Language is never, ever, ever, random

ADAM KILGARRIFF

Abstract

Language users never choose words randomly, and language is essentially non-random. Statistical hypothesis testing uses a null hypothesis, which posits randomness. Hence, when we look at linguistic phenomena in corpora, the null hypothesis will never be true. Moreover, where there is enough data, we shall (almost) always be able to establish that it is not true. In corpus studies, we frequently do have enough data, so the fact that a relation between two phenomena is demonstrably non-random, does not support the inference that it is not arbitrary. We present experimental evidence of how arbitrary associations between word frequencies and corpora are systematically non-random. We review literature in which hypothesis testing has been used, and show how it has often led to unhelpful or misleading results.

Keywords: hypothesis testing; language corpora; randomness assumption.

1. Introduction

Any two phenomena might or might not be related. The range of possibilities is that the association is Random, Arbitrary, Motivated or Predictable (R, A, M, P). The bulk of linguistic questions concern the distinction between A and M. A linguistic account of a phenomenon generally gives us reason to view the relation between, for example, a verb's syntax and its semantics, as motivated rather than arbitrary. However, it is not in general possible to model the A-M distinction mathematically. The distinction that can be modeled mathematically is between R and not-R, that is, between random, or uncorrelated, pairs and pairs where there is some correlation, be it arbitrary, motivated or predictable.[1] The mechanism here is hypothesis-testing. A null hypothesis, H_0 is constructed to model the situation in which there is no correlation between

Corpus Linguistics and Linguistic Theory 1–2 (2005), 263–276 1613-7027/05/0001–0263

the two phenomena. As the mathematics of the random is well understood, we can compute the likelihood of the null hypothesis given the data. If the likelihood is low, we reject H_0.

The problem for empirical linguistics is that language is not random, so the null hypothesis is never true. Language is not random because we speak or write with purposes. We do not, indeed, without computational help are not capable of, producing words or sounds or sentences or documents randomly. We do not always have enough data to reject the null hypothesis, but that is a distinct issue: wherever there is enough data, it is rejected. Using language corpora, we are frequently in the fortunate position of having very large quantities of data at our disposal. Then, even where pairs of corpora are set up to be linguistically identical, the null hypothesis is resoundingly defeated. In section 4, we present an experiment demonstrating this counterintuitive effect.

There are a number of papers in the empirical linguistics literature where researchers seemed to be testing whether an association was linguistically salient, or used the confidence with which H_0 could be rejected as a measure of salience, whereas in fact they were merely testing whether they had enough data to reject H_0 with confidence. Some such cases are reviewed in section 5. Hypothesis testing has been widely used in the acquisition of subcategorization frames from corpora and this literature is considered in some detail. Alternatives to inappropriate hypothesis-testing are presented.

Before proceeding, may I clarify that this paper is in no way critical of using probability models, all of which are based on assumptions of randomness, in empirical linguistics in general. Probability models have been responsible for a large share of progress in the field in the last decade and a half. The randomness assumptions are always untrue, but that does not preclude them from frequently being useful. Making false assumptions is often an ingenious way to proceed; the problem arises where the literal falsity of the assumption is overlooked, and inappropriate inferences are drawn.

2. The arbitrary and the random

In common parlance, *random* and *arbitrary* are synonyms, with dictionaries giving near-identical definitions: LDOCE (1995) defines *random* as

> happening or chosen without any definite plan, or pattern

and arbitrary as

> **1** decided or arranged without any reason or plan, often unfairly … **2** happening or decided by chance rather than a plan

Superficially, randomness, as defined here, is what the technical sense of *random* captures and makes explicit. The technical sense is defined in terms of statistical independence. First, we formalize the framework:

> For a population of events, the first phenomenon holds where x is true of the event, the second holds where y is true of the event.

