Evidence-Based Selection of Word Frequency Lists

Sélection de listes de fréquence de mots basée sur des données probantes

Christopher J. Lee

Abstract

There are numerous corpora that list the frequency with which particular words occur in the English language. One factor guiding the selection and use of a corpus is the total number of words sampled in compiling its frequencies. Although it has been suggested that many words are used with different frequency in print than in speech, the results of confirmatory factor analyses reported in this paper provide little justification for distinguishing between written English and spoken English when selecting a word frequency list. Those concerned with the appropriateness of various corpora should give priority to the size of the sample of words rather than the source of the sample of words.

Abrégé

Il existe plusieurs corpus qui énumèrent la fréquence d'utilisation de certains mots dans la langue anglaise. Un des facteurs qui guide la sélection et l'utilisation d'un corpus est le nombre total de mots échantillonnés lors de la compilation de fréquence d'utilisation des mots. Même s'il a été suggéré que plusieurs mots ont une fréquence différente d'utilisation à l'écrit qu'à l'oral, les résultats de la présente étude confirmant les facteurs d'analyse fournissent très peu de raisons justifiant l'utilisation différente de listes de fréquence de mots pour l'anglais écrit et l'anglais parlé. Ceux préoccupés par le choix approprié de corpus devraient plutôt donner priorité à la grandeur de l'échantillon de mots utilisés plutôt qu'à sa source.

Key Words: word frequency lists, measurement, language assessment

It is often desirable in the study and treatment of language impairment to design lists of words. These lists might represent different exercises for therapy or different sets of stimuli for research. It is a common practice to use word frequency as a basis for estimating the level of word difficulty (Breland, 1996; Breland, Jones, & Jenkins, 1994), and speech-language pathologists (SLPs) have often turned to word frequency lists, such as Kucera and Francis (1967) or Carroll, Davies, and Richman (1971), when designing therapy materials or when constructing research stimuli.

Over the years, numerous word frequency counts have been undertaken. The earliest of these lists were compiled by Horn (1926), French, Carter, and Koenig (1930), and Thorndike (1921; Thorndike & Lorge, 1944), and the more contemporary of these lists have been compiled by Kucera and Francis (1967), Carroll et al. (1971), Dahl (1979), Howes (1966), Brown (1984), and Zeno, Ivens, Millard, and Duvvuri (1995). As there are a considerable number of lists that one might consult in order to obtain an estimate of word frequency, the purpose of this paper is to suggest a basis for choosing among them.

Christopher J. Lee
The University of Western Ontario
School of Occupational Therapy
London, Ontario

Word frequency lists differ in two general respects. First, some lists have been compiled from written English, whereas others have been compiled from spoken English. Second, lists vary in the total number of words sampled. A greater number of lists have been compiled from written English than from spoken English, and the number of words sampled in compilation of a particular word frequency list has tended to be greater when sampling written English than when sampling spoken English. This apparent trade-off between the source of the sample and the size of the sample can be an important consideration in obtaining reliable estimates of word frequency. Breland, Jones, and Jenkins (1994), for instance, have suggested that even a sample of several million words can be insufficient for the study of college-level vocabulary given that many of the words of interest are used relatively infrequently.

Frequency estimation can be considered more accurate when based on a larger sample than when based on a smaller sample due to the fact that, statistically, the standard error varies as a function of the square root of sample size. This means that there is less random error associated with frequency estimation in larger samples. In this respect, when choosing among available corpora of word frequency, it would be desirable to select the one based on the largest sample. Thus, for example, if one had access to both *The Educator's Word Frequency Guide* (Zeno et al., 1995) and the *Word Frequencies of Spoken American English* (Dahl, 1979), it would be advisable to use the Zeno et al. corpus because it was compiled using a sample of more than 17 million words, whereas the Dahl corpus was compiled from a sample of approximately one million words.

However, it has been argued that a substantial number of words are used differently in print than in speech (Dahl, 1979; Howes, 1966; Tryk, 1968). Tryk noted that the usual sources of written English are materials produced by professional writers whose language is modified by editorial practices and aesthetic concerns that do not constrain ordinary speech. For example, Dahl suggests that there is a greater tendency to introduce synonyms as a means of avoiding repetition of the same word when writing than when speaking. In addition, some varieties of words, such as profanity, may tend to occur less frequently in print than in speech, whereas other varieties of words, such as arcane words, may tend to occur more frequently in print than in speech (Dahl, 1979).

