Behavior Research Methods, Instruments, & Computers
2002, 34 (2), 286-290

Using Internet search engines to estimate word frequency

IRENE V. BLAIR and GEOFFREY R. URLAND
University of Colorado, Boulder, Colorado

and

JENNIFER E. MA
University of Kansas, Lawrence, Kansas

The present research investigated Internet search engines as a rapid, cost-effective alternative for estimating word frequencies. Frequency estimates for 382 words were obtained and compared across four methods: (1) Internet search engines, (2) the Kučera and Francis (1967) analysis of a traditional linguistic corpus, (3) the CELEX English linguistic database (Baayen, Piepenbrock, & Gulikers, 1995), and (4) participant ratings of familiarity. The results showed that Internet search engines produced frequency estimates that were highly consistent with those reported by Kučera and Francis and those calculated from CELEX, highly consistent across search engines, and very reliable over a 6-month period of time. Additional results suggested that Internet search engines are an excellent option when traditional word frequency analyses do not contain the necessary data (e.g., estimates for forenames and slang). In contrast, participants' familiarity judgments did not correspond well with the more objective estimates of word frequency. Researchers are advised to use search engines with large databases (e.g., AltaVista) to ensure the greatest representativeness of the frequency estimates.

Word frequency is an important experimental variable, with the potential to influence any cognitive phenomenon that involves language. Domains that have been shown to benefit from a consideration of word frequency include word recognition (e.g., Balota & Rayner, 1991), lexical decision latency (e.g., Rubenstein, Garfield, & Millikan, 1970), memory (e.g., Chalmers, Humphreys, & Dennis, 1997), language acquisition (e.g., Brysbaert, Lange, & Wijnendaele, 2000), and unitization in reading (e.g., Peterzell, Sinclair, Healy, & Bourne, 1990). Even when word frequency is not the focus of attention, researchers routinely control for it in their experiments to eliminate a potentially powerful confound.

One of the most popular methods for determining word frequency has been to analyze a linguistic corpus, which consists of a sample of texts intended to be a representative reference for that language (McEnery & Wilson, 1996). Some of the most commonly used frequency analyses are Thorndike and Lorge (1944), Kučera and Francis (1967; see also the extended 1982 analysis of the same corpus by Francis & Kučera), and those obtained from the CELEX linguistic database (Baayen, Piepenbrock, & Gulikers, 1995). The popularity of these analyses, however, belies several potential problems in their usage. Because each corpus is composed of written texts, they may not fully represent the frequency with which some words are used in spoken English (Chomsky, 1965). In addition, a traditional corpus is unlikely to contain certain types of words, such as slang and forenames, that are not common in formal written texts but are used with some frequency in everyday life and are central for some experiments (e.g., Blair & Banaji, 1996; Devine, 1989; Greenwald, McGhee, & Schwartz, 1998).

Another potential problem concerns the age of the popular corpora analyses. Kučera and Francis (1967) is over 30 years old, and Thorndike and Lorge (1944) is almost 60 years old. In light of rapid changes in vocabulary, researchers may question whether those frequency counts are still valid (Gernsbacher, 1984). The CELEX database (Baayen et al., 1995) contains contemporary samples of written English, but its cost is more than some researchers may be willing to spend (currently $150), and without constant updating it too will go out of date. Finally, access to even the older frequency analyses is restricted because they are out of print and difficult to obtain. Many research institutions have only one or two copies that must be shared among all researchers.

In light of the limitations of traditional word frequency analyses, the goal of the present research was to examine the potential of using Internet search engines to provide valid and reliable information on word frequency. Search engines operate by sending automated agents (known as

This work was supported by NIH Grant MH63372 to I.V.B., a NSF Graduate Research Fellowship to G.R.U., and an NIH postdoctoral fellowship to J.E.M. We thank Alice Healy and the University of Colorado Stereotyping and Prejudice (CUSP) Lab for their helpful comments on an earlier draft of the paper. We also thank Gary McClelland and Lou McClelland for their assistance with the sampling analyses. Correspondence should be addressed to I. V. Blair, Department of Psychology, University of Colorado, Boulder, CO 80309-0345 (e-mail: irene.blair@colorado.edu).

"spiders") out on the Internet. These spiders, in turn, send information about a site's content back to a database, which can be searched by multiple users. For example, one may ask the engine to search for the word *desk*. When the search is completed, the user is provided with a report on the number of times the word was found in the database—commonly known as the "hit" count.[1] We argue that this count is analogous to a conventional word frequency estimate, and it can be compared with the hit count for other words of interest.

