Bulletin of the Psychonomic Society
1988, 26 (5), 410-412

On using norms for low-frequency words

EUGENE A. LOVELACE
Alfred University, Alfred, New York

When norms are used to select words of various frequency of occurrence in the language, great care must be exercised in the selection of low-frequency words. The stability of estimates of frequency for rare words hinges heavily on the size of the corpus on which the word count is based, and on whether the frequency index takes into account the distribution across samples making up the corpus. Although some frequency norms directly provide an index that takes dispersion into account (*The American Heritage Word Frequency Book*), others do not (e.g., Kučera & Francis, 1967). Researchers predominantly use the Kučera and Francis norms, and they routinely take the total frequency of occurrence in the corpus as their frequency index, with no correction for dispersion.

Norms detailing the frequency with which words occur in the printed language often have been employed by empirical researchers either to select materials that will have specified frequency characteristics or, as a control measure, to match two sets of materials so that performance differences are not attributable to differences in language frequency. There are currently three sets of norms that account for nearly all recent citations of language frequency in American journals. In historical order of publication, they are *The Teacher's Word Book of 30,000 Words* (Thorndike & Lorge, 1944), *Computational Analysis of Present-Day American English* (Kučera & Francis, 1967), and *The American Heritage Word Frequency Book* (Carroll, Davies, & Richman, 1971). A more recent set, *Frequency Analysis of English Usage: Lexicon and Grammar*, by Francis and Kučera (1982), is based on the same word counts reported in 1967 by Kučera and Francis.

In the case of selection from the norms, researchers typically select words from these norms so as to create sets of items that vary with respect to language frequency. When one undertakes to use the norms for words already selected, however, potential peculiarities of such norms with respect to low-frequency words become apparent. For example, I sought the language frequencies in the Kučera & Francis (1967) norms for a large pool of words I had chosen on other criteria. To my surprise, 104 of those words were not to be found in that word count. Although some of these words did seem quite rare to me (e.g., *littoral*, *exactitude*, *maize*, *twang*), others seemed decidely more common (e.g., *cucumber*, *fraternal*, *lettuce*, *nag*, *quilt*, *spool*, *toaster*), yet all had word counts of 0 in those norms. On the other hand, scanning a portion of the norms revealed that the following words all had counts of 10 occurrences: *autocoder*, *conjugate*, *eromonga*, *irradiation*, *mosque*, *phonologic*, *polynomials*, *text-form*, and *vagina*. (Because of the very skewed distribution of frequencies in such counts, this frequency count of 10 places them among the 20% of words with the *highest* frequency counts in those norms; the majority of words have counts so low as to be changed dramatically in frequency rankings by a change of only 2 or 3 in their total frequency of occurrence.)

There are a number of studies indicating that subjective frequency judgments correlate very highly with objective word counts (e.g., Carroll, 1971; Shapiro, 1969). My subjective assessments of word frequency were telling me that something was clearly amiss; those words with counts of 10 (listed above) seemed much less common than many of the 104 words that I had discovered did not exist in Kučera and Francis's (1967) count. To assess whether others shared my perceptions, I selected 50 of the words I could not find in those norms, then paired each with a word of about the same length that began with the same letter and that had a Kučera-Francis count of 3 to 7 (all but 6 having counts of 5 or greater in the norms). The pairs were randomly ordered in a single column on a page, with the word existing in the norms on the left in half the pairs. Twenty-five faculty, staff, and students were asked to simply work down the sheet and "for each pair, *circle* the word that you think occurs *more* frequently."

For each of these 50 pairs, the number of participants (out of 25) choosing as more common the word that does not appear in those norms ranged from 16 to 25 (M = 22.3, *Mdn* = 23). In every case, more people chose the word with 0 frequency than the one with a frequency around 6 in those norms. For 30 of the 50 word pairs, 23 or more of the 25 judges chose the word missing from the norms as the more common.

These data are not to be taken as a blanket condemnation of the particular set of norms. But they do reflect two problems, one with the norms themselves and the other with the way in which these norms typically have been employed. The problem with the norms themselves results from the frequencies being based on a corpus that is too small to permit accurate representations of the relative

Address correspondence to Eugene A. Lovelace, Psychology-Science Center, Alfred University, Alfred, NY 14802.

occurrence of rare words. The corpus for Kučera and Francis's (1967) count was 1,014,232 printed words. This was divided into 500 samples of about 2,000 words each. These samples were selected so as to represent some 15 categories or genres of writing. Although this may seem like a large number of words, it must be noted that this computer-coded count has a much smaller corpus than Thorndike and Lorge's (1944) older count that was done manually on 18 million words, or the *American Heritage* word count that is based on a corpus of over 5 million words. Kučera and Francis's corpus of a little over a million words is adequate for a number of uses one may have for these word counts, but it is clear that some words missing from those norms, such as *cucumber* and *quilt*, would occur given a substantially larger corpus. Although words of medium to high frequency are likely to have valid relative counts in the Kučera and Francis norms, estimates for those of quite low frequency are unstable because of the limited corpus.

Actually, there are three pieces of information provided for each item in those norms. The first is the total number of occurrences of that word in the corpus; the second and third provide information about the dispersion of those occurrences. The second indicates the number of genres in which the word appears and the third the number of the 500 samples in which it appears. Consider an extreme case of two items that both have total frequency counts of 10: the occurrences of one may have been in nine different passages representing eight different genres, whereas the other was in only a single passage in a single genre. Clearly, the former word is one of greater breadth of occurrence in the language; the latter has occurred several times in the count since it was a recurring word in a single passage. The list of words cited above as all having counts of 10 are, in fact, precisely such cases: words occurring 10 times in a *single* passage. The error of use of these norms is for researchers to attend only to the first number, the total frequency count, as the index of frequency. Although the data are provided that would permit one to compute some index that does take account of the degree of dispersion, no such index has been directly provided in these norms. It is undoubtedly because of this that researchers continue to refer only to the total counts.

