29

COMPARING CORPORA

Adam Kilgarriff

Source: *International Journal of Corpus Linguistics*, 6:1 (2001), 97–133.

Abstract

Corpus linguistics lacks strategies for describing and comparing corpora. Currently most descriptions of corpora are textual, and questions such as 'what sort of a corpus is this?', or 'how does this corpus compare to that?' can only be answered impressionistically. This paper considers various ways in which different corpora can be compared more objectively. First we address the issue, 'which words are particularly characteristic of a corpus?', reviewing and critiquing the statistical methods which have been applied to the question and proposing the use of the Mann-Whitney ranks test. Results of two corpus comparisons using the ranks test are presented. Then, we consider measures for corpus similarity. After discussing limitations of the idea of corpus similarity, we present a method for evaluating corpus similarity measures. We consider several measures and establish that a χ^2-based one performs best. All methods considered in this paper are based on word and ngram frequencies; the strategy is defended.

1 Introduction

There is a void at the heart of corpus linguistics. The name puts 'corpus' at the centre of the discipline.[1] In any science, one expects to find a useful account of how its central constructs are taxonomised and measured, and how the subspecies compare. But to date, corpus linguistics has measured corpora only in the most rudimentary ways, ways which provide no leverage on the different kinds of corpora there are. The terms it has used for taxonomising corpora have been unrelated to any measurement: a corpus is described as "Wall Street Journal" or "transcripts of business meetings" or "foreign learners' essays (intermediate grade)", but if a new corpus is to be compared with existing ones, there are no methods for quantifying how it stands in relation to them.

The lack arises periodically wherever corpora are discussed. If an interesting finding is generated using one corpus, for what other corpora does it hold? On the CORPORA mailing list (http://www.hd.vib.no/corpora), the issue is aired every few months. Recently it arose in relation to the size of corpus that had to be gathered to test certain hypotheses: a reference was made to Biber (1990 and 1993a) where corpus sizes for various tasks are discussed. The next question is, what sort of corpus did Biber's figures relate to? If the corpus is highly homogeneous, less data will be required. But there are no established measures for homogeneity.

Two salient questions are "how similar are two corpora", and "in what ways do two corpora differ?" The second question has a longer history to it, so is taken first. Researchers have wanted to answer it for a variety of reasons, in a variety of academic disciplines. In the first part of the paper, the statistical and other techniques used in linguistics, social science, information retrieval, natural language processing and speech processing are critically reviewed.

Almost all the techniques considered work with word frequencies. While a full comparison between any two corpora would of course cover many other matters, the concern of this paper is with the statistical framework. Reliable statistics depend on features that are reliably countable and, foremost amongst these, in language corpora, are words.

The first part of the paper surveys and critiques the statistical methods which have been applied to finding the words which are most characteristic of one corpus as against another, and identifies the Mann-Whitney ranks test as a suitable technique. The next part goes on to use it to compare British and American English, and male and female conversational speech.

We then move on to address corpus similarity directly. Measures are needed not only for theoretical and research work, but also to address practical questions that arise wherever corpora are used: is a new corpus sufficiently different from available ones, to make it worth acquiring? When will a grammar based on one corpus be valid for another? How much will it cost to port a Natural Language Processing (NLP) application from one domain, with one corpus, to another, with another? Various measures for corpus similarity (and homogeneity) are proposed, a strategy for evaluating the measures is presented and the measures are evaluated.

2 Which words are particularly characteristic of a corpus (or text)?

In the simplest formulation of the problem, we ignore the internal structure of the corpora, so the corpora could be single texts, and are referred to as such below. For two texts, which words best characterise their differences?

Table 1 Basic contingency table.

	X	*Y*	
w	*a*	*b*	*a* + *b*
not *w*	*c*	*d*	*c* + *d*
	a + *c*	*b* + *d*	*a* + *b* + *c* + *d* = N

For word *w* in texts X and Y, this might be represented in a contingency table as in Table 1.

There are *a* occurrences of *w* in text X (which contains *a* + *c* words) and *b* in Y (which has *b* + *d* words).

Measuring the distinctiveness of words in corpora is, for some purposes, homologous to measuring the distinctiveness of combinations of words in a corpus (e.g. bigrams and similar). Daille (1995) presents a review and assessment of measures of strength of co-occurrence, in a paper which can be seen as a complement to this one. She considers a wider range of measures, but her best candidates are considered here. For bigrams, the columns are for ω_2 and not-ω_2 rather than text X and text Y, and the window of words within which ω and ω_2 must both occur for it to count as a co-occurrence must also be defined.

2.1 The χ^2-test

We now need to relate our question to a hypothesis we can test. A first candidate is the null hypothesis that both texts comprise words drawn randomly from some larger population; for a contingency table of dimensions $m \times n$, if the null hypothesis is true, the statistic:

$$\sum \frac{(\mathrm{O} - \mathrm{E})^2}{\mathrm{E}}$$

(where O is the observed value, E is the expected value calculated on the basis of the joint corpus, and the sum is over the cells of the contingency table) will be χ^2-distributed with $(m - 1) \times (n - 1)$ degrees of freedom.[2] For our 2 × 2 contingency table the statistic has one degree of freedom and Yate's correction is applied, subtracting $^1/_2$ from $|\mathrm{O} - \mathrm{E}|$ before squaring. Wherever the statistic is greater than the critical value of 7.88, we conclude with 99.5% confidence that, in terms of the word we are looking at, X and Y are not random samples of the same larger population.

This is the strategy adopted by Hofland and Johansson (1982), Leech and Fallon (1992), to identify where words are more common in British than American English or vice versa. X was the Lancaster-Oslo-Bergen (LOB) corpus, Y was the Brown, and, in the table where the comparison is made, words are marked *a* where the null hypothesis was defeated with 99.9% confidence, *b* where it was defeated with 99% confidence, and *c* where it was defeated with 95% confidence. Rayson, Leech, and Hodges (1997) use the χ^2 similarly for the analysis of the conversation component of the British National Corpus (BNC).[3]

Looking at the LOB-Brown comparison, we find that this is true for very many words, and for almost all very common words. Much of the time, the null hypothesis is defeated. At a first pass, this would appear to demonstrate that all those words have systematically different patterns of usage in British and American English, the two types that the two corpora were designed to represent. A first experiment was designed to check whether this was an appropriate interpretation.

2.2 *Experiment: same-genre subsets of the BNC*

If the χ^2-test was picking up on interesting differences between the corpora, then, if there were no such differences, the null hypothesis would not be defeated. To test this, I took two corpora which were indisputably of the same language type: each was a random subset of the written part of the British National Corpus (BNC). The sampling was as follows: all texts shorter than 20,000 words were excluded. This left 820 texts. Half the texts were then randomly assigned to each of two subcorpora.

If we randomly assign words (as opposed to documents) to the one corpus or the other, then we have a straightforward random distribution, with the value of the χ^2-statistic equal to or greater than the 99.5% confidence threshold of 7.88 for just 0.5% of words. The average value of the error term,

$$(|O - E| - 0.5)^2/\mathrm{E}$$

is then 0.5.[4] The hypothesis can, therefore, be couched as: are the error terms systematically greater than 0.5? If they are, we should be wary of attributing high error terms to significant differences between text types, since we also obtain many high error terms where there are no significant differences between text types.

Frequency lists for word-POS pairs for each subcorpus were generated. For each word occurring in either subcorpus, the error term which would have contributed to a chi-square calculation was determined. As Table 2 shows, average values for the error term are far greater than 0.5, and tend to increase as word frequency increases.

As the averages indicate, the error term is very often greater than 0.5 × 7.88 = 3.94, the relevant critical value of the chi-square statistic. As in the LOB-Brown comparison, for very many words, including most common words, the null hypothesis is defeated.

Table 2 Comparing two same-genre corpora using χ^2: Mean error term is far greater than 0.5, and increases with frequency. POS tags are drawn from the CLAWS-5 tagset as used in the BNC: see http:/info.ox.ac.uk/bnc.