Now, the relation between the phenomena is random iff the probability of x, for that subset of events where y does hold, is identical to its probability for the subset where y does not hold, that is

$$P(x|y) = P(x|\neg y)$$

The relation is symmetric: $P(x|y) = P(x|\neg y)$ entails $P(y|x) = P(y|\neg x)$. Hereafter I use 'random' for the technical meaning and 'arbitrary' for the non-technical one.

Arbitrary events are very rarely random, and random events are very rarely arbitrary. It takes considerable ingenuity and sophisticated mathematics to produce a pseudo-random sequence algorithmically, and true randomness is not possible at all. Events happening "without any definite plan, aim or pattern" are, by definition, arbitrary, but are vanishingly unlikely to be random. Outside the sub-atomic realm, natural events are very rarely random.

Consider, for example, cat food purchases and shoe-polish purchases within the space of all UK supermarket-shopping events: does the fact that cat food was bought predict (positively or negatively) whether shoe polish was bought in the same shopping trip? There is no obvious reason why it should, and we can happily declare the relation arbitrary. But perhaps either cat food or shoe-polish are more (or less) often bought in hot (or cold) weather, or on Saturday nights, or Sunday mornings, or Monday lunchtimes, or by richer (or poorer) people, or by men (or women), or by people in (or out of) towns… There is an unlimited number of hypotheses connecting the two (positively or negatively); if just one of these has any validity, however weak, then the null hypothesis is false.

At this point, you may question why the null hypothesis is ever a useful construct.

For a wide range of tasks, although H_0 is false, there is only enough evidence to establish the fact if there is a strong relation between the two phenomena. Thus, given evidence from 1,000 shopping trips, it is unlikely we shall be able to reject H_0 concerning cat food and shoe-polish, whereas we shall be able to reject it concerning strawberry-buying and cream-buying. Given further evidence, perhaps from 1,000,000 shopping trips, we shall also be able to reject the null hypothesis regarding nappy[2]-buying and beer-sixpack-buying. (The correlation, the most newsworthy product of large-scale data mining by supermarkets, was widely reported in the British media.) But still not for catfood and shoe-polish. But, given 1,000,000,000 events, we shall in all likelihood also be able to reject it for catfood and shoe-polish.

Whether or not we can reject the null hypothesis (with eg. 95% confidence) is a function of sample size and level of correlation. Where sample size is held constant (and is not enormous), whether or not we can reject H_0 can be seen as a way of providing statistical support for distinguishing the arbitrary and the motivated. This is a role that hypothesis testing plays across the social sciences. However where the sample size varies by an order of magnitude, or where it is enormous, it is wrong to identify the accept-H_0/reject-H_0 distinction with the arbitrary/motivated one.

The uneasy relationship between hypothesis-testing, and quantity of data, is familiar to statisticians though frequently overlooked or misunderstood by users of statistics (Carver 1993, Stubbs 1995, Brandstätter 1999). One statistics textbook warns thus:

> 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, p 359).

The issue is particularly salient for empirical linguistics because, firstly, we have access to extremely large sample sizes, and secondly, the distribution of many language phenomena is Zipfian. *The* has 6,000,000 occurrences in the BNC whereas *cat food* (spelled as one word or two) has 66. For a vast number of third phenomena *X,* the null hypothesis that *the* and *X* are uncorrelated will be rejected, whereas the null hypothesis that *cat food* and *X* are uncorrelated will not. It would be wrong to draw inferences about what was arbitrary, what motivated.