Internet search engines solve many of the drawbacks of traditional frequency analyses. The Internet is ubiquitous and search engines are generally free sites, making issues of availability and cost nearly irrelevant. Moreover, the Internet is relatively comprehensive, including academic texts, commercial and personal information, and records from newsgroup postings. The latter source of information is similar to spoken language and gives the Internet an additional advantage over corpora that rely on more formal written language. Because anyone can post information on the Internet, it may also be more representative of "everyone's language." Likewise, the fast-paced nature of the Internet and the fact that search engines constantly update their databases provide a way for the search engines to reflect contemporary language usage. Thus, the Internet provides a linguistic database that is relatively comprehensive, representative, contemporary, and easily searched.

However, it is necessary to determine empirically whether search engines provide valid estimates of word frequency. In addition, the fluid nature of the Internet may undermine the reliability of estimates based on Internet databases. These questions were addressed in the following study by obtaining word frequency estimates for a large sample of words from four popular Internet search engines and comparing them with the estimates obtained from Kučera and Francis (1967) and the CELEX linguistic database (Baayen et al., 1995), as well as to participant ratings of the words' familiarity. The test–retest reliability of the search engines was also examined by conducting a second set of Internet searches, 6 months later.

METHOD

Test Sample

The test sample was composed of 382 words. The majority of the words ($n = 250$) were standard English words that included a selection of nouns, verbs, and adjectives (e.g., *attain*, *dishonest*, *nurse*, *welfare*). The frequency of these words, according to Kučera and Francis (1967), ranged from 0 to 1,303 ($M = 102.41$). In addition, a subsample of 132 nonstandard words was added. This subsample contained 36 slang terms (e.g., *bimbo*, *rad*, *reefer*, *yuppie*) and 96 forenames (48 male and female European American names and 48 male and female African American names, taken from Greenwald et al., 1998). According to Kučera and Francis, these words ranged in frequency from 0 to 92 ($M = 5.82$).

Frequency Estimates

Four methods were used to estimate the frequency of the 382 words in the sample: (1) Kučera and Francis (1967), (2) CELEX linguistic database (Baayen et al., 1995), (3) participant ratings of familiarity, and (4) Internet search engines. These methods are described below.

Kučera and Francis analysis. Frequency estimates for the words in the test sample were obtained from Kučera and Francis's (1967) *Computational Analysis of Present-Day American English*. If a word did not appear in the database, the frequency was recorded as zero.

CELEX linguistic database. Frequency estimates for the words in the test sample were obtained by electronically searching the CELEX linguistic database (Baayen et al., 1995). This CD-ROM database contains 160,594 words from 284 written texts. If a word did not appear in the database, the frequency was recorded as zero.

Participant ratings. Because pretesting is often used to select stimuli that are matched on a particular dimension (e.g., imageability), some researchers may use participants' subjective familiarity with words as an alternative to obtaining objective word frequency estimates. To explore the validity of this method, 33 undergraduates at the University of Kansas were asked to rate the familiarity of each word in the sample, using a 5-point scale with labeled endpoints (1 = *very unfamiliar*, 5 = *very familiar*). The participants were asked to consider "how common or frequently you have encountered the word, or how well you know the word." The 382 words were divided into two lists of equal length, with the words on each list presented in a single fixed order. The participants rated the familiarity of the words on one list and then, following a 5-min break, they rated the familiarity of the second list of words. The order of the two word lists was counterbalanced across the participants. Cronbach's alpha for interrater reliability was .97.

Internet search engines. Four search engines were selected for the study: AltaVista, Northern Light, Excite, and Yahoo! These search engines were chosen primarily for their popularity among Internet users. Database and search technique were also considered. AltaVista (www.altavista.com) and Northern Light (www.northernlight.com) were included because they are two of the largest and most comprehensive Internet search engines. When the analyses were conducted, AltaVista had a database of more than 250 million webpages and Northern Light had over 200 million webpages (AllSearchEngines.Com, 2000; Kansas City Public Library, 2000; Leita, 2000). During a search with these engines, the full text of webpages and articles in their databases are searched for word matches. In contrast, Excite (www.excite.com) and Yahoo! (www.yahoo.com) have relatively smaller databases (150 million and 2 million, respectively). In addition, Yahoo! conducts its searches in a very different manner from the other three engines. Specifically, it is an Internet subject directory that searches for general topics as opposed to keyword matches. As a consequence, its frequency estimates may not be as valid or reliable.