Class	*First item in class*		*Mean error term*
(Words in freq. order)	*word*	*POS*	*for items in class*
First 10 items	the	DET	18.76
Next 10 items	for	PRP	17.45
Next 20 items	not	XX	14.39
Next 40 items	have	VHB	10.71
Next 80 items	also	AV0	7.03
Next 160 items	know	VVI	6.40
Next 320 items	six	CRD	5.30
Next 640 items	finally	AV0	6.71
Next 1280 items	plants	NN2	6.05
Next 2560 items	pocket	NN1	5.82
Next 5120 items	represent	VVB	4.53
Next 10240 items	peking	NP0	3.07
Next 20480 items	fondly	AV0	1.87
Next 40960 items	chandelier	NN1	1.15

2.2.1 Discussion

This reveals a bald, obvious fact about language. Words are not selected at random. There is no *a priori* reason to expect them to behave as if they had been, and indeed they do not. The LOB-Brown differences cannot in general be interpreted as British-American differences: it is in the nature of language that any two collections of texts, covering a wide range of registers (and comprising, say, less than a thousand samples of over a thousand words each) will show such differences. While it might seem plausible that oddities would in some way balance out to give a population that was indistinguishable from one where the individual words (as opposed to the texts) had been randomly selected, this turns out not to be the case.

Let us look closer at why this occurs. A key word in the last paragraph is 'indistinguishable'. In hypothesis testing, the objective is generally to see if the population can be distinguished from one that has been randomly generated—or, in our case, to see if the two populations are distinguishable from two populations which have been randomly generated on the basis of the frequencies in the joint corpus. Since words in a text are not random, we know that our corpora are not randomly generated. The only question, then, is whether there is enough evidence to say that they are not, with confidence. In general, where a word is more common, there is more evidence. This is why a higher proportion of common words than of rare ones defeat the null hypothesis. As one statistics textbook puts it:

> None of the null hypotheses we have considered with respect to goodness of fit can be *exactly* true, so if we increase the sample size (and hence the value of χ^2) we would ultimately reach the point when all null hypotheses would be rejected. All that the χ^2 test can tell us, then, is that the sample size is too small to reject the null hypothesis!
>
> (Owen and Jones 1977: 359)

For large corpora and common words, the sample size is no longer too small.

The χ^2-test can be used for all sizes of contingency tables, so can be used to compare two corpora in respect of a set of words, large or small, rather than one-word-at-a-time. In all the experiments in which I have compared corpora in respect of a substantial set of words, the null hypothesis has been defeated (by a huge margin).

The original question was not about which words are random but about which words are most distinctive. It might seem that these are converses, and that the words with the highest values for the error term—those for which the null hypothesis is most soundly defeated—will also be the ones which are most distinctive to one corpus or the other. Where the overall frequency for a word in the joint corpus is held constant, this is valid, but as we have seen, for very common words, high χ^2 values are associated with the sheer quantity of evidence and are not necessarily associated with a pre-theoretical notion of distinctiveness (and for words with expected frequency less than 5, the test is not usable).

2.3 Mann-Whitney ranks test

The Mann-Whitney (also known as Wilcoxon) ranks test can be applied to corpus data if the two corpora are first divided into same-sized samples. Then the numbers of occurrences of a word are directly comparable across all samples in both corpora. The test addresses the null hypothesis—that all samples are from the same population—by seeing whether the counts from the samples in the one corpus are usually bigger than ones from the other, or usually smaller, or similarly spread. The frequencies of word w in each sample are labelled with the corpus they come from, put in rank order, and numbered from 1 to $m + n$ (where there are m samples in the one corpus and – in the other) according to their rank. All the ranks of items coming from the smaller corpus are summed. The sum is then compared with the figure that would be expected on the basis of the null hypothesis, as tabulated in statistics textbooks.

To demonstrate: suppose corpora X and Y have been divided into five equal-sized parts. We count the frequencies of a given word in each of the ten parts, five of X and 5 of Y, and rank them, as in Table 3.

Table 3 Mann-Whitney ranks test.

Count:	3	3	12	13	15	18	24	27	33	88	TOTAL
Corpus:	Y	Y	X	Y	X	X	X	Y	Y	X	
Rank (X)			3		5	6	7			10	31
Rank (Y)	1	2		4				8	9		24

The sum of ranks associated with the corpus with the smaller number of samples, here 24, is compared with the value that would be expected, on the basis of the null hypothesis. If the null hypothesis were true, the tables tell us that 95% of the time the statistic would be in the range 18.37–36.63: 24 is within this range, so there is no evidence against the null hypothesis.

Note that the one very high count of 88 has only limited impact on the statistic. This is the desired behaviour for our task, since it is of limited interest if a single document in a corpus has very many occurrences of a word.

Sections 6.1 and 6.2 describe the use of rank-based statistics to find characteristic words in LOB *vs*. Brown, and in male *vs*. female speech.

2.4 **t-*test***

The (unrelated) *t*-test, which operates on counts rather than on rank order of counts, could also be applied to frequency counts from two corpora divided into same-size samples. However the *t*-test is only valid where the data is normally distributed, which is not in general the case for word counts (see below). The Mann-Whitney test has the advantage of being non-parametric, that is, making no assumptions about the data obeying any particular distribution.

2.5 Mutual information

Another approach uses the Mutual Information (MI) statistic (Church and Hanks 1989). This simply takes the (log of the) ratio of the word's relative frequency in one corpus to its relative frequency in the joint corpus. In terms of Table 1:

$$MI_{\omega,X} = \log_2\left(\frac{a}{a+c} \times \frac{N}{a+b}\right)$$

This is an information theoretic measure (with relative frequencies serving as maximum likelihood estimators for probabilities) as distinct from one based in statistical hypothesis testing, and it makes no reference to hypotheses. Rather, it states how much information word *w* provides about corpus *X* (with respect to the joint corpus). It was introduced into language engineering as a measure for co-occurrence, where it specifies the information one word supplies about another.[5]

Church and Hanks state that MI is invalid for low counts, suggesting a threshold of 5. In contrast to χ^2, there is no notion in MI of evidence accumulating. MI, for our purposes, is a relation between two corpora and a word: if the corpora are held constant, it is usually rare words which give the highest MI. This contrasts with common words tending to have the highest χ^2 scores. Church and Hanks proposed MI as a tool to help lexicographers isolate salient co-occurring terms. Several years on, it is evident that MI overemphasises rare terms, relative to lexicographers' judgements of salience, while χ^2 correspondingly overemphasises common terms.

2.6 Log-likelihood (G^2*)*

Dunning (1993) is concerned at the invalidity of both χ^2 and MI where counts are low. The word he uses is 'surprise'; he wants to quantify how surprising various events are. He points out that rare events, such as the occurrence of many words and most bigrams in almost any corpus, play a large role in many language engineering tasks yet in these cases both MI and χ^2 statistics are invalid. He then presents the log-likelihood statistic, which gives an accurate measure of how surprising an event is even where it has occurred only once. For our contingency table, it can be calculated as:

$$\begin{aligned} G^2 = 2(&a \log(a) + b \log(b) + c \log(c) + d \log(d) \\ &- (a + b)\log(a + b) - (a + c)\log(a + c) \\ &- (b + d)\log(b + d) - (c + d)\log(c + d) \\ &+ (a + b + c + d)\log(a + b + c + d)) \end{aligned}$$

Daille (1995) determines empirically that it is an effective measure for finding terms. In relation to our simple case, of finding the most surprisingly frequent words in a corpus without looking at the internal structure of the corpus, G^2 is a mathematically well-grounded and accurate measure of surprisingness, and early indications are that, at least for low and medium frequency words such as those in Daille's study, it corresponds reasonably well to human judgements of distinctiveness.[6]

2.7 Fisher's exact test

Pedersen (1996) points out that log-likelihood approximates the probability of the data having occurred, given the null hypothesis, while there is a method for calculating it exactly: Fisher's Exact method. The machinery for computing it is available on various mathematical statistics packages. For very low counts, there is significant divergence between log-likelihood and the exact probability.

2.8 TF.IDF

The question, "Which words are particularly characteristic of a text?" is at the heart of information retrieval (IR). These are the words which will be the most appropriate key words and search terms. The general IR problem is to retrieve just the documents which are relevant to a user's query, from a database of many documents.[7]

A very simple method would be to recall just those documents containing one or more of the search terms. Since the user does not want to be swamped with 'potentially relevant' documents, this method is viable only if none of the search terms occur in many documents. Also, one might want to rank the documents, putting those containing more of the search terms at the top of the list. This suggests two modifications to the very simple method which give us the widely-used TF.IDF (term frequency by inverse document frequency) statistic (Salton 1989: 280 and references therein). Firstly a search term is of more value, the fewer documents it occurs in: IDF (inverse document frequency) is calculated, for each term in a collection, as the log of the inverse of the proportion of documents it occurs in. Secondly, a term is more likely to be important in a document, the more times it occurs in it: TF for a term and a document is simply the number of times the term occurs in the document.

