Behavior Research Methods, Instruments, & Computers
2004, 36 (3), 388-396

Case-sensitive letter and bigram frequency counts from large-scale English corpora

MICHAEL N. JONES and D. J. K. MEWHORT
Queen's University, Kingston, Ontario, Canada

We tabulated upper- and lowercase letter frequency using several large-scale English corpora (~183 million words in total). The results indicate that the relative frequencies for upper- and lowercase letters are not equivalent. We report a letter-naming experiment in which uppercase frequency predicted response time to uppercase letters better than did lowercase frequency. Tables of case-sensitive letter and bigram frequency are provided, including common nonalphabetic characters. Because subjects are sensitive to frequency relationships among letters, we recommend that experimenters use case-sensitive counts when constructing stimuli from letters.

Familiar patterns can be identified more easily and quickly than unfamiliar ones. Common examples include the advantage for high- over low-frequency words (e.g., Broadbent, 1967; Howes & Solomon, 1951; Krueger, 1975) and a corresponding advantage for high- over low-frequency letters (e.g., Appelman & Mayzner, 1981; