)INBURGH TEXTBOOKS IN EMPIRICAL LINGUISTICS

CORPUS LINGUISTICS
by Tony McEnery and Andrew Wilson

LANGUAGE AND COMPUTERS
A PRACTICAL INTRODUCTION TO THE COMPUTER ANALYSIS OF LANGUAGE
by Geoff Barnbrook

STATISTICS FOR CORPUS LINGUISTICS
by Michael Oakes

COMPUTER CORPUS LEXICOGRAPHY
by Vincent B.Y. Ooi

THE BNC HANDBOOK
EXPLORING THE BRITISH NATIONAL CORPUS WITH SARA
by Guy Aston and Lou Burnard

PROGRAMMING FOR CORPUS LINGUSTICS
HOW TO DO TEXT ANALYSIS WITH JAVA
by Oliver Mason

'ou would like information on forthcoming titles in this series, please contact
Edinburgh University Press, 22 George Square, Edinburgh EH8 9LF

EDINBURGH TEXTBOOKS IN EMPIRICAL LINGUISTICS

Series Editors: Tony McEnery and Andrew Wilson

Corpus Linguistics

An Introduction

Tony McEnery and Andrew Wilson

Second Edition

EDINBURGH UNIVERSITY PRESS

:cording to their theoretical perspectives on linguistic variation.

Having defined the population, one needs to determine which sample sizes 'e most representative of it, both in terms of the optimal *length* of each sample xt and the optimal *number* of texts which should be included in the corpus. oth these figures are ultimately dependent on the distribution of linguistic atures within the population, that is, what is the probability that y text nples of length n will contain proportionately the same number and distri- tion of examples of particular items as the total population? In a pilot study, ber found that frequent items are stable in their distributions and hence all samples are adequate for these. Rarer features on the other hand show ore variation in their distributions and consequently require larger samples hey are to be fully represented in the corpus, as de Haan (1992) has also served. In terms of such rarer features, therefore, we can perhaps admit that omsky's criticism of the small corpora of the 1950s was a valid one.

Biber notes that the standard statistical equations which are used to deter- ne these optimal sample lengths and sample numbers are problematic for pus building (1993b). This is because they require two statistical values ich cannot be computed for a corpus as a whole: **standard deviations**, ich must be calculated for each individual feature, and **tolerable error**, ich will vary according to the overall frequency of a feature. These values therefore, problematic, since a corpus, unless collected for one specific pose, is normally intended for use in research on many different features of guage. Biber's suggestion in this situation is that the most conservative way nsuring representative samples is to base the computations on the most ely varying feature. With regard to sample lengths, taking samples of sizes ch are representative of that feature should mean that the samples are also esentative of those features which show less variation in distribution. ilarly, with the number of texts within each genre, the degree of variation hat feature which occurs within given genres is used to scale the number xts required to represent each genre.

: will be appreciated, then, that corpus sampling is by no means a straight- ard exercise. However, the constant application of strict statistical proce- s should ensure that the corpus is as representative as possible of the larger ulation, within the limits imposed by practicality.

ne way of supplementing these procedures for enhancing the representa- ess of corpora is the use of dispersion statistics. Dispersion is a measure of evenly distributed the occurrence of a feature is in a text or corpus – for nple, whether its appearance is restricted mainly to a few places or whether curs much more widely. Frequency alone cannot be a measure of typical- n a corpus of ten genres, two words might both have a frequency of 20, ne of these words might have two occurrences in each of the ten genres eas the other's 20 occurrences might all be concentrated within a single . Dispersion can thus tell us how typical a word is and not just how often

it occurs: it can serve to counterbalance the concern, voiced amongst others by Chomsky, about the potential skewedness of corpora. To take a couple of fairly obvious examples, we might expect the words 'the' and 'and' to be very widely, and quite evenly, distributed, whereas a word such as 'autopsy' would occur only rarely outside certain text types – e.g. crime reporting, crime fiction and medical writing – but might occur very frequently within those text types. We can also use dispersion to examine the distribution of different senses of the same word – i.e. which are the most typical and which are more specialised.

The most reliable dispersion measure has been found to be Juilland's D coefficient (Lyne 1985). For this equation and a brief discussion of it in the context of other dispersion measures, see the volume by Oakes in this series (Oakes 1998: 189–92).

Dispersion measures, though used in early computer-based work on word frequencies (see the discussion of Juilland's work in Chapter 1), have been rather neglected in modern corpus linguistics, but have in the last few years come to prominence again – see, for example, the study of fixed expressions by De Cock *et al.* (1998).

Start Here

3.4. APPROACHING QUANTITATIVE DATA

In the preceding sections, we have seen the value of supplementing qualitative analyses of language with quantitative data. We have also seen why corpora in particular are of value for quantitative linguistic analysis. But it should be noted that the use of quantification in corpus linguistics typically goes well beyond simple counting: many sophisticated statistical techniques are used which can both provide a mathematically rigorous analysis of often complex data – one might almost say, colloquially, to bring order out of chaos – and be used to show with some degree of certainty that differences between texts, genres, languages and so on are real ones and not simply a fluke of the sampling procedure.