3. Objections to Maximum Likelihood Estimates (MLEs)

Church and Hanks (1990) inaugurated the research area of lexical statistics with their presentation of Mutual Information (I), a measure of how closely associated two phenomena are. It can be applied to finding words which occur together to a noteworthy degree, or to finding words which are particularly associated with one corpus as against another, or for various other purposes.[3] They define the mutual information between two words x and y as

$$I(x;y) = \log 2\left(\frac{p\,(x\text{-}and\text{-}y)}{p\,(x)\cdot p(y)}\right)$$

and then estimate probabilities directly from frequencies, that is using the 'Maximum Likelihood Estimate' (MLE) of $f\,(x)/N$ for p (x), $f\,(y)/N$ for $p\,(y)$, $f\,(x\text{-}and\text{-}y)/N$ for $p\,(x\text{-}and\text{-}y)$, thereby giving

$$I(x;y) = \log 2\left(\frac{N\cdot f\,(x\text{-}and\text{-}y)}{f\,(x)\cdot f\,(y)}\right)$$

Dunning (1993) presents a critique of the use of Mutual Information in empirical linguistics. His objection has been confused with the critique of hypothesis-testing I make here so I mention his work in order to clarify that the two objections, while both valid, are different in nature and independent.

Dunning demonstrates how MLEs fare poorly when estimating the probabilities of rare events. The problem is essentially this: if a word (or bigram, or trigram, or character-sequence etc.) occurs just once or twice in a corpus of N words (bigrams, etc.), then the simplest way to estimate the probability is the MILE, which gives 1/N or 2/N. However this does not factor in the arbitrariness of the word occurring at all in the corpus: in a corpus ten times the size, there would be roughly ten times the number of singletons and doubletons in the corpus, most of which would not have occurred at all in the original corpus. Thus some of the probability mass contributing to the 1/N or 2/N MLEs should have been put aside for the words (bigrams etc.) which did not occur at all in this particular corpus. Viewed another way, the 1/N and 2/N should be discounted to allow for the fact that one or two occurrences are very low bases of evidence on which to assert probabilities.

There are various ways in which the discounting can be done, for example the Good-Turing method (Good 1953), usefully applied to empirical linguistics in Gale and Sampson (1995), Bod (1995). Dunning presents and advocates the use of the log-likelihood statistic, which, like the χ^2 statistic, is χ^2-distributed,[4] but more accurately estimates probabilities where counts are low. The log-likelihood statistic still only estimates probabilities: since Dunning's work, Pedersen (1996) has shown how Fisher's Exact Method can be applied to the problem, to identify the exact probability of a word (bigram etc.) rather than estimating it at all.

Thus Dunning's objection to Mutual Information is that it fails to accurately represent probabilities when counts are low (where 'low' is generally taken as less than five). If the probabilities can be accurately represented, Dunning's anxieties will be set at ease.

The critique in this paper does not concern whether probabilities are accurately calculated. Rather, the objection is that the probability model, with its assumptions of randomness, is inappropriate, particularly where counts are high (eg, thousands or more).

Where the task is to determine whether there is an interesting association between two rare events, Dunning's concern must be heeded. Where it is to determine whether there is an interesting association between high-frequency events, the concerns of this paper must be.

4. Experiment

Given enough data, H_0 is almost always rejected however arbitrary the data, as the author discovered when grappling with the following data.

Two corpora were set up to be indisputably of the same language type, with only arbitrary differences between them: 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 corpora.

The null hypotheses were (1) that the two subcorpora, viewed as collections of words rather than documents, were random samples drawn from the same population; and consequently, (2) that the deviation in frequency of occurrence for each individual word between the two subcorpora was explicable as random fluctuation. The H_O were tested using the χ^2-test: is χ^2

$$\Sigma(|O - E| - 0.5)^2 / E$$

greater than the critical value? The sum is over the four cells of the contingency table

	Corpus 1	Corpus 2
word w	a	b
not word w	c	d

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,

$$(|O - E| - 0.5)^2 / E$$

is then 0.5.[5] 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 χ^2 calculation was determined. As Table 4 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 $0.5 \times 7.88 = 3.94$, the relevant critical value of the chi-square statistic. For very many words, including most common words, the null hypothesis is resoundingly defeated (as is the null hypthesis regarding the two subcorpora as wholes).

