NGL 33-022-090
CR-109148
p. 12

CORRECTING WORD FREQUENCIES TO TAKE ACCOUNT OF RANGE

G. Harry McLaughlin

August 1969

Occasional Paper No. 1

(NASA-CR-109148) CORRECTING WORD FREQUENCIES TO TAKE ACCOUNT OF RANGE (Syracuse Univ.) 12 p

N70-70661

Unclas
00/62 0278548

OCT 1 1969

Occasional Paper No. 1
Aug. 1969

" Correcting word frequencies to take account of range* "

G. Harry McLaughlin
Communications Research Center
Syracuse University

I am circulating this paper for constructive criticism before offering it for publication. If you have any observations please write them on this draft and return it to me at the Newhouse Communications Center by October 10--and, later, I will send you a copy of the final version.

* This study was supported in part by the Syracuse/NASA Program.

I would like to thank Messrs. Peter Fellows and Martin Steinbuch ("Research Methods" course students) for help in data processing.

Correcting word frequencies to take account of range

Words of a feather flock together. A speaker or writer will repeat certain key words for as long as he continues to discuss the same topic. Consequently, there are inherent biases in word frequency counts based on clustered samples, that is, in counts of all the immediately successive word tokens in each of the texts comprising a corpus. Such a count will tend to overestimate the frequency of word types associated with the topics discussed and to underestimate the frequencies of words germane to the vast number of topics necessarily left out of any reasonably-sized corpus. It is desirable to compensate for these biases even though it will be shown that their effect on the overall pattern of the word frequency distribution is slight. It will also be shown that an arbitrary weighting system can seriously distort the distribution, but a rationale will be offered for a convenient new correction method.

Before continuing the argument, a note of caution must be introduced. It is usually assumed that the frequencies observed in any corpus approximate the frequencies that characterize the language of which the corpus itself is merely a sample. This assumption is justified, however, only if the corpus is in fact a representative sample. A corpus will be representative if its texts are randomly selected from some well-defined universe, such as the works of a particular author, the recorded utterances of a particular speech community, or the materials read by a clearly specified group of readers.

An arbitrary assortment of texts on different topics and in different styles, however large, is most unlikely to provide an unbiased sample of the language used by any clearly identifiable group of people; it would not, then, be legitimate to make generalizations from a word count based on such a corpus. Therefore in the following discussion it must be understood that the

term corpus refers to a collection of texts representative of the type of language used by some clearly specified type of person. Only if the term is used in this sense is it meaningful to say that counts of immediately successive words occurring in a corpus produce biased estimates of word frequencies in the language.

That shortcoming could be overcome in principle by using instead a truly random sampling procedure. This would involve counting only every nth word token in each sample. However, it would be necessary to set n sufficiently large, so as to ensure that the tokens counted in a single sample are statistically independent, making such a sample indistinguishable in character from a sample consisting of tokens each sampled at random from a different text. As a guess, n would have to be at least 1,000. Thus, to obtain a count of 1,000,000 words, it would be necessary to assemble a corpus of 1,000,000,000 words. The volume of work involved in counting every 1,000th word in such a corpus would render it impractical with present-day techniques.

It is, however, possible to correct frequency counts for the effects of clustered sampling by taking account of differences in range. Here range refers to the proportion of samples which include a given word. Clearly, if a word's range is so wide that it occurs in almost every sample, its observed frequency is likely to be a reasonably good estimator of its frequency in the language of that corpus. On the other hand, another word having the same frequency as the first, but with a range restricted to only one or two samples may be presumed to be highly specific to the topic of those samples: its observed frequency in the corpus will therefore almost certainly be an overestimate of its frequency in the language.

It follows that better estimates of frequencies in the language could be obtained by weighting the observed frequencies of words in proportion to

their range.

The weighting function must be chosen with considerable care. An arbitrary method of weighting can so overcorrect the frequency distribution as to distort it seriously. This can be seen from a detailed consideration of the word count published by Ernest Horn (1926). This publication is of importance because the word frequencies it reports provide the data base for the most ambitious attempt to analyse occurrences of English speech sounds statistically, which was made by A. Hood Roberts (1965).