Now, rather than simply registering a hit if there are any matches between a query term and a term in a document, we accumulate the TF.IDF scores for each match. We can then rank the hits, with the documents with the highest summed TF.IDF coming at the top of the list. This has been found to be a successful approach to retrieval (Robertson and Sparck Jones 1994).[8]

Two considerations regarding this scheme are:

- As described so far, it does not normalise for document length. In IR applications, TF is usually normalised by the length of the document. The discussion above shows that this is not altogether satisfactory. A single use of a word in a hundred-word document is far less noteworthy than ten uses of the word in a thousand-word document, but, if we normalise TF, they become equivalent.
- Very common words will be present in all documents. In this case, IDF = log 1 = 0 and TF.IDF collapses to zero. This point is not of particular interest to IR, as IR generally puts very common words on a stop list and ignores them, but it is a severe constraint on the generality of TF.IDF.

3 Probability distributions for words

As Church and Gale (1995) say, words come in clumps; unlike lightening, they often strike twice. Where a word occurs once in a text, you are substantially more likely to see it again than if it had not occurred once. Once a corpus is seen as having internal structure—that is, comprising distinct texts—the independence assumption is unsustainable.

Some words are ‘clumpier’ or ‘burstier’ than others; typically content words are clumpier than grammatical words. The ‘clumpiness’ or ‘burstiness’ of a particular word is an aspect of its behaviour which is important for many language-engineering and linguistic purposes, and in this section we sketch various approaches to modelling and measuring it.

The three probability distributions which are most commonly cited in the literature are the poisson, the binomial, and the normal. (Dunning refers to the multinomial, but for current purposes this is equivalent to the binomial.) The normal distribution is most often used as a convenient approximation to the binomial or poisson, where the mean is large, as justified by the Central Limit Theorem. For all three cases (poisson, binomial, or normal approximating to either) the distribution has just one parameter. Mean and variance do not vary independently: for the poisson they are equal, and for the binomial, if the expected value of the mean is p, the expected value of the variance is $p(1 - p)$.

To relate this to word-counting, consider the situation in which there are a number of same-length text samples. If words followed a poisson or binomial distribution then if a word occurred, on average, c times in a sample, the expected value for the variance of hits-per-sample is also c (or, in the binomial case, slightly less: the difference is negligible for all but the most common words). As various authors have found, this is not the case. Most of the time, the variance is greater than the mean. This was true for all but two of the 5,000 most common words in the BNC.[9]

3.1 Poisson mixtures

Following Mosteller and Wallace (1964), Gale and Church identify Poisson models as belonging to the right family of distributions for describing word frequencies, and then generalise so that the single poisson parameter is itself variable and governed by a probability distribution. A ‘poisson mixture’ distribution can then be designed with parameters set in such a way that, for a word of a given level of clumpiness and overall frequency in the corpus, the theoretical distribution models the number of documents it occurs in and the frequencies it has in those documents.

They list a number of ways in which clumping—or, more technically, ‘deviation from poisson’—can be measured. IDF is one, variance another, and they present three more. These empirical measures of clumpiness can then be used to set the second parameter of the poisson-mixture probability model. They show how these improved models can be put to work within a Bayesian approach to author identification.

3.2 Termight and Katz's model

Justeson and Katz (1995) and Katz (1996) present a more radical account of word distributions. The goal of their Termight system was to identify and index all the terms worth indexing in a document collection. They note the simple fact that a topical term—a term denoting what a document or part of a document was about—occurs repeatedly in documents about that topic. Termight identifies terms by simply finding all those words and phrases of the appropriate syntactic shape (noun phrases without subordinate clauses) which occur more than once in a document. Katz (1996) takes the theme further. He argues that word frequencies are not well modelled unless we take into account the linguistic intuition that a document is or is not about a topic, and that that means documents will tend to have zero occurrences of a term, or multiple occurrences. For terms, documents containing exactly one occurrence of a term will not be particularly common. Katz models word probability distributions with three parameters: first, the probability that it occurs in a document at all (document frequency), second, the probability that it will occur a second time in a document given that it has occurred once, and third, the probability that it will occur another time, given that it has already occurred *k* times (where $k > 1$). Thus the first parameter (which is most closely related to the pre-theoretical idea of a word being in more or less common usage) is independent of the second and third (which address how term-like the word is). Katz argues that, for true terms, the third parameter is very high, approaching unity: where a term has already occurred twice or more in a document, it is the topic of the document, so we should not be surprised if it occurs any number of further times.

Katz establishes that his model provides a closer fit to corpus data than a number of other models for word distributions that have been proposed, including Poisson mixtures.

3.3 Adjusted frequencies

The literature includes some proposals that word counts for a corpus should be adjusted to reflect clumpiness, with a word's frequency being adjusted downwards, the clumpier it is. The issues are described in Francis and Kučera (1982: 461–464). Francis and Kučera use a measure they call AF, attributed (indirectly) to J. Lanke of Lund University. It is defined as:

$$\mathrm{AF} = \left(\sum_{i=1}^{n} (d_i x_i)^{1/2} \right)^2$$

where the corpus is divided into *n* categories (which could be texts but, in Francis and Kučera's analysis, are genres, each of which contain numerous texts); d_i is the proportion of the corpus in that category; and x_i is the count for the word in the category.

Adjusting frequencies is of importance where the rank order is to be used directly for some purpose, for example, for choosing vocabulary for language-teaching, or in other circumstances where a single-parameter account of a word's distribution is wanted. Here, I mention it for purposes of completeness. A two- or three-parameter model as proposed by Church and Gale or Katz gives a more accurate picture of a word's behaviour than any one-parameter model.

4 Summary statistics for human interpretation

4.1 Content analysis

Content analysis is the social science tradition of quantitative analysis of texts to determine themes. It was particularly popular in the 1950s and 60s, a landmark being the General Enquirer (Stone *et al.* 1966), an early computerised system. Studies using the method have investigated a great range of topics, from analyses of propaganda and of changes in the tone of political communiqués over time, to psychotherapeutic interviews and the social psychology of interactions between management, staff and patients in nursing homes. The approach is taught in social science 'methods' courses, and used in political science (Fan 1988), psychology (Smith 1992) and market research (Wilson and Rayson 1993). The basic method is to:

- identify a set of 'concepts' which words might fall into, on the basis of a theoretical understanding of the situation;
- classify words into these concepts, to give a content analysis dictionary;
- take the texts (these will often be transcribed spoken material);
- for each text, count the number of occurrences of each concept.

One recent scheme, Minnesota Contextual Content Analysis (McTavish and Pirro 1990, MCCA), uses both a set of 116 concepts and an additional, more general level of 4 'contexts'. Norms for levels of usage of each concept come with the MCCA system, and scores for each concept are defined by taking the difference between the norm and the count for each concept-text pair (and dividing by the standard deviation of the concept across contexts). The concept scores are then directly comparable, between concepts and between texts. The approach is primarily descriptive: it provides a new way of describing texts, which it is then for the researcher to interpret and explain, so MCCA does nothing more with the concept scores.

It does however also provide the context scores. These serve several purposes, including

> [to] contribute to a kind of "triangulation", which would help to locate any potential text in relation to each of the "marker" contexts. (p. 250)

The validity of this kind of analysis is to be found in its predictive power. A content analysis study of open-ended conversations between husbands and wives was able to classify the couples as 'seeking divorce', 'seeking outside help', or 'coping' (McDonald and Weidetke 1979, quoted in McTavish and Pirro: 260).

4.2 Multi-dimensional analysis

A major goal of sociolinguistics is to identify the main ways in which language varies, from group to group and context to context. Biber (1988 and 1995) identifies the main dimensions of variation for English and three other languages using the following method:

- Gather a set of text samples to cover a wide range of language varieties;
- Enter them ("the corpus") into the computer;
- Identify a set of linguistic features which are likely to serve as discriminators for different varieties;
- Count the number of occurrences of each linguistic feature in each text sample;
- Perform a factor analysis (a statistical procedure) to identify which linguistic features tend to co-occur in texts. The output is a set of "dimensions", each of which carry a weighting for each of the linguistic features;
- Interpret each dimension, to identify what linguistic features, and what corresponding communicative functions, high-positive and high-negative values on the dimension correspond to.