There is no *a priori* reason to expect words to behave as if they had been selected at random, and indeed they do not. It is in the nature of language that any two collections of texts, covering a wide range of

Table 1. *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:/natcorp.ox.ac.uk/bnc).*

Class (Words in freq. order)	First item in class		Mean error term for items in class
	word	POS	
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*	AVO	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

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.

The 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, and the hypothesis test confirms the fact.

5. Re-analysis of previous work

5.1. Brown and LOB

Hofland and Johansson (1982) wanted to find words which were significantly different in their frequencies between British and American English, as represented in the Brown corpus for American English and LOB corpus for British. For each word, they tested the null hypothesis that the difference in frequency between the two corpora could be explained as random variation, with the samples being random samples from the same source, and in their frequency lists, they mark words where the null hypothesis was defeated (at a 95, 99 or 99.9% confidence level). Looking at these lists suggests that virtually all common words are markedly different in their levels of use between the US and the UK: they are all marked as such. By contrast, most of the rarer marked words are words we know to be American or British, or to refer to items that are more common or more salient in the US or the UK.

As the argument of the previous section explains, most of the marked high-frequency words are marked simply as a consequence of the essentially non-random nature of language. It would not be surprising for a high-frequency word marked as British English in these lists to be marked as American English in a repeat of the experiment using new data.

Similar strategies are used by, and a similar critique is applicable to, Leech and Fallon (1992) (again, for comparing LOB and Brown), Rayson, Leech, and Hodges (1997) for comparing the conversation of different social groups, and Rayson and Garside (2000) for contrasting the language of a specialist genre with ‘general language’, as represented by the British National Corpus.

5.2 Subcategorization frame (SCF) learning

Hypothesis-testing has been used in a number of papers concerning the automatic acquisition of subcategorization frames (SCFs) for verbs from corpora. The problem is this. Dictionaries, even where they do present explicit and accurate SCFs for verbs, are not complete: they do not present all the frames for each verb. This gives rise to many parsing errors. Researchers including Brent (1993), Briscoe and Carroll (1997) and Korhonen (2000) have developed methods for SCF acquisition. However, their methods are inevitably noisy, suffering, for example, from just those parser errors that the whole process is designed to address, and they do not wish to accept any SCF for which there is any evidence as a true SCF for the verb. They wish to filter out those SCFs where the evidence is not strong enough. Brent and Briscoe and Carroll used hypothesis testing to this end. However, problems are noted:

> Further evaluation of the results ...reveals that the filtering phase is the weak link in the system … The performance of the filter for classes with less than 10 exemplars is around chance, and a simple heuristic of accepting all classes with more then 10 exemplars would have produced broadly similar results for these verbs (Briscoe and Carroll 1997: 360−36).

Korhonen, Correll, and McCarthy (2000) explore the issue in detail. Using Briscoe and Carroll’s SCF acquisition system, they explore the impact of four different strategies for filtering out noise:

Baseline No filter

BHT binomial hypothesis test: reject the SCF if H_o *is* not defeated[6]

LLR hypothesis test using log-likelihood ratio: reject the SCF if H_o *is* not defeated

MLE threshold based on the relative frequency (which is also the maximum likelihood estimate (MLE) of the probability) of the verb occurring in the SCF given the verb, with the threshold determined empirically

They observe

> MLE thresholding produced better results than the two statistical tests used. Precision improved considerably, showing that the classes occurring in the data with the highest frequency are often correct … MLE is not adept at finding low frequency SCFs … (Korhonen, Correll, and McCarthy 2000: 202)

This concurs with the theoretical argument above. Hypothesis tests are inappropriate for the task, because the relations between verb and SCF will never be random and the hypothesis test will merely reject the null hypothesis wherever there is enough data, in a manner not closely correlated with whether the SCF-verb link is motivated. Where there is enough data, then the relationship between verb and SCF is easy to see so even a simple threshold method will identify the verb's SCFs. Where data is very sparse, no method works well.