But is this a matter of much concern? Are these particular norms very widely used, and if so are they often used to assess the frequency of rare words? This was assessed by examining the references in all the articles in 2 recent years for four journals where researchers might be likely to have use of these norms (*Journal of Memory & Language*, *Memory & Cognition*, and *Journal of Experimental Psychology: Human Perception and Performance* for 1985 and 1986, plus *Journal of Experimental Psychology: Learning, Memory and Cognition* for 1984 and 1985). The numbers of references to Kučera and Francis (1967), *American Heritage*, and Thorndike and Lorge (1944) were 43, 9, and 10, respectively. The Kučera and Francis norms are clearly the most widely employed word frequency counts. In all cases the total frequency counts were employed; no study reported adjusting these for dispersion.

For those articles referencing the Kučera and Francis norms, how were the norms being used? These norms were the basis for selecting stimulus words so as to have certain frequency properties in 10 of the 43 studies. In 13 studies they were used to match items (pairwise or the means of sets). In another 13 studies the frequency information was simply reported for the word stimuli employed, but frequency did not serve as a basis for selecting or matching items; in nearly all cases, these frequencies were high and were provided to establish that these were "common" words. The remaining 7 studies include cases in which word frequency was dependent measure for words the subjects generated, was included as one of a set of variables being correlated with some aspect of performance, or allowed selection of high- or low-frequency consonant clusters. Thus in more than half the cases, the norms were used for selecting or matching materials (23 of 43).

How often did these include low-frequency items? In 9 of the 10 studies in which words of certain frequencies were selected and in 5 of the 13 in which the norms were used to match items, items with frequencies less than 10 in Kučera and Francis's count were used (14 of 23, combined). Some experiments used words with frequencies as low as 1 or 2, both in matching word pairs (e.g., Millis, 1986) and in selecting low-frequency words (e.g., Dobbs, Friedman, & Lloyd, 1985; Forster & Davis, 1984). In at least one case the value of 0 is given because some of the words, chosen for other reasons, were not found in these norms (Kirsner, Milech, & Stumpfel, 1986).

In one study, where sets of items were selected so as to be matched in frequency, and eight sets of items had mean frequencies between 1.8 and 5.0, the author acknowledged the potential problem as follows: "All of the whole word stimuli were of low frequency according to Kučera and Francis (1967) but it is possible that sampling error causes the frequency estimates to be unreliable" (Andrews, 1986, p. 732). Andrews reported subsequently checking the frequencies of these words in the *American Heritage* frequency count, which is based on a corpus five times as large, and found that it could affect the interpretation of performance in at least some conditions where the sets were then found to be *not* closely matched.

Clearly, no actual corpus employed in any word count can contain all words of the language. Of the 104 words I could not find in the Kučera and Francis norms, 19 are also missing from the *American Heritage* word count. In general, however, the larger the corpus, the greater will be the proportion of all English words that are included in the count, and the more stable the relative frequencies of the less common words.

When items are selected from the Kučera and Francis norms simply to be "low in frequency" and these are to be contrasted to items with much higher frequency counts, the peculiarities that result from the small corpus will rarely be of much consequence. When sets of items are said to be *matched* for frequency, however, and those fre-

quencies are very low, this can be seriously misleading. And when the sets are chosen for other reasons, and one attempts to match on frequency using these norms, the words may be found to have counts of 0, which is clearly problematic.

I believe a practical implication for such cases is that the researcher might employ the *American Heritage Word Frequency Book* (Carroll et al., 1971) in preference to Kučera and Francis's norms. For one thing, the *American Heritage* word counts for rare words are more stable since they are based on a corpus five times as large. Furthermore, these norms provide, along with the total frequency count for each word, a "standard frequency index" (SFI). The SFI value is derived form the actual count total for that word *with* an adjustment for dispersion, a measure of the proportion of the 17 categories of samples in which that word occurred.

REFERENCES

ANDREWS, S. (1986). Morphological influences on lexical access: Lexical or nonlexical effects? *Journal of Memory & Language*, **25**, 726-740.

CARROLL, J. B. (1971). Measurement properties of subjective magnitude estimates of word frequency. *Journal of Verbal Learning & Verbal Behavior*, **10**, 722-729.

CARROLL, J. B., DAVIES, P., & RICHMAN, B. (1971). *The American Heritage word frequency book*. New York: American Heritage.

DOBBS, A. R., FRIEDMAN, A., & LLOYD, J. (1985). Frequency effects in lexical decisions: A test of the verification model. *Journal of Experimental Psychology: Human Perception & Performance*, **11**, 81-92.

FORSTER, K. I., & DAVIS, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory & Cognition*, **10**, 680-698.

FRANCIS, W. N., & KUČERA, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

KIRSNER, K., MILECH, D., & STUMPFEL, V. (1986). Word and picture identification: Is representational parsimony possible? *Memory & Cognition*, **14**, 398-408.

KUČERA, H., & FRANCIS, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

MILLIS, M. L. (1986). Syllables and spelling units affect feature integration in words. *Memory & Cognition*, **14**, 409-419.

SHAPIRO, B. J. (1969). The subjective estimation of relative word frequency. *Journal of Verbal Learning & Verbal Behavior*, **8**, 248-251.

THORNDIKE, E. L., & LORGE, I. (1944). *The teacher's word book of 30,000 words*. New York: Teacher's College Press, Columbia University.

(Manuscript received for publication February 24, 1988.)