For English, Biber identifies seven dimensions, numbered in decreasing order of significance (so dimension 1 accounts for the largest part of the non-randomness of the data, dimension 2, the next largest, etc.) The first he calls "Involved versus Informational Production". Texts getting high positive scores are typically spoken and typically conversations. Texts getting high negative scores are academic prose and official documents. The linguistic features with the highest positive weightings are "private" verbs (*assume*, *believe* etc.), *that*-deletion, contractions, present tense verbs, and second person pronouns. The linguistic features with the highest negative weightings are nouns, word length, prepositions, and type-token ratio. The two books cited above present the case for the explanatory power of the multidimensional approach.

Any text can be given a score for any dimension, by counting the numbers of occurrences of the linguistic features in the text, weighting, and summing.

The approach offers the possibility of "triangulation", placing a text within the space of English language variation, in a manner comparable to MCCA's context scores but using linguistic rather than social-science constructs, and using a statistical procedure rather than theory to identify the dimensions.

The methods described in Section 2 all take each word as a distinct data point, so each word defines a distinct dimension of a vector describing the differences. Biber first reduces the dimensionality of the space to a level where it is manageable by a human, and then offers contrasts between texts, and comments about what is distinctive about a text, in terms of these seven dimensions.[10] He thereby achieves some generalisation: he can describe how classes of features behave, whereas the other methods can only talk about the behaviour of individual words.

5 Discussion

Clearly, people working in the area of measuring what is distinctive about a text have had a variety of goals. Some have been producing figures primarily for further automatic manipulation, others have had human scrutiny in mind. Some have been comparing texts with texts, others, texts or corpora with corpora, and others again have been making comparisons with norms for the language at large. Some (Biber, Mosteller and Wallace) have looked more closely at high-frequency, form words; others (McTavish and Pirro, Dunning, Church and Gale) at medium and low frequency words.

The words in a corpus approximate to a Zipfian distribution, in which the product of rank order and frequency is constant. So, to a first approximation, the most common word in a corpus is a hundred times as common as the hundredth most common, a thousand times as common as the thousandth, and a million times as common as the millionth. This is a very skewed distribution. The few very common words have several orders of magnitude more occurrences than most others. The different ends of the range tend to have very different statistical behaviour. Thus, as we have seen, high-frequency words tend to give very high χ_2 error terms whereas very high MI scores come from low-frequency words. Variance, as we have seen, is almost always greater than the mean, and the ratio tends to increase with word frequency.

Linguists have long made a distinction approximating to the high/low frequency contrast: form words (or 'grammar words' or 'closed class words') *vs*. content words (or 'lexical words' or 'open class words'). The relation between the distinct linguistic behaviour, and the distinct statistical behaviour of high-frequency words is obvious yet intriguing.

It would not be surprising if we cannot find a statistic which works well for both high and medium-to-low frequency words. It is far from clear what a comparison of the distinctiveness of a very common word and a rare word would mean.

6 Finding words that vary across text-type: experiments

In this section I describe two experiments in which the Mann-Whitney ranks test is used to find words that are systematically used more in one text type than another.

6.1 LOB-Brown comparison

The LOB and Brown corpora both contain 2,000-word-long texts, so the numbers of occurrences of a word are directly comparable across all samples in both corpora. Had all 500 texts from each of LOB and Brown been used as distinct samples for the purposes of the ranks test, most counts would have been zero for all but very common words and the test would have been inapplicable. To make it applicable, it was necessary to agglomerate texts into larger samples. Ten samples for each corpus were used, each sample comprising 50 texts and 100,000 words. Texts were randomly assigned to one of these samples (and the experiment was repeated ten times, to give different random assignments, and the results averaged.) Following some experimentation, it transpired that most words with a frequency of 30 or more in the joint LOB and Brown had few enough zeroes for the test to be applicable, so tests were carried out for just those words, 5,733 in number.

The results were as follows. For 3,418 of the words, the null hypothesis was defeated (at a 97.5% significance level). In corpus statistics, this sort of result is not surprising. Few words comply with the null hypothesis, but then, as discussed above, the null hypothesis has little appeal: there is no intrinsic reason to expect any word to have exactly the same frequency of occurrence on both sides of the Atlantic. We are not in fact concerned with whether the null hypothesis holds: rather, we are interested in the words that are furthest from it. The minimum and maximum possible values for the statistic were 55 and 155, with a mean of 105, and we define a threshold for 'significantly British' (sB) of 75, and for 'significantly American' (sA), of 135.

The distribution curve was bell-shaped, one tail being sA and the other sB. There were 216 sB words and 288 sA words. They showed the same spread of frequencies as the whole population: the inter-quartile range for joint frequencies for the whole population was 44–147; for the sA it was 49–141 and for sB, 58–328. In contrast to the chi-square test, frequency-related distortion is avoided.

The sA and sB words were classified as in Table 4, according to a scheme close to that of Leech and Fallon (1992).

The items with distinct spellings occupied the extreme tails of the distribution. All other items were well distributed.

The first four categories serve as checks: had the items in these classes not been identified as sA and sB, the method would not have been working. It is the items in the 'others' category which are interesting. The three

Table 4 Classes of significantly British, and significantly American, words from the LOB/Brown comparison.

Mnemonic	*Example*	*sA*	*sB*
Spelling	color/colour; realise/realize	30	23
Equivalent	toward/towards; flat/apartment	15	17
Name	los, san, united; london, africa, alan	45	24
Cultural	negro, baseball, jazz; royal, chap, tea	38	26
Format	e, m, w	6	10
Other		154	116
Totals		288	216

highest-scoring sA 'others' are *entire*, *several* and *location*. None of these are identified as particularly American (or as having any particularly American uses) in any of four 1995 learners' dictionaries of English (LDOCE3, OALDCE5, CIDE, COBUILD2) all of which claim to cover both varieties of the language. Of course it does not follow from the frequency difference that there is a semantic or other difference that a dictionary should mention, but the 'others' list does provide a list of words for which linguists or lexicographers might want to examine whether there is some such difference.

6.2 Malelfemale conversation

The spoken, conversational part of the BNC was based on a demographic sample of the UK population, sampled for age, gender, region and social class. It is a rich resource for investigating how speech varies across these parameters. For details of its composition and collection see Crowdy (1993), Rayson, Leech, and Hodges (1997). Here we use it as a resource for exploring male/female differences, and for contrasting lists of most different words gathered using χ^2 with those gathered using the Mann-Whitney test.

Speaker turns where the gender of the speaker was available were identified, giving two corpora, one of male speech (M), the other, of female (F). Each corpus was divided into 25,000-word chunks. The order of texts in the BNC was retained in M and F, and the chunking took, first, the first 25,000 words, then the next 25,000, and so on, so the text relating to a single conversation would never be found in more than two chunks. The organisation of the BNC also ensured that a single speaker's words were unlikely to occur in more than two chunks. There were 31 M chunks and 50 F chunks. These chunks were then randomly combined into 150,000 word 'slices', giving five M slices and eight F slices. For each word with frequency of 20 or greater in M and F combined, the frequency in each slice was calculated, frequencies were ranked, and the Mann-Whitney statistic was calculated twice, once with the M slice always given the higher rank in cases of ties, once with the F, and the average taken.

Table 5 Mann-Whitney "consistently male" and "consistency female" words, compared with Rayson *et al.* lists.

M (MW)	*M (Ray)*	*F (MW)*	*F (Ray)*
a	a	Alice	and
against	ah	apples	because
ah	aye	child	Christmas
Ahhh	do	children	cos
bloke	er	clothes	did
can	four	cooking	going
Dad	fuck	curtains	had
Er	fucking	dish	he
fast	guy	her	her
itself	hundred	hers	him
mate	is	husband	home
one	mate	kitchen	I
quid	no	likes	lovely
record	number	Liz	me
right	of	lounge	mm
shoot	okay	made	n't
shot	one	morning	nice
slowly	quid	ours	oh
square	right	She	really
That	that	she	said
The	the	shopping	school
These	three	such	she
This	two	thinks	think
virtually	which	thought	thought
way	yeah	wardrobe	to

The 25 words which are most consistently more common in M and F are presented in Table 5, alongside the equivalent lists from Rayson *et al.*[11] All lists have been alphabeticised, for ease of comparison. Of the 25 commonest words in the joint corpus (unnormalised for case), twelve are in Rayson *et al.*'s lists, whereas just one (*a*) is in either of the Mann-Whitney lists. The Rayson *et al.* lists display a bias towards high-frequency items which is not generally useful for corpus linguistics and which the Mann-Whitney lists do not share.

