpus data: namely, that

[8] Recall the confidence interval estimate obtained from Oliver Twist in Section 3. While it encompasses most of Dickens' novels, the frequency of passives in the Brown corpus and the (written part of the) BNC is far outside the interval.

the *unit of sampling* is almost always different from the *unit of measurement*. In the library metaphor, the latter is a word, phrase or sentence, while the former is an entire book or a large connected fragment of a book (which was referred to as a 'document' in Section 4.1).

As pointed out at the end of Section 3, the random sample model assumes that the contents of a book (when cut into paper slips according to the unit of measurement and jumbled around) are sufficiently similar to a handful of such paper slips grabbed from a large heap.[9] This assumption is violated whenever instances of the target phenomenon (whether passive verb forms, a specific noun, some syntactic construction or multi-word expression, etc.) have a tendency to lump together within individual books. This is a well-known effect for terms and multi-word expressions in technical terminology, where it is referred to as *term clustering* or *burstiness* (see e.g. Church 2000). It is quite plausible that similar effects exist for other phenomena such as passives, nominalisations or particular syntactic constructions. For instance, the frequency of passive verb forms might be particularly high in a certain document because of the author's individual style or because the passives are used as a rhetorical device.

The consequences of clustering effects are intuitively obvious when one thinks of a lower-frequency word such as *frictional*, which should occur at most once in a book-sized random sample (and will not be found at all in most of the samples). However, when this word is topical in a document, it is likely to be used much more often than just once. For example, in the BNC *frictional* occurs about once in every two million words on average. Therefore, its frequency in a representative sample of one million words (from the same sublanguage as the BNC) should be one or zero, but *frictional* is found as often as ten times in a single document (HRG, a 42,000-word excerpt from a textbook on polymers). If this document happens to be included in the sample (in which case it will be included in its entirety because documents are the unit of sampling for most balanced corpora), the observed frequency will be inflated to 10 or more. Since such a high value is extremely unlikely to be found in a true random sample, it cannot be corrected for by the statistical methods based on random sampling. A similar case is made for imperative verb forms in the ICE-GB corpus by Gries (this volume, XXX).

Another way of looking at this problem is to state that the variation of observed frequencies between documents (up to a value of 10 in the case of *frictional* in the BNC) is much larger than predicted for random samples of the same size (where the unit of sampling coincides with the unit of measurement, and where *frictional* should occur either once or not at all in a sample). Higher-frequency phenomena (such as our previous example of passive verb forms) are also affected by clustering effects, leading to the same inflation of frequency estimates, although this is less obvious than for the low-frequency words.

5. Validating the randomness assumption

So far, we have convinced ourselves – with the help of the library metaphor – that the random sample model is applicable to corpus frequency data, at least to a certain extent. We have identified an 'external' (Section 4.1) and an 'internal' (Section 4.2) source of non-randomness, and we have seen how this non-randomness increases the variation of observed frequencies between documents beyond what could be corrected for by statistical methods. It is now

[9] It is not essential that this heap contains material from all books in the library. When there is a separate heap for each section of the library, a representative sample can be compiled by combining samples taken from each of the heaps. Mathematically speaking, if each part of a balanced corpus is a random sample from the corresponding section of the library and if the relative sizes of these parts match the relative sizes of the library sections, then the whole corpus is a random sample from the entire library (at least to a very good approximation).

obvious that there will be some amount of non-randomness in any corpus. This could only be avoided by sampling at the unit of measurement, i.e. individual words or sentences from the entire library, which is impracticable because it would require each word or sentence to be taken from a different book.[10] The inherent non-randomness of corpus data renders statistical estimates unreliable, since the random variation on which they are based will be smaller than the true variation of the observed frequencies. However, in order to understand whether these effects are negligible or highly problematic, we need to determine exactly how much non-randomness there is in the corpus data.