To begin with, it should be noted that Horn's count involved a series of estimates. In 1922, he made a compilation of some ten counts carried out by various investigators, including himself. Horn claimed that this compilation was based upon the examination of 864,334 running words (tokens). However, two of the investigators reported only estimates of the total number of tokens examined, and these two account for approximately 265,000 tokens out of the grand total of 864,334. As Horn remarked:

> "The scientific accuracy of this compilation is marred by the lack of completeness in the data reported in some investigations, as well as by lack of uniformity among the various investigations in the method of tabulating words." (p.17)

Neither _a_ nor _I_ was tabulated in the bulk of the investigations: 36 additional types were omitted from the published lists of one or more investigations. For any word omitted from an investigation

> "its probable frequency in that investigation was computed on the basis of the data for the word in the investigation which most nearly duplicates that in which it was omitted." (p.18)

Nonetheless, Horn used this compilation in his 1926 publication to estimate the frequencies of 372 word types, namely all words less than four

letters long and 41 other high-frequency words. Horn explained his method of estimation thus:

> "The sum of the frequencies of these 372 words in the compilation of 1922 was found to be 526,149. This number subtracted from the grand total 864,334 gives 338,185, which is the sum of the frequencies for all words when these 372 words are excluded. It was assumed that, if these special words had been tabulated in the new investigation, they would have occurred with approximately the same relative frequency. Thus, in the new investigation the ratio of the total number of running words to the sum of the frequencies of all words excluding these special words was assumed to be the same as in the compilation of 1922. According to this assumption, the total number of running words examined in the new investigation was computed to be 4,272,482. When to this number are added the 864,334 running words of the compilation of 1922, the estimated total number of running words examined in making the present compilation is found to be 5,136,816 . . ." (p.41)
>
> "The total number of running words scored in the 1922 investigation (864,334) was then divided into 5,136,816 . . . The quotient is 5.94. Since the final compilation is 5.94 times as extensive as the earlier one, the frequency of each word not tabulated in the new investigations was found in the 1922 compilation and multiplied by 5.94. The result is the estimated unweighted frequency with which each of these 372 words would most likely have been found had it been scored in the new investigations." (p.42)

This extensive quotation has been given to establish beyond doubt that, including about 526,149 x 5.94 = 3,125,325 "omitted" word tokens, Horn's corpus consisted of little more than

5,000,000 tokens. Yet the frequencies of the 10,000 types which he reported add up to 15,465,010 tokens!

The explanation for this glaring discrepancy is that Horn did not, as Roberts (p.28) stated, give "the actual frequencies of occurrence of all words in the word count": the figures he published are not the observed frequencies: they are "corrected" versions thereof.

Horn went into many pages of detail concerning his method of weighting the frequencies of each word type actually counted. Basically, his principle was to multiply the actual frequency of a type occurring in each group of samples by the approximate square root of the number of samples in that group in which it occurred. He then summed the weighted frequencies derived from each group (p.26). For example, the actual total number of occurrences of the word "around" in Horn's corpus was no more than 1,870, according to Tables 5 through 9 of his introductory Chapter 2. However, the total number of the occurrences reported in the frequency list is the weighted total, namely, 5,442.

If the groups of samples had been equated for size before Horn applied the square root transformation, its effect would at least have been similar for all groups. Unfortunately, they each varied in the number of their constituent samples, and each of the samples contained a different number of tokens. Thus a variety of unsystematic biases were introduced into the weighting system. Its net effect, however, was simply to treble the frequencies of the most common words, to double the frequencies of the less common, but to leave the frequencies of the rare words unchanged (Horn, p.42).

In any case, the square root transformation considerably depresses the frequency of less common words relative to the most common, simply because the former occur in fewer texts. As it is only to be expected that the less common words would have a restricted range, there is no justification for

"correcting" the relative frequencies of all words, other than the thousand most common, by slashing them to one-third of their observed values. Yet this is what Horn's transformation virtually does. Using unweighted frequencies, Kucera and Francis (1967) show in their Table I that word-types other than the 1,007 first in rank order account for 31% of the million tokens in the Brown University corpus. By contrast, Roberts, relying on Horn's weighted frequencies , is led in Appendix V to the conclusion that types other than the first 1,007 account for no more than 10 per cent of the tokens in the corpus. Naturally, all the phoneme frequency distributions given by Roberts are correspondingly distorted.

What seems to be required is a method of weighting which would reduce the observed frequencies only for words used repeatedly in a disproportionately small number of samples. The rationale for one suitable correction system will now be derived.

Suppose that we were to obtain an homogenous corpus of enormous size, that we sampled only one word from each text and then arbitrarily collected the words into a number of equal sized samples. We would then have a perfectly random sampling in which any occurrence of a word would have an equal probability of falling into any one of the samples. This condition is all that is required to produce a Poisson distribution for each word-type. The Poisson distribution is characterized by the fact that the probability of a particular type being in any sample uniquely determines the proportion of samples which will contain no occurrences of that type, exactly one occurrence, exactly two occurrences, exactly three occurrences, etc.

