

Comparing word frequencies across corpora: Why chi-square doesn't work, and an improved LOB-Brown comparison

Adam Kilgarriff

Research Fellow, Information Technology Research Institute, University of Brighton, Lewes Road Brighton BN2 4AT, UK

AFFILIATION: ITRI, University of Brighton

E-MAIL: Adam.Kilgarriff@itri.bton.ac.uk
 FAX NUMBER: (44) 1273 642908
 PHONE NUMBER: (44) 1273 642919

1. Introduction

We are often interested in discovering which words are markedly different in their distribution between two texts or two corpora. In this paper I show that one statistic which has sometimes been used for this purpose, chi-square, is inappropriate. I present an alternative, the Mann-Whitney ranks test. I apply the test to finding the words which are most different between the LOB and Brown corpora and show that it produces output that is well suited to the interests of lexicographers and humanities scholars.

2. A simple framework

For two texts, which words best characterise their differences? For word *w* in texts *X* and *Y*, this might be represented in a contingency table as follows:

| | | | |
|-------|-----|-----|-----------|
| | X | Y | |
| w | a | b | a+b |
| not w | c | d | c+d |
| | a+c | b+d | a+b+c+d=N |

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

3. The chi-square test

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

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

(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 chi-square-distributed with (*m*-1)×(*n*-1) degrees of freedom (provided all expected values are over a threshold of 5.) For our 2×2 contingency table the statistic has one degree of freedom and we apply Yates' correction, subtracting 0.5 from *O-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) to identify where words are more common in British than American English or vice versa. *X* was the LOB corpus, *Y* was the Brown, and, in the table where they make the comparison, the chi-square value for each word is given, with the values marked where they exceeded critical values (at any of three levels of significance) so one might infer that the LOB-Brown difference was non-random.

Looking at the LOB-Brown comparison, we find that this is true for very many words, and for almost all very common words. Most of the time, the null hypothesis is defeated. Does this show that all those words have systematically different patterns of usage in British and American English?

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, for each of which a frequency list for the first 20,000 running words was generated. Half the lists were then randomly assigned to each of two subcorpora. Frequency lists for each subcorpus were generated. For each word occurring in either subcorpus, the

$$\frac{(O-E-0.5)^2}{E}$$

term which would have contributed to a chi-square calculation was determined. If the two corpora were random samples of words – not texts – drawn from the same population, the error term would not vary systematically with the frequency of the word, and the average error term would be 0.5. In fact, as the table shows, average values for the error term are far greater than that, and tend to increase as word frequency increases.

| Class (Words in freq. order) | First item in class Word | POS | Mean error term for items in class |
|---------------------------------------|--------------------------------|--------|--|
| First 10 items | the | DET | 18.76 |
| Next 10 items | for | PREP | 17.45 |
| Next 20 items | not | NOT | 14.39 |
| Next 40 items | have | V-BASE | 10.71 |
| Next 80 items | also | ADV | 7.03 |
| Next 160 items | know | V-INF | 6.40 |
| Next 320 items | six | CARD | 5.30 |
| Next 640 items | finally | ADV | 6.71 |
| Next 1280 items | plants | N-PL | 6.05 |
| Next 2560 items | pocket | N-SING | 5.82 |
| Next 5120 items | represent | V-BASE | 4.53 |
| Next 10240 items | peking | PROPER | 3.07 |
| Next 20480 items | fondly | ADV | 1.87 |
| Next 40960 items | chandelier | N-SING | 1.15 |

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. As in the LOB-Brown comparison, for very many words, including most common words, the null hypothesis is defeated.

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.

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 chi-square statistic – 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 chi-square values are associated with the sheer quantity of evidence and are not necessarily associated with a pre-theoretical notion of distinctiveness.

4. Burstiness

As Church and Gale (1995) say, words come in bursts; unlike lightning, 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. A single document containing w is relatively likely to contain a ‘burst’ of w 's, so whichever corpus contains that document, will contain more w 's than is compatible with the null hypothesis. We require a test which does not give undue weight to single documents with a high count for w .

A test meeting this criterion is the Mann-Whitney (also known as Wilcoxon) ranks test¹. To perform this test, we use frequency of occurrence to rank the data, and then use ranks rather than frequency to compute the statistic. The test proceeds as follows. The corpora to be compared are each divided into a number of equal-sized parts (for purposes of illustration, we use five). Suppose the frequencies for X are

12 24 15 18 88

and for Y are

3 3 13 27 33

As the subcorpora that these frequencies are based

on are all of the same size, the figures are directly comparable. They are now placed in rank order, a record being kept of the corpus they come from:

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

The ranks associated with the corpus with the smaller number of samples (or either, where, as here, there are equal numbers for each) are summed: for Y, 1+2+4+8+9=24. This sum is compared with the value that would be expected, on the basis of the null hypothesis. These values are tabulated (at various significance levels) in statistics textbooks. If the null hypothesis were true, 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.

A complication arises where two samples have the same number of hits so they cannot be straightforwardly ranked. Recommended practice here is to, first, give all X's higher ranks, and then repeat giving all Y's higher ranks. If the two methods give different conclusions, the test is not applicable.

5. LOB-Brown comparison

The LOB and Brown 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 the null hypothesis has little appeal: there is no a priori 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 had been avoided.

The sA and sB words were classified as follows:

| Code | Mnemonic | Example | sA | sB |
|--------|------------|---|-----|-----|
| s | Spelling | color/colour; realise/realize | 30 | 23 |
| e | Equivalent | toward/towards; flat/apartment | 15 | 17 |
| n | Name | los, san, united; london, africa, alan | 45 | 24 |
| c | Cultural | negro, baseball, jazz; royal, chap, tea | 38 | 26 |
| ? | Unclear | e, m, w ... (to be investigated) | 10 | 10 |
| o | Other | | 154 | 116 |
| Totals | | | 288 | 216 |

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: if we had not identified the items in these classes as sA and sB, then our method would not have been working. It is the items in the 'others' category which are interesting. The three 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 lexicographers might want to examine whether there is some such difference.

Notes

¹ A survey of other statistics which have been used for this purpose is available in Kilgarriff (1996).

Acknowledgements

This work is supported by the EPSRC, Grant K18931, SEAL. The idea of using the Mann-Whitney test emerged from a discussion with Ted Dunning and Mark Lauer.

References

- Kenneth Church and William Gale. Poisson Mixtures. *Journal of Natural Language Engineering*, 1(2):163–190.
- Knut Hofland and Stig Johansson. 1982. *Word Frequencies in British and American English*. The Norwegian Computing Centre for the Humanities, Bergen, Norway.
- Adam Kilgarriff. 1996. Which words are particularly characteristic of a text? A survey of statistical approaches. In: *Language Engineering for Document Analysis and Recognition*. Proceedings, AISB Workshop, Falmer, Sussex.