7 Measures for similarity and homogeneity

The explorations described above have skirted around the issue of corpus similarity, looking at particular ways in which corpora are notably different. In the remainder of the paper, we look directly at corpus similarity, and particularly at how it might be measured.

What are the constraints on a measure for corpus similarity? The first is simply that its findings correspond to unequivocal human judgements. It must match our intuition that, for instance, a corpus of syntax papers is more like one of semantics papers than one of shopping lists. The constraint is key but is weak. Direct human intuitions on corpus similarity are not easy to come by, firstly, because large corpora, unlike coherent texts, are not the sorts of things people read, so people are not generally in a position to have any intuitions about them. Secondly, a human response to the question, "how similar are two objects", where those objects are complex and multi-dimensional, will themselves be multi-dimensional: things will be similar in some ways and dissimilar in others. To ask a human to reduce a set of perceptions about the similarities and differences between two complex objects to a single figure is an exercise of dubious value.

This serves to emphasise an underlying truth: corpus similarity is complex, and there is no absolute answer to "is Corpus 1 more like Corpus 2 than Corpus 3?". All there are are possible measures which serve particular purposes more or less well. Given the task of costing the customisation of an NLP system, produced for one domain, to another, a corpus similarity measure is of interest insofar as it predicts how long the porting will take. It could be that a measure which predicts well for one NLP system, predicts badly for another. It can only be established whether a measure correctly predicts actual costs, by investigating actual costs.[12]

Having struck a note of caution, we now proceed on the hypothesis that there is a single measure which corresponds to pre-theoretical intuitions about 'similarity' and which is a good indicator of many properties of interest— customisation costs, the likelihood that linguistic findings based on one corpus apply to another, etc. We would expect the limitations of the hypothesis to show through at some point, when different measures are shown to be suited to different purposes, but in the current situation, where there has been almost no work on the question, it is a good starting point.

7.1 Similarity and homogeneity

How homogeneous is a corpus? The question is both of interest in its own right, and is a preliminary to any quantitative approach to corpus similarity. In its own right, because a sublanguage corpus, or one containing only a specific language variety, has very different characteristics to a general corpus (Biber 1993b) yet it is not obvious how a corpus's position on this scale can be assessed. It is of interest as a preliminary to measuring corpus similarity, because it is not clear what a measure of similarity would mean if a homogeneous corpus (of, e.g., software manuals) was being compared with a heterogeneous one (e.g., Brown). Ideally, the same measure can be used for similarity and homogeneity, as then, Corpus 1/Corpus 2 distances will be directly comparable with heterogeneity (or "within-corpus distances") for Corpus 1 and Corpus 2. This is the approach adopted here.

Table 6 Interactions between homogeneity and similarity: a similarity measure can only be interpreted with respect to homogeneity.

	Corpus 1	*Corpus 2*	*Distance*	*Interpretation*
1	equal	equal	equal	same language variety/ies
2	equal	equal	high	different language varieties
3	high	high	low	impossible
4	high	low	high	corpus 2 is homogeneous and falls within the range of corpus 1
5	high	low	higher	corpus 2 is homogeneous and falls outside the range of corpus 1
6	low	low	slightly higher	similar varieties
7	high	high	slightly higher	overlapping; share some varieties

Like corpus distance, heterogeneity is multi-dimensional, and, in looking for a single measure for it, we are inevitably ignoring much. Two aspects which will be combined within any single measure are the heterogeneity that arises from a corpus comprising texts of different types, and the heterogeneity that arises from a single text type, where, for instance, a wide variety of grammar and lexis is used. These are clearly different things, and it would be desirable to develop measures which address them separately, but that remains as future work.

Some of the permutations of homogeneity and similarity scores, and their interpretations, are shown in Table 6. In the table, *high* scores means a large distance between corpora, or large within-corpus distances, so the corpus is heterogeneous or the corpora are dissimilar; *low*, that the distances are low, so the corpus is homogeneous, or the corpora are similar. (Thus we have a distance measure rather than a similarity measure, which would have opposite polarity.) *High*, *low* and *equal* are relative to the other columns in the same row. In row 1, all three scores are equal, implying that both corpora are of the same text type. In row 2, 'equal' in the first two columns reads that the within-corpus distance (homogeneity) of Corpus 1 is roughly equal to the within-corpus distance of Corpus 2, and 'high' in the Distance column reads that the distance between the corpora is substantially higher than these within-corpus distances. Thus a comparison between the two corpora is straightforward to interpret, since the two corpora do not differ radically in their homogeneity, and the outcome of the comparison is that they are of markedly different language varieties.

Not all combinations of homogeneity and similarity scores are logically possible. For example, two corpora cannot be much more similar to each other than either is to itself (row 3).

Rows 4 and 5 indicate two of the possible outcomes when a relatively heterogeneous corpus (corpus 1) is compared with a relatively homogeneous one (corpus 2). It is not possible for the distance between the corpora to be much lower than the heterogeneity of the more heterogeneous corpus 1. If the distance is roughly equal to corpus 1 heterogeneity, the interpretation is that corpus 2 falls within the range of corpus 2; if it is higher, it falls outside.

The last two rows point to the differences between general corpora and specific corpora. High and low values in the first two columns are to be interpreted relative to norms for the language. Particularly high within-corpus distance scores will be for general corpora, which embrace a number of language varieties. Corpus similarity between general corpora will be a matter of whether all the same language varieties are represented in each corpus, and in what proportions. Low within-corpus distance scores will typically relate to corpora of a single language variety, so here, scores may be interpreted as a measure of the distance between the two varieties.

7.2 Related work

There is very little work which explicitly aims to measure similarity between corpora. The one clearly relevant item is Johansson anf Hofland (1989), which aims to find which genres, within the LOB corpus, most resemble each other. They take the 89 most common words in the corpus, find their rank within each genre, and calculate the Spearman rank correlation statistic ('spearman'):

Rose, Haddock, and Tucker (1997) explore how performance of a speech recognition system varies with the size and specificity of the training data used to build the language model. They have a small corpus of the target text type, and experiment with 'growing' their seed corpus by adding more same-text-type material. They use Spearman and log-likelihood (Dunning 1993) as measures to identify same-text-type corpora. Spearman is evaluated below.

Sekine (1997) explores the domain dependence of parsing. He parses corpora of various text genres and counts the number of occurrences of each subtree of depth one. This gives him a subtree frequency list for each corpus, and he is then able to investigate which subtrees are markedly different in frequency between corpora. Such work is highly salient for customising parsers for particular domains. Subtree frequencies could readily replace word frequencies for the frequency-based measures below.

In information-theoretic approaches, perplexity is a widely-used measure. Given a language model and a corpus, perplexity "is, crudely speaking, a measure of the size of the set of words from which the next word is chosen given that we observe the history of [. . .] words" (Roukos 1996). Perplexity is most often used to assess how good a language modelling strategy is, so is used with the corpus held constant. Achieving low perplexity in the language model is critical for high-accuracy speech recognition, as it means there are fewer high-likelihood candidate words for the speech signal to be compared with.

Perplexity can be used to measure a property akin to homogeneity if the language modelling strategy is held constant and the corpora are varied. In this case, perplexity is taken to measure the intrinsic difficulty of the speech recognition task: the less constraint the domain corpus provides on what the next word might be, the harder the task. Thus Roukos (1996) presents a table in which different corpora are associated with different perplexities. Perplexity measures are evaluated below.

8 "Known-Similarity Corpora"

Proposing measures for corpus similarity is relatively straightforward: determining which measures are good ones, is harder. To evaluate the measures, it would be useful to have a set of corpora where similarities were already known. In this section, we present a method for producing a set of "Known-Similarity Corpora" (KSC).

A KSC-set is built as follows: two reasonably distinct text types, A and B, are taken. Corpus 1 comprises 100% A; Corpus 2, 90% A and 10% B; Corpus 3, 80% A and 20% B; and so on. We now have at our disposal a set of fine-grained statements of corpus similarity: Corpus 1 is more like Corpus 2 than Corpus 1 is like Corpus 3. Corpus 2 is more like Corpus 3 than Corpus 1 is like Corpus 4, etc. Alternative measures can now be evaluated, by determining how many of these 'gold standard judgements' they get right. For a set of n Known-Similarity Corpora there are

$$\sum_{i=1}^{n}(n-i)\left(\frac{i(i+1)}{2}-1\right)$$

gold standard judgements (see Appendix 1 for proof) and the ideal measure would get all of them right. Measures can be compared by seeing what percentage of gold standard judgements they get right.