In the following experiments, we focus on 'internal' non-randomness caused by clustering effects and a mismatch between the unit of sampling and the unit of measurement. The variation between samples taken from different sections of our metaphorical library is not a problem of the statistical analysis but of corpus design, which has to ensure that the composition of the corpus mirrors the sections of the library (cf. the remarks at the end of Section 4.1). If the random sample model holds *within* each section, its inferences are also valid for a balanced corpus collected from the entire library.[11] We will thus look at the variation within a single section, i.e. a homogeneous subset of the language being studied.

We can test the validity of the random sample model directly by comparing the theoretical normal or binomial distribution of frequencies in the documents with their empirical distribution. The latter is obtained by tabulating the observed frequencies for a large number of documents sampled from the relevant section. For mathematical reasons, it is much easier to carry out the experiment when all these documents have approximately the same size. Such a test collection of documents is provided by the Brown corpus, which contains 500 text fragments of approx. 2,000 words each. However, the Brown corpus cannot be understood as a sample from a single section in the library because it combines material from 15 different categories, ranging from religious texts to humorous writing. We will instead interpret the main division of the corpus into 'informative prose' and 'imaginative prose' as two library sections and test the randomness within each section.[12] Once again, we use the frequency of passive verb forms as an example.

---

[10] Just imagine how difficult it would have been to compile the Brown corpus by sampling one word each from a million books, rather than taking 2,000-word samples from only 500 books. Using the Web as a corpus may soon enable us to obtain random samples of individual word tokens from a very large population, but it will first be necessary to solve a number of methodological problems surrounding the Web as corpus approach (see e.g. Lüdeling, Evert and Baroni forthcoming).

[11] Another apparent contradiction is hiding here: What, you might say, if I make the sections of the library so small (by narrowing down their intensional description) that each of them contains only a single book? Then any balanced corpus automatically comprises all the material from each section selected by the compiler, so that the random sampling assumption is trivially satisfied (because taking all the books from a section gives exactly the same result as taking all the paper slips from a heap on the floor). The fallacy here is that a balanced corpus has to contain a certain number of documents from every section of the library (though it is mathematically difficult to determine exactly how many are required), which can no longer be satisfied when the sections become too numerous and too small.

[12] In addition, the categories 'learned writing' (J), 'press reviews' (C) and 'miscellaneous' (H) were excluded from the informative prose section in order to improve its homogeneity. Ideally, each of the 15 categories should be studied individually as a library section, but the number of documents in a single category (between 6 and 80) is too small to analyse the empirical distribution of observed frequencies.

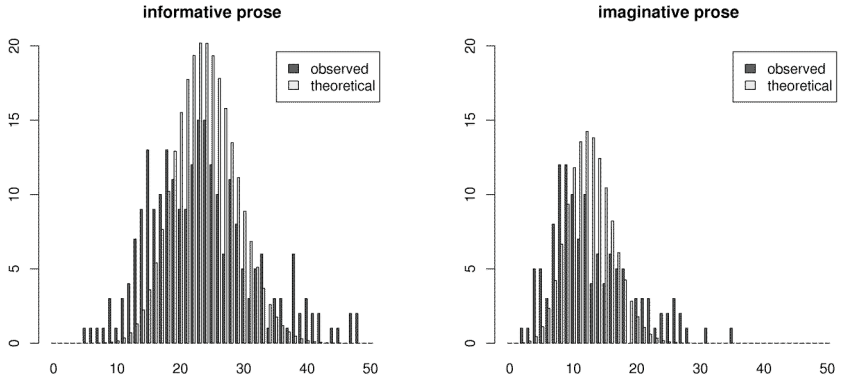

Figure 1: Empirical distribution (dark bars) of the frequencies of passive verb forms in the documents of the Brown corpus vs. theoretical binomial distribution (light bars) predicted by the random sample model

Figure 1 tabulates the (absolute) frequencies of passives in the documents of the Brown corpus, which is done separately for the two parts. Each dark bar represents a certain frequency value (as indicated by the labels on the x-axis), and the length of the bar corresponds to the number of documents with this number of passives (indicated by the labels on the y-axis). For instance, we can see from the left panel that there are 15 documents that contain exactly 23 passives in the informative prose section, and another 13 documents that contain 15 passives. For comparison, the light bars show the binomial distribution predicted by the random sample model. According to this theoretical distribution, there should be some 20 documents with exactly 23 passives, but only 4 that contain 15 passives.