Thus the Poisson distribution constitutes a standard by which it is possible to judge the effects of clustered sampling on a corpus. A word's observed range will probably be smaller than the number of samples through

which it would be expected to be spread if the samples were perfectly random rather than clustered. It is therefore reasonable to correct (which will generally mean to reduce) an observed word frequency by a factor equal to the ratio between the range of the word and the range that would be expected in representative, truly random samples.

The procedure for making this correction is as follows. Divide the frequency of each word type by the total number of (equal-sized) samples in the corpus. This gives the average number of times that the type is represented in each ~~text~~ sample. Refer this average to a table of the Poisson distribution to find the proportion of samples in which the occurrence of that type is expected to be 0. Multiply the expected proportion by the number of samples to find the expected number of samples in which the type <u>will not</u> occur. Subtract this number from the total number of samples: obviously the result is the expected number of samples in which the type <u>will</u> occur, in other words, its expected range. Divide the observed range by the expected range to obtain the required correction factor. Finally correct the frequency by multiplying it by the correction factor, rounding off the product to the nearest whole number.

When this procedure was applied to the Brown University data, the results were somewhat surprising. Nearly every word appears in fewer samples than would be expected under truly random sampling. The correction was applied to every one of the 1,000 most frequent word types: only 14 of them had a range as great as that predicted by the Poisson model. It may therefore be surmised that an individual's literary style limits his selection of function words (which is what most of the very common types happen to be), just as his topic will limit his selection of content words.

Not only was nearly every word affected, but the impact of the correction procedure was roughly uniform throughout the entire corpus. As can be seen from the accompanying table, most corrected frequencies are about 70 per cent of the observed frequencies: much of the variance apparent in the table probably arises from the fact that the correction was applied to only a proportion of the types beyond the first 1,000.

The uniform effect of the correction means that the relative frequencies are little changed. The corrected frequencies add up to only 82.25% of the number of tokens in the original corpus. The reduced corpus is designed to be equivalent to a truly random sampling. It is therefore legitimate to compare observed frequencies, expressed as cumulative percentages of the original corpus, with corrected frequencies, expressed as cumulative percentages of the reduced corpus. The table shows that the two sets of figures are very similar indeed.

Nonetheless, many content words will be over-represented in any clustered sample, so it is still desirable to make the correction, especially considering that it is trivial to program a computer to do the calculations. In any case, calculations will be needed for only the first few thousand frequencies, because the observed range can be treated as the expected frequency of all rare words. In the Brown University count, which has 500 samples, it was found that the corrected frequency was equal to the observed range of all words occurring 23 times or fewer per million--that is, all types beyond the first 4,270. This phenomenon occurs because a rare word will not be expected to occur more than once per sample, so that the observed frequency and the expected range will be nearly equal, and will in effect, cancel out in the formula:

corrected frequency = observed frequency x observed range ÷ expected range

What is a rare word for this purpose will depend upon the number of samples and their size.

Results of applying the Poisson correction for range to groups of word types ranked in order of decreasing observed frequency

Lowest rank in group of types (1)	Lowest observed frequency in group of types (2)	Mean corrected frequency as percentage of mean observed frequency (3)	Cumulative sum of tokens accounted for as percentage of total in corpus		Percentage of number of types used to calculate results in colums 3 and 5 (6)
			Raw data (4)	Corrected (5)	
1,000	106	87.2	68.9	72.9	100
2,000	56	67.9	76.2	79.0	20
3,000	37	74.6	80.6	82.8	10
4,000	26	62.6	83.6	85.3	4
5,000	19	65.0	85.9	87.0	4
6,000	15	63.6	87.5	88.3	4
7,000	12	65.4	88.9	89.4	4
7,421	12	66.7	89.4	89.9	4
7,919	11	72.7	90.0	90.3	4
8,478	10	70.0	90.5	90.8	4
9,173	9	66.7	91.1	91.3	4
9,998	8	56.2	91.8	91.7	4
11,119	7	71.4	92.6	92.4	1
12,398	6	65.1	93.3	93.0	1
14,218	5	74.7	94.2	93.8	1
16,683	4	77.1	95.2	94.8	1
20,630	3	79.4	96.3	95.9	1
27,863	2	81.6	97.8	97.3	1
50,406	1	100.0	100.0	100.0	-

Based on data given in the Rank List and Table B 1 by Kucera and Francis (1967).

References

Horn, Ernest. A basic writing vocabulary: University of Iowa Monographs in Education, 1926, 7, No. 4.

Kucera, Henry and Francis, W. Nelson. Analysis of present-day American English. Providence, Rhode Island: Brown University Press, 1967.

Roberts, A. Hood. A statistical linguistic analysis of American English. The Hague: Mouton, 1965.