Korhonen (2000) extends this line of work, exploring thresholding methods where a more accurate estimate of the probability is obtained by using data from semantically similar but higher frequency verbs. She achieves modest improvements over the baseline which uses Korhonen, Correll and McCarthy's MLE, particularly when combining the frequencies of the target verb and its semantic neighbour using a linear method based on the quantity of evidence available for each.

The problem is not one of distinguishing random and non-random relationships, but of sparseness of data. Where the data is not sparse, the difference between arbitrary and motivated connections is evident in greatly differing relative frequencies. This makes the moral of the story plain. Data is abundant. A modest-frequency verb like *devastate* occurs (Google tells us) in well over a million web pages. With just 1 % of them, devastate becomes one of the verbs for which we have plenty of data, and crude thresholding methods will distinguish associated SCFs from noise. It is possible that parsing errors are systematic and thus that the same errors occur very often in very large corpora although our experience from looking at large corpora in the Word Sketch Engine (Kilgarriff et al 2004) suggests not. Harvesting the web (or other huge corpora) is the way to build an accurate SCF lexicon.[7]

6. Conclusion

Language is non-random and hence, when we look at linguistic phenomena in corpora, the null hypothesis will never be true. Moreover, where there is enough data, we shall (almost) always be able to establish that it is not true. In corpus studies, we frequently do have enough data, so the fact that a relation between two phenomena is demonstrably non-random, does not support the inference that it is not arbitrary. Hypothesis testing is rarely useful for distinguishing associated from non-associated pairs of phenomena in large corpora. Where used, it has often led to unhelpful or misleading results.

Hypothesis testing has been used to reach conclusions, where the difficulty in reaching a conclusion is caused by sparsity of data. But language data, in this age of information glut, is available in vast quantities. A better strategy will generally be to use more data Then the difference between the motivated and the arbitrary will be evident without the use of compromised hypothesis testing. As Lord Rutherford put it: “If your experiment needs statistics, you ought to have done a better experiment.”

Received July 2004
Revisions received May 2005
Final acceptance May 2005

Lexical Computing Ltd.

Notes

* This work was supported by the UK EPSRC, under grants GR/K18931 (SEAL) and GR/M54971 (WASPS).

1. In this paper we do not consider the distinction between the predictable and the ‘merely’ motivated.
2. Diapers, in American English.
3. There is some confusion over names. In information theory, Mutual Information is usually defined over a whole population of words, rather than being specified for a particular word-pair, as here, and the definition incorporates information from all cells of the contingency table. Church and Hanks only use a subset of that information. Church-and-Hanks Mutual Information has been called Pointwise Mutual Information. See Manning and Schütze (1999: 66 ff.) for a fuller discussion. Here we use Church and Hanks’s definition and name.
4. This sentence will be confusing to non-mathematicians. The χ^2 statistic is a statistic, that is, it can be calculated from a data sample using actual numbers. The χ^2 distribution is a theoretical construct. If a sufficiently large number of chi-square statistics are calculated, all from true random samples of the same population, then this population of χ^2 statistics will, provably, fit a χ^2 distribution. This is also true for other statistics: that is, if a sufficiently large number of log-likelihood statistics are calculated, all from true random samples of the same population, then this population of log-likelihood statistics will, provably, fit a χ^2 distribution. Some texts call the statistic x^2 rather than χ^2 to distinguish it more clearly from the distribution, but this practice is in the minority and is not adopted here.
5. See appendix.
6. The model used was a sophisticated one incorporating evidence about type frequencies of verbs from the ANLT lexicon: see Briscoe and Carroll (1997) or Korhonen, Correll, and McCarthy (2000) for details.
7. See Kilgarriff and Grefenstette (2003) and papers therein. The web is a vast resource for many languages. See also Banko and Brill (2001) for the benefits of large data over sophisticated mathematics.

Appendix

The average value of the error term is 0.5. We explain this as follows.