Two limitations on the validity of the method are, first, there are different ways in which corpora can be different. They can be different because each represents one language variety, and these varieties are different, or because they contain different mixes, with some of the same varieties. The method only directly addresses the latter model.

Second, if the corpora are small and the difference in proportions between the corpora is also small, it is not clear that all the 'gold standard' assertions are in fact true. There may be a finance supplement in one of the copies of the *Guardian* in the corpus, and one of the copies of *Accountancy* may be full of political stories: perhaps, then, Corpus 3 *is* more like Corpus 5 than Corpus 4. It is necessary to address this by selecting the two text types with care so they are similar enough so the measures are not all 100% correct yet dissimilar enough to make it likely that all gold-standard judgements are true, and by ensuring there is enough data and enough KSC-sets so that oddities of individual corpora do not obscure the picture of the best overall measure.

9 Experiment to evaluate measures

We now describe an experiment in which KSC-sets were used to evaluate four candidate measures for corpus similarity.

9.1 The measures

All the measures use spelled forms of words. None make use of linguistic theories. The comment has been made that lemmas, or word senses, or syntactic constituents, were more appropriate objects to count and perform computations on than spelled forms. This would in many ways be desirable. However there are costs to be considered. To count, for example, syntactic constituents requires, firstly, a theory of what the syntactic constituents are; secondly, an account of how they can be recognised in running text; and thirdly, a program which performs the recognition. Shortcomings or bugs in any of the three will tend to degrade performance, and it will not be straightforward to allocate blame. Different theories and implementations are likely to have been developed with different varieties of text in focus, so the degradation may well affect different text types differently. Moreover, practical users of a corpus-similarity measure cannot be expected to invest energy in particular linguistic modules and associated theory. To be of general utility, a measure should be as theory-neutral as possible.

In these experiments we consider only raw word-counts. Two word frequency measures were considered. For each, the statistic did not dictate which words should be compared across the two corpora. In a preliminary investigation we had experimented with taking the most frequent 10, 20, 40 . . . 640, 1280, 2560, 5120 words in the union of the two corpora as data points, and had achieved the best results with 320 or 640. For the experiments below, we used the most frequent 500 words.

Both word-frequency measures can be directly applied to pairs of corpora, but only indirectly to measure homogeneity. To measure homogeneity:

1 divide the corpus into ‘slices’;
2 create two subcorpora by randomly allocating half the slices to each;
3 measure the similarity between the subcorpora;
4 iterate with different random allocations of slices;
5 calculate mean and standard deviation over all iterations.

Wherever similarity and homogeneity figures were to be compared, the same method was adopting for calculating corpus similarity, with one subcorpus comprising a random half of Corpus 1, the other, a random half of Corpus 2.

9.1.1 Spearman rank correlation co-efficient

Ranked wordlists are produced for Corpus 1 and Corpus 2. For each of the *n* most common words, the difference in rank order between the two corpora is taken. The statistic is then the normalised sum of the squares of these differences,

$$1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

9.1.2 Comment

Spearman is easy to compute and is independent of corpus size: one can directly compare ranked lists for large and small corpora. However the following objection seemed likely to play against it. For very frequent words, a difference of rank order is highly significant: if *the* is the most common word in corpus 1 but only third in corpus 2, this indicates a high degree of difference between the genres. But at the other end of the scale, the opposite is the case: if *bread* is in 400th position in the one corpus and 500th in the other, this is of no significance; yet Spearman counts the latter as far more significant than the former.

χ^2

For each of the *n* most common words, we calculate the number of occurrences in each corpus that would be expected if both corpora were random samples from the same population. If the size of corpora 1 and 2 are N_1, N_2 and word w has observed frequencies $o_{\omega,1}$, $o_{\omega,2}$, then expected value $e_{\omega,1} = \frac{N_1 \times (o_{\omega,1} + o_{\omega,2})}{N_1 + N_2}$ and $e_{\omega,2} = \frac{N_2 \times (o_{\omega,1} + o_{\omega,2})}{N_1 + N_2}$; then

$$x^2 = \sum \frac{(o - e)^2}{e}$$

9.1.3 Comment

The inspiration for the statistic comes from the χ^2-test for statistical independence. As shown above, the statistic is not in general appropriate for hypothesis-testing in corpus linguistics: a corpus is never a random sample of words, so the null hypothesis is of no interest. But once divested of the hypothesis-testing link, χ^2 is suitable. The $(o - e)^2/e$ term gives a measure of the difference in a word's frequency between two corpora, and the measure tends to increase slowly with word frequency in a way that is compatible with the intuition that higher-frequency words are more significant in assessments of corpus similarity that lower-frequency ones.

The measure does not directly permit comparison between corpora of different sizes.

9.1.4 Perplexity and cross-entropy

From an information-theoretic point of view, *prima facie*, entropy is a well defined term capturing the informal notion of homogeneity, and the cross-entropy between two corpora captures their similarity. Entropy is not a quantity that can be directly measured. The standard problem for statistical language modelling is to aim to find the model for which the cross-entropy of the model for the corpus is as low as possible. For a perfect language model, the cross-entropy would be the entropy of the corpus (Church and Mercer 1993, Charniak 1993).

With language modelling strategy held constant, the cross-entropy of a language model (LM) trained on Corpus 1, as applied to Corpus 2, is a similarity measure. The cross-entropy of the LM based on nine tenths of Corpus 1, as applied to the other 'held-out' tenth, is a measure of homogeneity. We standardised on the 'tenfold cross-validation' method for measures of both similarity and homogeneity: that is, for each corpus, we divided the corpus into ten parts[13] and produced ten LMs, using nine tenths and leaving out a different tenth each time. (Perplexity is the log of the cross-entropy of a corpus with itself: measuring homogeneity as self-similarity is standard practice in information theoretic approaches.)

To measure homogeneity, we calculated the cross-entropy of each of these LMs as applied to the left-out tenth, and took the mean of the ten values. To measure similarity, we calculated the cross-entropy of each of the Corpus 1 LMs as applied to a tenth of Corpus 2 (using a different tenth each time). We then repeated the procedure with the roles of Corpus 1 and Corpus 2 reversed, and took the mean of the 20 values.

All LMs were trigram models. All LMs were produced and calculations performed using the CMU/Cambridge toolkit (Rosenfeld 1995).

The treatment of words in the test material but not in the training material was critical to our procedure. It is typical in the language modelling community to represent such words with the symbol UNK, and to calculate the probability for the occurrence of UNK in the test corpus using one of three main strategies.

Closed vocabulary The vocabulary is defined to include all items in training and test data. Probabilities for those items that occur in training but not test data, the 'zerotons', are estimated by sharing out the probability mass initially assigned to the singletons and doubletons to include the zerotons.

Open, type 1 The vocabulary is chosen independently of the training and test data, so the probability of UNK may be estimated by counting the occurrence of unknown words in the training data and dividing by N (the total number of words).
Open, type 2 The vocabulary is defined to include all and only the training data, so the probability of UNK cannot be estimated directly from the training data. It is estimated instead using the discount mass created by the normalisation procedure.

All three strategies were evaluated.

9.2 Data

All KSC sets were subsets of the British National Corpus (BNC). A number of sets were prepared as follows.

For those newspapers or periodicals for which the BNC contained over 300,000 running words of text, word frequency lists were generated and similarity and homogeneity were calculated (using χ^2; results are shown in Appendix 2.) We then selected pairs of text types which were moderately distinct, but not too distinct, to use to generate KSC sets. (In initial experiments, more highly distinct text types had been used, but then both Spearman and χ^2 had scored 100%, so 'harder' tests involving more similar text types were selected.)

For each pair *a* and *b*, all the text in the BNC for each of *a* and *b* was divided into 10,000-word tranches. These tranches were randomly shuffled and allocated as follows:

first 10 of *a*	into	b0a
next 9 of *a*, first 1 of *b*	into	b1a
next 8 of *a*, next 2 of *b*	into	b2a
next 7 of *a*, next 3 of *b*	into	b3a
. . .		

until either the tranches of *a* or *b* ran out, or a complete eleven-corpus KSC-set was formed. A sample of KSC sets are available is the web.[14] There were 21 sets containing between 5 and 11 corpora. The method ensured that the same piece of text never occurred in more than one of the corpora in a KSC set.

The text types used were:

> *Accountancy* (acc); *The Art Newspaper* (art); *British Medical Journal* (bmj); *Environment Digest* (env); *The Guardian* (gua); *The Scotsman* (sco); and *Today* ('low-brow' daily newspaper, tod).