It is obvious from the graphs in Figure 1 that the true distribution is much broader than the theoretical one. In other words, the random sample model underestimates the true variation of frequencies, which means that its confidence intervals are too narrow (because the model has more confidence in its estimate than is justified) and that it will too easily report significant differences between two experimental conditions (e.g. in a frequency comparison study). In order to obtain a quantitative measure for the inaccuracy of the random sample model, we compare the standard deviations (DeGroot and Schervish 2002, 198) of the observed and theoretical distributions. With standard deviations of 9.87 (empirical) vs. 4.87 (theoretical) for informative prose and 6.59 (empirical) vs. 3.54 (theoretical) for imaginative prose, we find that the random sample model underestimates the true variation by a factor of two in both cases.

This very intuitive approach for testing the randomness assumption can only be applied to relatively frequent phenomena. It is difficult to specify the exact limits, but instances of the relevant phenomenon should occur at the very least in 20 different documents from the test collection. For lower frequencies, which are typical when studying lexical items, Baayen (2001, 164-167) suggests a so-called dispersion test. Consider a word (or other phenomenon) that occurs only twice in the entire test collection (referred to as a *dis legomenon*). Under the random sample model, it is very unlikely that both occurrences should be in the same document. Therefore, a word with this property is called 'underdispersed'. As a consequence of Zipf's law, there will be many *dis legomena* in any corpus and it comes as no surprise that some of them are underdispersed. For example, the Brown corpus contains 7,592 word form types that have exactly two occurrences in the corpus. According to the random sample model, we would expect that approx. 15 of these types are underdispersed, i.e. both of their instances appear in the same document. In reality, however, there are 2,351 underdispersed *dis legomena*, indicating a high level of non-randomness (of the term clustering type). Similar observations

can be made for more frequent words: of the 1,943 word form types with five occurrences in the Brown corpus, 1,123 are underdispersed (i.e. they appear in four or fewer documents), while only 39 underdispersed word forms are predicted by the random sample model.[13]

The mathematics of the dispersion test are worked out by Evert (2004, 60-63), and the accompanying software includes an easy-to-use implementation (which was used to obtain the results above).

6. Conclusion and outlook

The aim of this paper was to show that even though a sentence is not a random bag of words, and even though a text is not a random sequence of sentences, it is sensible to apply statistical methods based on a random sample model to corpus frequency data. The metaphor of language (defined extensionally) as a gigantic library, with each book corresponding to a language fragment (or corpus), explains how randomness finds its way into any quantitative corpus study. It is not inherent in the language itself but is introduced by the choice of a particular corpus for the study, which can be likened to picking a book from one of the shelves in a library.

A deeper analysis of the library metaphor has revealed two major sources of non-randomness, which result in a greater variation of the observed corpus frequencies than is predicted by the random sample model. Obviously, this increased variation will distort the quantitative results of a corpus study (in the form of statistical estimates), possibly leading researchers to erroneous conclusions (such as spurious claims about differences between genres).

One cause of non-randomness is the (lack of) representativeness of a homogeneous corpus. This is not an inherent problem of the statistical models, though, and can largely be avoided by compiling a balanced corpus from different sections of the library. A much more problematic cause of non-randomness is the discrepancy between the unit of measurement (typically words or sentences) and the unit of sampling (entire documents or long connected fragments). Since word types and many other linguistic phenomena tend to cluster together in the same document, observed frequencies are inflated when such documents are included in the corpus (compared to the frequencies that would be obtained for random samples from the same library).

Having identified the two most important causes of non-randomness, we looked at two practicable methods for quantifying the degree to which frequency estimates are affected by non-randomness in the corpus data. These methods are based on test collections of documents from relevant sections of the metaphorical library. An illustration of the procedure, using passive verb forms as an example and the Brown corpus as a test collection, revealed the actual variation of corpus frequencies to be twice as large as predicted by the random sample model. Preliminary experiments with low-frequency data indicate even larger discrepancies between empirical and theoretical distributions.