Despite these claims, there has not been a systematic examination of the extent to which words are used differently in print than in speech. It is certainly possible to compare two corpora and find differences in the reported frequencies. For example, the word *system* is reported to occur in print about 414 times per million (Kucera & Francis, 1967), but to occur in speech only about 16 times per million (Dahl, 1979). Undoubtedly, one could find other cases of disparity by culling various corpora. But these apparent differences can belie substantial commonality given the ubiquitous effect of sampling error, and a more general examination of the relationship between written-frequency and spoken-frequency is needed.

In the present study, confirmatory factor analysis was used to examine the general relationship between written-frequency and spoken-frequency. Confirmatory factor analysis allows one to test an assumed model of the relationship between measured variables and one or more underlying theoretical constructs or *factors*. A researcher specifies which variables serve as indicators of specific factors and evaluates the extent to which this conceptual model fits the observed pattern of relationships among the variables. In other words, a researcher evaluates the extent to which the data confirm the hypothesized model. An evaluation of the fit of the model is made on the basis of factor loadings and indices of overall fit. Factor loadings are correlations between an underlying construct and the variables modelled as indicators of the construct. Higher loadings indicate a stronger relationship between a construct and a variable. Various fit indices can be used to gauge the overall fit between the model and the data. Indices such as CFI and GFI are scaled to have a maximum value of one, and values of .90 or greater are usually taken to indicate a good fit between the model and the data. A chi square statistic is also reported summarily, but its magnitude is not readily interpretable.

In the present study, a one-factor model in which the written-frequency and spoken-frequency corpora are treated as indicators of a single, common factor is compared to a two-factor model in which the written-frequency and spoken-frequency corpora are treated as indicators of separate factors. To the extent that the two-factor model is found to fit the data substantially better than the one-factor model, there will be empirical evidence to support the claim that differences in word use in print and in speech have a substantive general effect on the estimation of word frequencies, and there will be reason to recommend selecting among word frequency corpora on the basis of the written or spoken nature of the sample. On the other hand, if it is found that the one-factor model fits the data just as well as the two-factor model, there will be evidence of a considerable degree of commonality among the corpora, and there will be reason to consider the size of the sample as being a more suitable basis for choosing among available word frequency corpora.

Method

Three spoken-frequency corpora (Brown, 1984; Dahl, 1979; Howes, 1966) and three written-frequency corpora (Carroll et al., 1971; Kucera & Francis, 1967; Zeno et al., 1995) were selected for analysis. All three of the spoken-frequency corpora were derived from adult speech, whereas a diverse range of reading levels, from early grades to adult levels, is reflected in the three written-frequency corpora. A sample of 500 words was randomly drawn from each of the corpora, yielding an initial pool of 3000 words. Some words, however, occurred in two or more of the individual samples, and after removing replications, the final sample consisted of 2807 different words. The frequency listed for these words in each of the six corpora was recorded. All frequencies were transformed to natural logarithms to correct for skewness, and as some of the words had a frequency of zero in one or more corpora, an arbitrary increment of 0.01 was added to the frequency of each word to allow transformation.

Table 2
Factor Loadings in the One-Factor and Two-Factor Models

	One-Factor Model	Two-Factor Model		Size
		Written	Spoken	
Corpus				
1. Zeno et al. (1995)	.94	.95		17.27
2. Carroll et al. (1971)	.88	.88		5.09
3. Kucera & Francis (1967)	.85	.84		1.01
4. Dahl (1979)	.81		.85	1.06
5. Howes (1966)	.77		.81	0.20
6. Brown (1984)	.73		.77	0.19

Note. The values listed for size indicate the number of words of text or speech sampled, expressed in millions.

Results and Discussion

The correlations among the six corpora of word frequency are presented in Table 1. The correlations among the three written-frequency corpora, ranging from .73 to .84 (mean $r = .79$), were higher than the correlations among the three spoken-frequency corpora, ranging from .64 to .68 (mean $r = .66$). The correlations between the written-frequency and spoken-frequency corpora ranged from .61 to .74 (mean $r = .67$). It is worth noting that the spoken-frequency corpora correlated with the written-frequency corpora to approximately the same extent as they correlated with each other.

Confirmatory factor analyses were performed on the covariances among the frequency corpora using maximum likelihood estimation. A one-factor model was estimated in which the written-frequency and spoken-frequency corpora served as indicators of a single factor, and a two-factor model was estimated in which the written-frequency corpora (Carroll et al., 1971; Kucera & Francis, 1967; Zeno et al., 1995) served as indicators of a written-frequency factor and the spoken-frequency corpora (Brown, 1984; Dahl, 1979; Howes, 1966) served as indicators of a spoken-frequency factor. In the two-factor model, the factors were allowed to correlate. In both the one-factor and two-factor models, measurement errors were estimated orthogonally.