To the extent that some text types differ in content, whereas others differ in style, both sources of variation are captured here. *Accountancy* and *The*

Art Newspaper are both trade journals, though in very different domains, while *The Guardian* and *Today* are both general national newspapers, of different styles.

9.3 Results

For each KSC-set, for each gold-standard judgement, the 'correct answer' was known, e.g., "the similarity 1,2 is greater than the similarity 0,3". A given measure either agreed with this gold-standard statement, or disagreed. The percentage of times it agreed is a measure of the quality of the measure. Results for the cases where all four measures were investigated are presented in Table 7.

Table 7 Comparison of four measures.

	spear	χ^2	*closed*	*type 1*	*type 2*
KSC-set					
acc_gua	93.33	91.33	82.22	81.11	80.44
art_gua	95.60	93.03	84.00	83.77	84.00
bmj_gua	95.57	97.27	88.77	89.11	88.77
env_gua	99.65	99.31	87.07	84.35	86.73

The word frequency measures outperformed the perplexity ones. It is also salient that the perplexity measures required far more computation: ca. 12 hours on a Sun, as opposed to around a minute.

Spearman and χ^2 were tested on all 21 KSC-sets, and χ^2 performed better for 13 of them, as shown in Table 8.

Table 8 Spearman/χ^2 comparison on all KSCs.

	spear	χ^2	*tie*	*total*
Highest score	5	13	3	21

The difference was significant (related *t*-test: $t = 4.47$, 20DF, significant at 99.9% level). χ^2 was the best of the measures compared.

10 Conclusions and further work

Corpus linguistics lacks a vocabulary for talking quantitatively about similarities and differences between corpora. The paper aims to begin to meet this need.

One way of describing differences between corpora is by highlighting the words which have consistently been used more in the one corpus that the other. This is a matter that has been looked at by a variety of authors, in a variety of disciplines, and the methods which have been developed are reviewed. At a first pass, it may appear that the χ^2-test is suitable for identifying the words which are used most differently, but we show that this is not the case. A more suitable test is the Mann-Whitney ranks test. This test was used to compare the Brown and LOB corpora, and the male and female conversational components of the BNC; the results are presented.

We then address the more ambitious goal of measuring corpus similarity. We argue that corpus linguistics is in urgent need of such a measure: without one, it is very difficult to talk accurately about the relevance of findings based on one corpus, to another, or to predict the costs of porting an application to a new domain. We note that corpus similarity is complex and multifaceted, and that different measures might be required for different purposes. However, given the paucity of other work in the field, at this stage it is enough to seek a single measure which performs reasonably.

The Known-Similarity Corpora method for evaluating corpus-similarity measures was presented, and measures discussed in the literature were compared using it. For the corpus-size used and this approach to evaluation, χ^2 and Spearman both performed better than any of three cross-entropy measures. These measures have the advantage that they are cheap and straightforward to compute. χ^2 outperformed Spearman.

Thus χ^2 is presented as a suitable measure for comparing corpora, and is shown to be the best measure of those tested. It can be used for measuring the similarity of a corpus to itself, as well as the similarity of one corpus to another, and this feature is valuable as, without self-similarity as a point of reference, a measure of similarity between corpora is uninterpretable.

There are, naturally, some desiderata it does not meet. Unlike cross-entropy, it is not rooted in a mathematical formalism which provides the prospect of integrating the measure with some wider theory. Also, an ideal measure would be scale-independent, supporting the comparison of small and large corpora. This is an area for future work.

Acknowledgements

This work has been supported by the EPSRC, Grants K18931 and M54971 and a Macquarie University Research Grant. I am particularly grateful to Tony Rose, who undertook the language modelling and cross-entropy calculations. The work has also benefited from comments from Harald Baayen, Ted Dunning and Mark Lauer.

Appendix 1

The proof is based on the fact that the number of similarity judgements is the triangle number of the number of corpora in the set (less one), and that each new similarity judgement introduces a triangle number of gold standard judgements (once an ordering which roles out duplicates is imposed on gold standard judgements).

- A KSC set is ordered according to the proportion of text of type 1. Call the corpora in the set 1 . . . *n*.
- A similarity judgement ('sim') between *a* and *b* (*a*, *b*) compares two corpora. To avoid duplication, we stipulate that $a < b$. Each sim is associated with a number of steps of difference between the corpora: dif(*a*, *b*) = $b - a$.
- A gold standard judgement ('gold') compares two sims; there is only a gold between *a*, *b* and *c*, *d* if $a < b$ and $c < d$ (as stipulated above) and also if $a <= c$, $b >= d$, and not ($a = c$ and $b = d$). Each four-way comparison can only give rise to zero or one gold, as enforced by the ordering constraints. Each gold has a difference of difs ('difdif') of $(b - a) - (d - c)$ (so, if we compare 3, 5 with 3, 4, difdif = 1, but where we compare 2, 7 with 3, 4, difdif = 4). difdif(*X*, *Y*) = dif(*X*) – dif(*Y*).
- Adding an *n*th corpus to a KSC set introduces $n - 1$ sims. Their difs vary from 1 (for $(n - 1)$, *n*) to $n - 1$ (for 1, *n*).
- The number of golds with a sim of dif m as first term is a triangle number less one, $\sum_{i=2}^{m} i$ or $\frac{\mathrm{m}(\mathrm{m}+1)}{2} - 1$. For example, for 2, 6 (dif = 4) there are 2 golds of difdif 1 (e.g. with 2, 5 and 3, 6), 3 of difdif 2 (with 2, 4, 3, 5, 4, 6), and 4 of difdif 3 (with 2, 3, 3, 4, 4, 5, 5, 6).
- With the addition of the nth corpus, we introduce $n - 1$ sims with difs from 1 to $n - 1$, so we add $\sum_{i=1}^{n-1} \frac{i(i+1)}{2} - 1$ golds. For the whole set, there are $\sum_{i=1}^{n} \sum_{j=1}^{i-1} \frac{j(j+1)}{2} - 1$ and collecting up repeated terms gives $\sum_{i=1}^{n} (n - i)\left(\frac{i(i+1)}{2} - 1\right)$

Appendix 2

Subcorpora were extracted from the BNC for all document sources for which there was a large quantity of data. Table l0 shows the distance and heterogeneity scores (based on the 500 most common words in the joint corpus) for twelve of these sources (as described in Table 9). The numbers presented are χ^2-scores normalised by the degrees of freedom ("Chi-by-degrees-of-freedom" or CBDF).

Table 9 Corpora for first experiment.

Short	*Title*	*Description*
GUA	The Guardian	Broadsheet national newspaper
IND	The Independent	Broadsheet national newspaper
DMI	Daily Mirror	Tabloid national newspaper
NME	New Musical Express	Weekly pop/rock music magazine
FAC	The Face	Monthly fashion magazine
ACC	Accountancy	Accountancy periodical
DNB	Dictionary of National Biography	Comprises short biographies
HAN	Hansard	Proceedings of Parliament
BMJ	British Medical Journal	Academic papers on medicine
GRA	Computergram	Electronic computer-trade newsletter

Table 10 CBDF homogeneity and similarity scores for twelve 200,000-word corpora.

	ACC	*ART*	*BMJ*	*DMI*	*DNB*	*ENV*	*FAC*	*GRA*	*GUA*	*HAN*	*IND*	*NME*
ACC	4.62											
ART	21.40	3.38										
BMJ	20.16	23.50	8.08									
DMI	21.56	26.19	32.08	2.47								
DNB	40.56	30.07	40.14	35.15	1.86							
ENV	22.68	23.10	28.12	34.65	41.50	2.60						
FAC	20.49	25.14	31.14	7.76	36.92	36.93	3.43					
GRA	27.75	29.96	33.50	31.40	45.26	28.96	34.35	2.20				
GUA	14.06	18.37	22.68	11.41	31.06	23.24	12.04	32.25	3.92			
HAN	24.13	33.76	33.00	32.14	52.25	32.03	31.23	36.21	22.62	3.65		
IND	12.76	17.83	22.96	13.96	30.10	21.69	14.45	28.06	4.11	23.27	4.44	
NME	21.18	25.99	30.05	9.77	39.41	34.77	5.84	31.39	15.09	33.25	16.56	3.10

Note that the lowest distances are between the Guardian and the Independent (two broadsheet newspapers) and between NME and The Face (a pop/rock music magazine and a style magazine).