In this section we introduce briefly some of the quantitative methods which are of most value in working practically with corpora. But before we move on we must raise two notes of caution. First, this section is of necessity incomplete. There are very many statistical techniques which have been, or can potentially be, applied to corpora and space precludes coverage of all of them. Instead we have chosen to concentrate on those which we consider to be the most important and most widely used. Second, we do not aim here to provide a complete step-by-step 'how to do it' guide to statistics. Many of the statistical techniques used in corpus linguistics are very complex and most require the use of computer software for them to be made manageable. To explain the mathematics fully we would need something approaching a full chapter for each technique. What we have done instead is to try to outline with as little mathematics as possible what each technique does and why it is of practical value to the corpus linguist. Other books in the Edinburgh Textbooks in

Empirical Linguistics series – notably *Language and Computers* and *Statistics for Corpus Linguistics* – pick up again on these methods and present them in much more detail than we have space to do here. Our recommendation is that the student reads what we have to say here as a brief introduction to how the techniques may be used and then progresses to the more detailed treatments in the other texts for further explanation. In what follows, precise references are given to the more detailed treatments in these two volumes.

3.4.1. Frequency counts

The most straightforward approach to working with quantitative data is simply to classify items according to a particular scheme and to perform an arithmetical count of the number of items (or *tokens*) within the text which belong to each classification (or *type*) within the scheme. So, for instance, we might set up a classification scheme to look at the frequency of four major parts of speech – noun, verb, adjective and adverb. These four classes would constitute our types. Every time we met a word in the corpus which belonged to one of these categories – a token of one of the types – we would simply add 1 to the count for corresponding category type. Sometimes the classification scheme used in frequency counts may be a simple one-to-one mapping of form on to classification. This can be the case with word frequency analysis, where each graphical word form is equivalent to one type within the classification scheme. More often, however, the use of a classification scheme implies a deliberate act of categorisation on the part of the investigator. This is even sometimes the case with word frequency analysis, in that variant forms of the same lexeme may be lemmatised before a frequency count is made: for instance, *loved*, *loving* and *loves* might all be considered to be instances of the lexeme LOVE. Very often the classification scheme will correspond to the types of linguistic annotation which have already been introduced into the corpus at some earlier stage (see Chapter 2). An example of this might be an analysis of the incidence of different parts of speech in a corpus, where the parts of speech have been previously classified by automated or manual part-of-speech tagging.

3.4.2. Working with proportions

Simple frequency counts are a useful approach to quantifying linguistic data and they have often been used in corpus-based research. However, they have certain disadvantages. The main disadvantage arises when one wishes to compare one data set with another – for example, a corpus of spoken language with a corpus of written language. Arithmetical frequency counts simply count occurrences: they do *not* indicate the prevalence of a type in terms of a proportion of the total number of tokens within the text. This can be problematic when the two or more texts or corpora which are being compared are different in size. Where a disparity in size occurs, a simple frequency count of a type in one text, although it is larger than the count for the same

type in another text, may actually indicate a smaller *proportion* of the type in that text than the smaller count indicates for the other text. For instance, assume that we have a corpus of spoken English with a size of 50,000 words and a corpus of written English with a size of 500,000 words. We may find that, in the corpus of spoken English, the word *boot* occurs 50 times whereas, in the corpus of written English, it occurs 500 times. So it looks at first glance as if *boot* is more frequent in written than in spoken English. But let us now look at these data in a different way. This time we shall go one step further beyond the simple arithmetical frequency and calculate the frequency of occurrence of the type *boot* as a percentage of the total number of tokens in the corpus, that is, the total size of the corpus. So we do the following calculations:

spoken English: 50 / 50,000 x 100 = 0.1%
written English: 500 / 500,000 x 100 = 0.1%

Looking at these figures, we see that, far from being 10 times more frequent in written English than in spoken English, *boot* has the same frequency of occurrence in both varieties: it makes up 0.1 per cent of the total number of tokens in each sample. It should be noted, therefore, that if the sample sizes on which a count is based are different, then simple arithmetical frequency counts cannot be compared directly with one another: it is necessary in those cases to normalise the data using some indicator of proportion. Even where disparity of size is not an issue, proportional statistics are a better approach to presenting frequencies, since most people find it easier to understand and compare figures such as percentages than fractions of unusual numbers such as 53,000.

There are several ways of indicating proportion, but they all boil down to a ratio between the size of the sample and the number of occurrences of the type under investigation. The most basic involves simply calculating the ratio:

ratio = number of occurrences of the type / number of tokens in entire sample

The result of this calculation may be expressed as a fraction or, more commonly, as a decimal. Usually, however, when working with large samples such as corpora and potentially many classifications, this calculation gives unwieldy looking small numbers. For example, the calculation we performed above would give a simple ratio of 0.0001. Normally, therefore, the ratio is scaled up to a larger, more manageable number by multiplying the result of the above equation by a constant. This is what we did with the example: in that case the constant was 100 and the result was therefore a percentage. Percentages are perhaps the most common way of representing proportions in empirical linguistics but, with a large number of classifications, or with a set of classifications in which the first few make up something like half the entire sample, the numbers can still look awkward, with few being greater than 1. It may sometimes be sensible, therefore, to multiply the ratio formula by a larger constant, for example 1,000 (giving a proportion *per mille* (‰)) or 1,000,000 (giving a

proportion in parts per million (p.p.m.)). It is not crucial which option is selected: what is important is to indicate clearly which has been used.

3.4.3. Tests of significance

Let us suppose now that we are interested in examining the Latin versions of the Gospel of Matthew and the Gospel of John. We are interested in looking at how third person singular speech is represented and specifically in comparing how often the present tense form of the verb 'to say' is used (i.e. *dicit*) and how often the perfect form of the verb used (i.e. *dixit*). So we decide to make a simple count of each of these two verb forms in each of the two texts. Having done this, we arrive at the following frequencies:

	dicit	*dixit*
Matthew	46	107
John	118	119

From these figures, it looks as if John uses the present tense form (*dicit*) proportionally more often than Matthew does. But with what degree of certainty can we infer that this is a genuine finding about the two texts rather than a result of chance? From these figures alone we cannot decide: we need to perform a further calculation – a test of **statistical significance** – to determine how high or low the probability is that the difference between the two texts on these features is due to chance.

There are several significance tests available to the corpus linguist – the chi-squared test, the [Student's] t-test, Wilcoxon's rank sum test and so on – and we will not try to cover each one here. As an example of the role of such tests we will concentrate on just one test – **the chi-squared test** (see Oakes 1998: 24–9). The chi-squared test is probably the most commonly used significance test in corpus linguistics and also has the advantages that (1) it is more sensitive than, for example, the t-test; (2) it does not assume that the data are 'normally distributed' – this is often not true of linguistic data; and (3) in 2 x 2 tables such as the one above – a common calculation in linguistics – it is very easy to calculate, even without a computer statistics package (see Swinscow 1983). Note, however, in comparison to Swinscow, that Oakes (1998: 25) recommends the use of Yates's correction with 2 x 2 tables. The main disadvantage of chi-square is that it is unreliable with very small frequencies. It should also be noted that proportional data (percentages etc.) *cannot* be used with the chi-squared test: disparities in corpus size are unimportant, since the chi-squared test itself compares the figures in the table proportionally.

Very simply, the chi-squared test compares the difference between the actual frequencies which have been observed in the corpus (the *observed* frequencies) and those which one would expect if no factor other than chance had been operating to affect the frequencies (the *expected* frequencies). The closer the expected frequencies are to the observed frequencies, the more likely it is that the observed frequencies are a result of chance. On the other hand, the greater the difference between the observed frequencies and the expected frequencies, the more likely it is that the observed frequencies are being influenced by something other than chance, for instance, a true difference in the grammars of two language varieties.

Let us for the present purpose omit the technicality of calculating the chi-square value and assume that it has already been calculated for our data. Having done this, it is then necessary (if not using a computer program which gives the information automatically) to look in a set of statistical tables to see how significant our chi-square value is. To do this one first requires one further value – the number of **degrees of freedom** (usually written **d.f.**). This is very simple to work out. It is simply:

(number of columns in the frequency table – 1) x (number of rows in the frequency table – 1)

We now look in the table of chi-square values in the row for the relevant number of degrees of freedom until we find the nearest chi-square value to the one which has been calculated, then we read off the probability value for that column. A probability value close to 0 means that the difference is very strongly significant, that is, it is very unlikely to be due to chance; a value close to 1 means that it is almost certainly due to chance. Although the interval between 1 and 0 is a continuum, in practice it is normal to assign a cut-off point which is taken to be the difference between a 'significant' result and an 'insignificant' result. In linguistics (and most other fields) this is normally taken to be a probability value of 0.05: probability values of less than 0.05 (written as $p < 0.05$) are assumed to be significant, whereas those greater than 0.05 are not.

Let us then return to our example and find out whether the difference which we found is statistically significant. If we calculate the chi-square value for this table (using Yate's correction) we find that it is 14.04. We have two columns and two rows in the original frequency table, so the number of degrees of freedom in this case is (2 – 1) x (2 – 1) = 1 d.f. For 1 d.f. we find that the probability value for this chi-square value is 0.0002. Thus the difference which we found between Matthew and John is significant at $p < 0.05$, and we can therefore say with quite a high degree of certainty that this difference is a true reflection of variation in the two texts and is not due to chance.

As an alternative to the chi-squared test, Dunning (1993) proposes the use of the log-likelihood test (also G^2 or λ), which, he argues, is more reliable with low frequencies and with samples that have a comparatively large discrepancy in size.

3.4.4. Significant collocations

The idea of **collocations** – the characteristic co-occurrence patterns of words – is an important one in many areas of linguistics. (For detailed treatments of collocation, see Oakes 1998, Chapter 4, Section 3 and Barnbrook 1996,