Each word in the sample was entered into the search function of each of the four search engines. The number of hits returned was then recorded as the frequency estimate for that word. To examine the reliability of frequency estimates obtained from the search engines, the same search process was repeated 6 months later.

RESULTS

As expected for frequency data, the word counts from the search engines, Kučera and Francis (1967), and CELEX (Baayen et al., 1995) were positively skewed. Thus, a standard square-root transformation was applied before further analyses (Judd & McClelland, 1989). In contrast, the participants' ratings of familiarity were negatively skewed. Because this skew was relatively minor and we believed that it reflected an important psychological reality for the participants (see Discussion), those data were left untransformed.

Due to differences in database size, the two larger search engines, AltaVista and Northern Light, returned

Table 1
Correlation Coefficients Among Frequency Estimates for the Full Word Sample (N = 382)

	CELEX	PR	AV	NL	EX	YH
Kučera & Francis (1967)	.92	.48	.81	.89	.78	.69
CELEX		.46	.76	.81	.71	.62
Participant ratings (PR)			.45	.49	.46	.47
Search engines						
AltaVista (AV)			.94	.91	.81	.73
Northern Light (NL)				.96	.84	.76
Excite (EX)					.94	.88
Yahoo! (YH)						.84

Note—Coefficients on the diagonal for the search engines are the test–retest reliability estimates. All correlations are significant at $p < .0001$.

significantly higher estimates, on average, than the two smaller search engines (M = 2.6 million vs. 0.7 million). And all of the search engines produced higher estimates than Kučera and Francis (M = 69.03) or CELEX (M = 925.64). Of greater importance, however, was the validity and reliability of the search engines as determined by comparisons of word frequency estimates *among* the words in the test sample. The following tests were conducted.

First, correlations were calculated among the frequency estimates obtained using each of the four methods. Table 1 shows that the estimates obtained from the four search engines were highly correlated with those obtained from Kučera and Francis (mean r = .79) and with the estimates provided by CELEX (mean r = .72). In contrast, the participants' word ratings were only moderately correlated with the other frequency counts.

Second, correlations were calculated among the four search engines. As shown in Table 1, the search engines provided highly consistent estimates of word frequency on the Internet (mean r = .82).

Finally, the test–retest reliabilities of the search engines were examined by calculating correlations between the word estimates obtained at Time 1 and at Time 2. These correlations, provided in Table 1, showed that the search engines produced highly reliable estimates over the 6-month period of time (mean r = .92).

Frequency Estimates for Subsamples of Words

As noted, the full test sample was composed of 250 standard words and 132 nonstandard words. To examine the congruence among the frequency methods for the two subsamples, the analyses were repeated within each sample. These analyses showed that for both subsamples, the Internet search engines produced frequency estimates that were very reliable in terms of their consistency with one another (mean r = .79 and .89, for the standard and nonstandard samples, respectively) and their consistency across time (mean r = .91 and .89, for the standard and nonstandard samples, respectively). However, the congruence between the search engines and the other three methods was different for the two subsamples of words (Table 2).

First, the congruence between the search engines and the two traditional linguistic databases was higher for the standard than the nonstandard words (mean r = .78 vs. .57, $z = 3.65$, $p < .001$ for Kučera & Francis; mean r = .70 vs. .44, $z = 3.64$, $p < .001$ for CELEX). One explanation for this discrepancy is that the linguistic databases did not contain estimates for many of the nonstandard words. Specifically, the Kučera and Francis database was missing 51% of the nonstandard words (54% of the forenames and 42% of the slang terms), compared with only 3% missing for the standard words; CELEX was missing 66% of the nonstandard words (84% of the forenames and 17% of the slang terms), compared with only 2% missing for the standard words. A very large number of zero estimates could have attenuated the correlation for the nonstandard words. However, even when all words with a zero estimate were eliminated from the analysis, the correlation between the search engines and the standard databases continued to be higher for the standard than the nonstandard words (mean r = .77 vs. .56, z = 2.72, $p < .01$ for Kučera & Francis; mean r = .70 vs. .50, $z = 1.90$, $p = .058$ for CELEX). We cannot be certain why these differences exist. However, it doesn't seem so surprising that the type of "common" English used on the

Table 2
Correlation Coefficients Between the Internet Search Engines and the Other Three Methods of Obtaining Frequency Estimates, for the Standard and Nonstandard Word Samples

	Standard Word Sample					Nonstandard Word Sample				
Method	PR	AV	NL	EX	YH	PR	AV	NL	EX	YH
Kučera & Francis (1967)	.40	.85	.88	.76	.64	.46	.46	.61	.60	.59
CELEX	.36	.78	.79	.68	.56	.46	.44	.45	.44	.43
Participant ratings		.40	.43	.39	.40		.59	.61	.60	.63

Note—All correlations are significant at $p < .0001$. PR, participant ratings; AV, AltaVista; NL, Northern Light; EX, Excite; YH, Yahoo!