Heterogeneity scores are generally lower than distance scores. The Dictionary of National Biography, which comprises paragraph-long biographies in a fairly standardised form, is the most homogeneous, but is very unlike any other source. The most academic source, British Medical Journal, is by far the least homogeneous.

Notes

[1] Alternative names for the field (or a closely related one) are "empirical linguistics" and "data-intensive linguistics". By using an adjective rather than a noun, these seem not to assert that the corpus is an object of study. Perhaps it is equivocation about what we can say about corpora that has led to the coining of the alternatives.

[2] Provided all expected values are over a threshold of 5.

[3] See http://info.ox.ac.uk/bnc

[4] In this case the null hypothesis is true, so the average value of the sum of the error terms over the four cells of the contingency table is 1 (from the definition of the χ^2 distribution). Of the four cells, the two error-terms associated with the absence of the word (cells *c* and *d* in Table 1) will be vanishingly small, as E is large—almost as large as the number of words in the corpus—whereas $(|O - E| - 0.5)^2$ is small, so the result of dividing it by E is vanishingly small. The two cells corresponding to the presence of the word (cells *a* and *b* in Table 1) will both have the same average value, since, by design, the two corpora are the same size. Thus the four-way sum is effectively shared between cells *a* and *b*, so the average value of each is 0.5.

[5] The usage of the term "mutual information" within information theory is different: $MI(X;Y) = \Sigma_{x,y} \log \frac{p(x,y)}{p(x)p(y)}$. However, in language engineering, the Church–Hanks definition has been widely adopted so here, MI will refer to that simpler term.

[6] Several corpus interface packages provide functionality for computing one or more of these statistics. For example, WordSmith (Scott 1997 and Sardinha

1996) and the associated KeyWords tool allows the user to generate lists using Mutual Information, chi-square and log-likelihood.

[7] We assume full-text searching. Also, issues such as stemming and stop lists are not considered, as they do not directly affect the statistical considerations.

[8] They also provide a 'tuning constant' for adjusting the relative weight given to TF and IDF to optimise performance.

[9] Figures based on the standard-document-length subset of the BNC described above.

[10] Reducing the dimensionality of the problem has also been explored in IR: see Schütze and Pederson (1995), Dumais *et al.* (1988).

[11] Rayson *et al.* The comparisons are normalised for case, so this is one point at which direct comparison is not possible.

[12] Cf. Ueberla (1997), who looks in detail at the appropriateness of perplexity as a measure of task difficulty for speech recognition, and finds it wanting.

[13] For the KSC corpora, we ensured that each tenth had an appropriate mix of text types, so that, e.g. each tenth of a corpus comprising 70% Guardian, 30% BMJ, also comprised 70% Guardian, 30% BMJ.

[14] http://www.itri.bton.ac.uk/~Adam.Kilgarriff/KSC/

References

Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. 1990. "Methodological issues regarding corpus-based analyses of linguistic variation." *Literary and Linguistic Computing* 5: 257–269.

Biber, D. 1993a. "Representativeness in corpus design." *Literary and Linguistic Computing* 8: 243–257.

Biber, D. 1993b. "Using register-diversified corpora for general language studies." *Computational Linguistics* 19(2): 219–242.

Biber, D. 1995. *Dimensions in Register Variation*. Cambridge: Cambridge University Press.

Charniak, E. 1993. *Statistical Language Learning*. Cambridge, Mass, MIT Press.

Church, K. and W. Gale. 1995. "Poisson mixtures." *Journal of Natural Language Engineering* 1(2): 163–190.

Church, K. and P. Hanks. 1989. "Word association norms, mutual information and lexicography." *ACL Proceedings, 27th Annual Meeting*, 76–83. Vancouver: ACL.

Church, K. W. and R. L. Mercer. 1993. "Introduction to the special issue on computational linguistics using large corpora." *Computational Linguistics* 19(1): 1–24.

Crowdy, S. 1993. "Spoken corpus design." *Literary and Linguistic Computing* 8: 259–265.

Daille, B. 1995. *Combined approach for terminology extraction: lexical statistics and linguistic filtering*. Technical Report 5. Lancaster University: UCREL.

Dumais, S., G. Furnas, T. Landauer, S. Deerwester, and R. Harshman. 1988. "Using latent semantic analysis to improve access to textual information." *Proceedings of CHI '88*, 281–285. Washington DC: ACM.

Dunning, T. 1993. "Accurate methods for the statistics of surprise and coincidence." *Computational Linguistics* 19(1): 61–74.

Fan, D. P. 1988. *Predictions of public opinion from the mass media: computer content analysis and mathematical modeling*. New York: Greenwood Press.

Francis, W. N. and H. Kučera. 1982. *Frequency Analysis of English Usage: lexicon and grammar*. Boston: Houghton Mifflin.

Hofland, K. and S. Johansson (eds). 1982. *Word Frequencies in British and American English*. Bergen: The Norwegian Computing Centre for the Humanities.

Johansson, S. and K. Hofland (eds). 1989. *Frequency Analysis of English vocabulary and grammar, based on the LOB corpus*. Oxford: Clarendon.

Justeson, J. S. and S. M. Katz. 1995. "Technical terminology: some linguistic properties and an algorithm for identification in text." *Natural Language Engineering* 1(1): 9–27.

Katz, S. 1996. "Distribution of content words and phrases in text and language modelling." *Natural Language Engineering* 2(1): 15–60.

Leech, G. and R. Fallon. 1992. "Computer corpora—what do they tell us about culture?" *ICAME Journal* 16: 29–50.

McDonald, C. and B. Weidetke. 1979. *Testing marriage climate*. Masters thesis. Ames, Iowa: Iowa State University.

McTavish, D. G. and E. B. Pirro. 1990. "Contextual content analysis." *Quality and Quantity* 24: 245–265.

Mosteller, F. and D. L. Wallace. 1964. *Applied Bayesian and Classical Inference—The Case of The Federalist Papers*. Springer Series in Statistics. London: Springer-Verlag.

Owen, F. and R. Jones. 1977. *Statistics*. Stockport: Polytech Publishers.

Pedersen, T. 1996. "Fishing for exactness." *Proceedings of the Conference of South-Central SAS Users Group*. Texas: SAS Users Group. Also available from CMP-LG E-Print Archive as #9608010.

Rayson, P., G. Leech, and M. Hodges. 1997. "Social differentiation in the use of English vocabulary: some analysis of the conversational component of the British National Corpus." *International Journal of Corpus Linguistics* 2(1): 133–152.

Robertson, S. E. and K. Sparck Jones. 1994. "Simple, proven approaches to text retrieval." Technical Report 356. Cambridge: Cambridge University.

Rose, T., N. Haddock, and R. Tucker. 1997. "The effects of corpus size and homogeneity on language model quality." *Proceedings of ACL SIGDAT workshop on Very Large Corpora*, 178–191. Beijing and Hong Kong: ACL.

Rosenfeld, R. 1995. "The CMU Statistical Language Modelling Toolkit and its use in the 1994 ARPA CSR Evaluation." *Proceedings of Spoken Language Technology Workshop*. Austin, Texas: Arpa.

Roukos, S. 1996. *Language Representation*, chapter 1.6. National Science Foundation and European Commission, www.cse.ogi/CSLU/HLTsurvey.html.

Salton, G. 1989. *Automatic Text Processing*. London: Addison-Wesley.

Berber Sardinha, T. 1996. WordSmith tools. *Computers and Texts* 12: 19–21.

Schütze, H. and J. O. Pederson. 1995. "Information retrieval based on word senses." *Proceedings of ACM Special Interest Group on Information Retrieval*, 161–175. Las Vegas: ACM.

Scott, M. 1997. "PC analysis of key words—and key key words." *System* 25: 233–245.

Sekine, S. 1997. "The domain dependence of parsing." *Proceedings of Fifth Conference on Applied Natural Language Processing*, 96–102. Washington DC: ACL.

Smith, C. P. 1992. *Motivation and personality: handbook of thematic content analysis*. Cambridge: Cambridge University Press.

Stone, P. J., D. C. Dunphy, M. S. Smith, and D. M. Ogilvie. 1966. *The General Enquirer: A Computer approach to content analysis*. Cambridge, Mass: MIT Press.

Ueberla, J. 1997. *Towards an improved performance measure for language models*. Technical Report DERA/CIS/CIS5/TR97426, DERA, cmp-lg/9711009.

Wilson, A. and P. Rayson. 1993. "Automatic Content Analysis of Spoken Discourse." In: C. Souter and E. Atwell (eds.), *Corpus Based Computational Linguistics*,