The obvious next step is to learn more about the nature and the amount of non-randomness in corpus data by applying the methods presented in Section 5 to other linguistic phenomena as well as to other test collections. There can be no doubt, however, that non-

[13] Note that we have pooled the data from the entire Brown corpus for this experiment. Since it is designed to detect term clustering effects within documents, the differences in relative frequency between the fifteen categories are irrelevant. A different application of the dispersion test – which measures dispersion at the category level – can be used to estimate how homogeneous the distribution of low-frequency items is across the categories.

randomness is a ubiquitous problem in corpus linguistics. In a longer perspective, it will thus be necessary to develop new statistical methods for the analysis of corpus frequency data. These methods should be able to deal with non-random data and correct for the larger amount of variation in some way. First tentative steps have been taken – either by adjusting the unit of sampling (Kilgarriff 2001) or by modifying the random sampling model itself (e.g. Katz 1996) – but much work still needs to be done.

Works Cited

Aston, Guy and Lou Burnard (1998). *The BNC Handbook*. Edinburgh: Edinburgh University Press. See also the BNC homepage <http://www.natcorp.ox.ac.uk/> (January 8, 2006).

Baayen, R. Harald (2001). *Word Frequency Distributions*. Dordrecht: Kluwer.

Baayen, R. Harald (2003). "Probabilistic approaches to morphology." R. Bod, J. Hay and S. Jannedy, eds. *Probabilistic Linguistics*. Cambridge, MA: MIT Press, 229-287.

Biber, Douglas (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Christ, Oliver (1994). "A modular and flexible architecture for an integrated corpus query system." *Papers in Computational Lexicography (COMPLEX '94)*. Budapest, Hungary, 22-32. See also the homepage of the IMS Corpus Workbench <http://cwb.sourceforge.net/> (January 20, 2006).

Church, Kenneth W. (2000). "Empirical estimates of adaptation: the chance of two Noriegas is closer to p/2 than p^2." *Proceedings of COLING 2000*. Saarbrücken, Germany, 173-179.

DeGroot, Morris H. and Mark J. Schervish (2002). *Probability and Statistics*. 3rd ed. Boston: Addison Wesley.

Evert, Stefan (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Published in 2005, URN urn:nbn:de:bsz:93-opus-23714. Available from <http://www.collocations.de/phd.html> (January 20, 2006).

Francis, W. N. and H. Kucera (1964). *Manual of Information to Accompany a Standard Sample of Present-day Edited American English, for Use with Digital Computers*. Technical report, Department of Linguistics, Brown University, Providence, RI. Revised ed. 1971; revised and augmented 1979.

Hoffmann, Sebastian and Stefan Evert (forthcoming). "BNCweb (CQP Edition): the marriage of two corpus tools." S. Braun, K. Kohn and J. Mukherjee, eds. *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. (English Corpus Linguistics 3). Frankfurt am Main: Lang.

Katz, Slava M. (1996). "Distribution of content words and phrases in text and language modelling." *Natural Language Engineering* 2.2, 15-59.

Kilgarriff, Adam (2001). "Comparing corpora." *International Journal of Corpus Linguistics* 6.1, 1-37.

Kucera, H. and W. N. Francis (1967). *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

Lüdeling, Anke, Stefan Evert and Marco Baroni (forthcoming). "Using Web data for linguistic purposes." M. Hundt, C. Biewer and N. Nesselhauf, eds. *Corpus Linguistics and the Web*. Amsterdam: Rodopi.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz (1993). "Building a large annotated corpus of English: the Penn treebank." *Computational Linguistics* 19.2, 313-330.

R Development Core Team (2003). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3. See also <http://www.r-project.org/> (January 8, 2006).

Zuraw, Kie (2003). "Probability in language change." R. Bod, J. Hay and S. Jannedy, eds. *Probabilistic Linguistics*. Cambridge, MA: MIT Press, 139-176.