Web and the more formal writing contained in the traditional linguistic corpora may differ in the frequency with which various slang terms and forenames are used.

The second difference observed between the two word samples was that the congruence between the search engines and the participants' ratings was *lower* for the standard than for the nonstandard words (mean $r = .40$ vs. $.61$, $z = 2.62$, $p < .01$). This finding suggests that the participants may have had an easier time making familiarity distinctions among relatively unfamiliar words. Nonetheless, in neither case was the correlation very high.

Because researchers may question whether they can rely on Internet search engines for small samples of words, correlational analyses were conducted on 100 randomly selected samples of 30 standard words. Median correlations and their interquartile ranges based on these analyses are presented in Table 3. These numbers showed that even for relatively small samples of words, the Internet search engines produced word frequency estimates that were highly consistent with the two standard databases, highly consistent with one another, and highly consistent across time. The two smaller search engines (EX and YH), however, returned results that were a little less consistent with the standard databases than those obtained from the two larger search engines (AV and NL).

DISCUSSION

The present research demonstrates that Internet search engines provide word frequency estimates that are both valid and reliable. The estimates obtained from the four search engines showed good convergent validity with both Kučera and Francis (1967) and CELEX (Baayen et al., 1995), were highly consistent with one another, and showed excellent test–retest reliability over a 6-month period of time. These results ought to encourage researchers to take advantage of this highly accessible and easy-to-use method.

The high convergence between the search engines and Kučera and Francis (1967) also suggests that despite its age, Kučera and Francis is still a valid source for word frequencies. We have shown, however, that one of the greatest drawbacks of that method—and similar databases, such as CELEX—is missing data. The lack of data for forenames is especially problematic for social psychologists who frequently employ forenames as stimuli and have few available means to estimate their frequency (for discussions of this problem, see Dasgupta, McGhee, Greenwald, & Banaji, 2000; Kasof, 1993). The lack of data for slang terms suggests that Kučera and Francis and CELEX may also not be a good source of frequency data when the words are relatively new to the lexicon or appear more often in speech than writing. The Internet search engines, in contrast, produced highly consistent and reliable word frequency estimates for both the standard and nonstandard words, suggesting that they can be used where other methods fail.

In addition to being an easy-to-use, cost-effective method of obtaining word frequencies, search engines may also open up other avenues for research. For example, by treating the Web as a linguistic database, researchers can conduct analyses of the contexts surrounding certain words. Such analyses could be informative in regard to the typical user of a word (e.g., age, education, culture) and the objects and social roles that are most often associated with it. An analysis of the surrounding context may also provide the researcher with a better sense of how familiar people really are with a particular word. If a word is most often listed in technical or otherwise specialized webpages, then it may not be as familiar to the average person as a word that is found on more mainstream webpages. Another advantage of using search engines for frequency analyses is the potential to search for phrases as well as for single words. For example, one might wonder if *baseball bat* or *hockey stick* occurs with greater frequency, or whether people are more familiar with "To be or not to be" or "I think therefore I am." (In both cases, the former phrase occurs with much greater frequency than the latter.) Finally, it is important to note that search engines can be used to estimate word frequencies for languages other than English (see New, Pallier, Ferrand, & Matos, in press). Researchers who use words from more than one language may find it useful to conduct word frequency analyses with the same basic method. However, we caution that the validity of such searches would depend on the extent to which speakers of the language use the Web.

Table 3
Median Correlation Coefficients (Mdns) and Interquartile Ranges (IRs) Based on Analyses of 100 Random Samples of 30 Standard Words

	CELEX		PR		AV		NL		EX		YH	
Method	Mdn	IR	Mdn	IR	Mdn	IR	Mdn	IR	Mdn	IR	Mdn	IR
Kučera & Francis (1967)	.91	.06	.42	.14	.85	.09	.89	.07	.77	.26	.65	.36
CELEX			.41	.11	.78	.10	.80	.10	.72	.29	.63	.36
Participant ratings (PR)					.44	.11	.47	.13	.43	.12	.43	.12
Search engines												
AltaVista (AV)					.98	.06	.93	.06	.85	.27	.71	.36
Northern Light (NL)							.99	.03	.82	.26	.72	.39
Excite (EX)									.98	.09	.97	.23
Yahoo! (YH)											.92	.30

Note—Coefficients on the diagonal for the search engines are the test–retest reliability estimates.