If we do in fact have a random distribution, then by the definition of the χ^2 distribution, the sum of the cells in the contingency table is 1:

$$a + b + c + d = 1$$

Each of these error terms is calculated as

$$(O - E - 0.5)^2/E$$

In our situation, there are very large datasets and the phenomenon of interest only accounts for a very small proportion of cases. The frequency of *not word w* is very high. Thus the expected values, *E,* for *not word w* to be used when calculating *c* and *d* for the contingency table are very high. As we divide by very large *E, c* and *d* are vanishingly small, so

$$a + b + c + d = 1$$

reduces to

$$a + b = 1$$

Since we have set the situation up symmetrically, *a* and *b* are the same size, so each will be, on average, 0.5.

References

Banko, Michele and Eric Brill
2001 Scaling to very very large corpora for natural language disambiguation. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter of the Association for Computational Linguistics.*

Bod, Rens
1995 Enriching linguistics with statistics: performance models of natural language. Ph.D. dissertation, University of Amsterdam.

Brandstätter, E.
1999 Confidence intervals as an alternative to significance testing. *Methods of Psychological Research Outline* 4(2), 33–46.

Brent, Michael R.
1993 From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics* 19(2), 243–262.

Briscoe, Ted and John Carroll
1997 Automatic extraction of subcategorization from corpora. *Proceedings of the Fifth Conference on Applied Natural Language Processing*, 356–363.

Carver, Ronald P.
1993 The case against statistical significance testing, revisited. *Journal of Experimental Education* 61, 287−292.

Church, Kenneth and Patrick Hanks
1990 Word association norms, mutual information and lexicography. *Computational Linguistics* 16(1), 22−29.

Dunning, Ted
1993 Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1), 61−74.

Gale, William and Geoffrey Sampson
1995 Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics* 2(3),

Good, I. J.
1953 The population frequencies of species and the estimation of population parameters. *Biometrika* 40, 237−264.

Grefenstette, Gregory and Julien Nioche
2000 Estimation of English and non-English language use on the www. In *Proceedings of RIAO (Recherche d'Informations Assistée par Ordinateur)*, 237−246.

Hofland, Knud and Stig Johanson (Eds.)
1982 *Word Frequencies in British and American English*. Bergen: The Norwegian Computing Centre for the Humanities.

Kilgarriff, Adam and Gregory Grefenstette
2003 Introduction to a special issue on web as corpus. *Computational Linguistics* 29(3), 333−348.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell
2004 *The Sketch Engine*. Proceedings of EURALEX, European Association for Lexicography, 105−116.

Korhonen, Anna
2000 Using semantically motivated estimates to help subcategorization acquisition. *Proceedings of the Joint Conference on Empirical Methods in NLP and Very Large Corpora,* 216−223.

Korhonen, Anna, Genevieve Gorrell, and Diana McCarthy
2000 Statistical filtering and subcategorization frame acquisition. *Proceedings of the Joint Conference on Empirical Methods in NLP and Very Large Corpora,* 199−206.

LDOCE
1995 *Longman Dictionary of Contemporary English, 3rd Edition*. Ed. Della Summers. Harlow: Longman.

Leech, Geoffrey and Roger Fallon
1992 Computer corpora — what do they tell us about culture? *ICAME Journal* 16, 29−50.

Manning, Christopher and Hinrich Schütze
1999 *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Owen, Frank and Ronald Jones
1977 *Statistics*. Polytech Publishers.

Pedersen, Ted
1996 Fishing for exactness. *Proceedings of the Conference of the South-Central SAS Users Group*, 188−200.

Rayson, Paul and Roger Garside
2000 Comparing corpora using frequency profiling. *Proceedings of the Workshop on Comparing Corpora, 38th ACL*, 1–6.

Rayson, Paul, Geoffrey Leech, and Mary 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.

Stubbs, Michael
1995 Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language* 2(1), 23–55.