Table 1
Correlations Among Corpora of Written and Spoken Word Frequency

Corpus	Variance	1	2	3	4	5
1. Zeno et al. (1995)	7.23					
2. Carroll et al. (1971)	10.40	.84				
3. Kucera & Francis (1967)	9.79	.80	.73			
4. Dahl (1979)	13.20	.74	.67	.70		
5. Howes (1966)	10.72	.70	.65	.65	.68	
6. Brown (1984)	10.86	.66	.61	.61	.65	.64

Note. The first three corpora were compiled from written English, and the latter three were compiled from spoken English

The one-factor model provided an excellent fit to the data (CFI = .97; GFI = .95; $\chi^2_{(9)} = 382.79$). As shown in Table 2, the loadings of the written-frequency and spoken-frequency corpora on the single factor ranged from .73 to .94, with written-frequency corpora exhibiting stronger factor loadings than the spoken-frequency corpora. In comparison to the one-factor model, the two-factor model provided a slightly better fit to the data (CFI = .99; GFI = .99; $\chi^2_{(8)} = 84.52$). The loadings of the three spoken-frequency corpora were modestly stronger in the two-factor model than in the one-factor model, whereas the loadings of the three written-frequency corpora differed very little in the two analyses. In the two-factor model, the written-frequency and spoken-frequency factors were found to be very highly correlated ($r = .92$), suggesting that

the written-frequency and spoken-frequency factors are largely indistinguishable. Given the excellent fit of the one-factor model, the negligible change in factor loadings and overall fit afforded by the two-factor model, and the high correlation between the written-frequency and spoken-frequency factors in the two-factor model, there is little empirical basis for distinguishing between written and spoken frequency.

Table 2 also indicates the size of the sample used in compiling each corpus. It is worth noting that there is a strong correspondence between the magnitude of a factor loading and the size of the sample. As apparent in Table 2, stronger factor loadings were consistently obtained with larger samples. As statistically the standard error varies as a function of the square root of sample size, the correlation between the factor loadings and the square root of the number of words sampled can provide a means of appraising the extent to which differences in the magnitudes of these factor loadings are attributable to the size of the sample. In the one-factor model, the magnitude of a factor loading correlated .91 with the square root of the number of words sampled, and in the two-factor model, the magnitude of a factor loading correlated .95 with the square root of the number of words sampled. These correlations suggest that corpora based on a larger sample are much better indicators of word frequency.

In sum, there is little basis for systematically distinguishing between written and spoken frequency given the excellent fit of the one-factor model, the negligible change in factor loadings and overall fit afforded by the two-factor model, and the high correlation between the written-frequency and spoken-frequency factors in the two-factor model. In light of the evident effect of sample size, it would be appropriate to give priority to the size of the sample rather than the source of the sample when selecting an appropriate corpus for determining the frequency with which particular words occur in the English language. Those concerned with the appropriateness of various corpora should choose the corpus with the largest sample size without regard to whether it was derived from spoken or written English.

Author Note

Please address all correspondence to Christopher Lee, PhD, School of Occupational Therapy, The University of Western Ontario, London, Ontario, Canada, N6G 1H1; email: cjlee@uwo.ca.

References

Breland, H. M. (1996). Word frequency and word difficulty: A comparison of counts in four corpora. *Psychological Science, 7*, 96-99.

Breland, H. M., Jones, R. J., & Jenkins, L. (1994). *The College Board vocabulary study* (College Board Report No. 94-4; Educational Testing Service Research Report No. 94-26). New York: College Entrance Examination Board.

Brown, G. D. A. (1984). A frequency count of 190,000 words in the London-Lund corpus of English conversation. *Behaviour Research Methods, Instruments, & Computers, 16*, 502-532.

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American Heritage word frequency book*. New York: American Heritage.

Dahl, H. (1979). *Word frequencies of spoken American English*. Essex, CT: Verbatim.

French, N., Carter, C. W., & Koenig, W. (1930). The words and sounds of telephone conversations. *Bell System Technical Journal, 9*, 290-324.

Horn, E. (1926). *A basic writing vocabulary: 10,000 frequently used words in writing*. [Monograph First Series, No. 4.] Iowa City: University of Iowa.

Howes, D. (1966). A word count of spoken English. *Journal of Verbal Learning and Verbal Behavior, 5*, 572-606.

Kucera, H. & Francis, W. H. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University.

Thorndike, E. L. (1921). *The teacher's word book*. New York: Teachers College, Columbia University.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Teachers College, Columbia University.

Tryk, H. E. (1968). Subjective scaling of word frequency. *American Journal of Psychology, 81*, 170-177.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. New York: Touchstone Applied Science Associates.

Manuscript received: *January 3, 2003*
Accepted: *April 29, 2003*