The present research also provided evidence on the validity of participants' subjective ratings of familiarity as an alternative measure of word frequency. The inconsistencies between such ratings and the other methods suggest that subjective familiarity is not equivalent to more objective measures of word frequency (see also Gernsbacher, 1984). In particular, the present data showed that the familiarity ratings were negatively skewed, whereas the other estimates were positively skewed. That negative skew reveals that the raters did not make distinctions among words that are relatively familiar but have very different objective frequencies in the language (e.g., *lazy* vs. *school*). This discrepancy was especially pronounced in the standard word sample, where the negative skew was greater than in the nonstandard word sample (−2.34 vs. −0.56, respectively). Other researchers have cautioned, however, that subjective familiarity should not be discounted as an important variable in cognition despite its differences from objective word frequency (Gernsbacher, 1984).

Although the present data provided strong evidence in favor of Internet search engines as a method of estimating word frequency, two caveats are in order. First, the two smaller search engines (Excite and Yahoo!) produced somewhat less consistent estimates with a relatively small sample of words. Thus, it is recommended that the larger search engines be used as a general rule because they have more representative databases. Second, Internet search engines are best used when relative word frequency estimates are satisfactory. With 463,830 hits in AltaVista for *brush* and 4,860,810 hits for *earth*, we know that the latter word occurs more frequently than the former word. However, with only a rough estimate of the total database (approximately 250 million) and the fact that it is always changing, the absolute frequencies of those words cannot be determined with any certainty. For many research purposes, relative word frequencies are the only estimates of interest, and it is for those studies that Internet search engines provide an excellent option.

REFERENCES

AllSearchEngines.Com homepage (May, 2000). Available: http://www.allsearchengines.com.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* [CD-ROM]. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Balota, D. A., & Rayner, K. (1991). Word recognition processes in foveal and parafoveal vision: The range of influence of lexical variables. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 198-232). Hillsdale, NJ: Erlbaum.

Blair, I. V., & Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality & Social Psychology*, **70**, 1142-1163.

Brysbaert, M., Lange, M., & Wijnendaele, I. V. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition: Further evidence from the Dutch language. *European Journal of Cognitive Psychology*, **12**, 65-85.

Chalmers, K. A., Humphreys, M. S., & Dennis, S. (1997). A naturalistic study of the word frequency effect in episodic recognition. *Memory & Cognition*, **25**, 780-784.

Chomsky, N. (1965) *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Dasgupta, N., McGhee, D. E., Greenwald, A. G., & Banaji, M. R. (2000). Automatic preference for White Americans: Eliminating the familiarity explanation. *Journal of Experimental Social Psychology*, **36**, 316-328.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality & Social Psychology*, **56**, 680-690.

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

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, **113**, 256-281.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality & Social Psychology*, **74**, 1464-1480.

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model comparison approach*. San Diego: Harcourt Brace Jovanovich.

Kansas City Public Library (2000, March). *Introduction to search engines*. Available: http://www.kcpl.lib.mo.us/search.

Kasof, J. (1993). Sex bias in the naming of stimulus persons. *Psychological Bulletin*, **113**, 140-163.

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

Leita, C. (2000, May). InfoPeople Search Tools Chart. Available: 2000 InFoPeople Project at http://infopeople.org/src/chart.html.

McEnery, T., & Wilson, A. (1996). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

New, B., Pallier, C., Ferrand, L., & Matos, R. (in press). Une base de données lexicales du français contemporain sur internet: LEXIQUE [A lexical database of contemporary French on the Internet: LEXIQUE]. *L'Année Psychologique*.

Peterzell, D. H., Sinclair, G. P., Healy, A. F., & Bourne, L. E. (1990). Identification of letters in the predesignated target paradigm: A word superiority effect for the common word *the*. *American Journal of Psychology*, **103**, 299-315.

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning & Verbal Behavior*, **9**, 487-494.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words* (3rd ed.). New York: Columbia University, Teachers College Press.

NOTE

1. Some search engines report both the number of exact word matches and the number of websites that contain the word. It is the former hit count that provides the more accurate word frequency count.

(Manuscript received February 15, 2001;
revision accepted for publication December 15, 2001.)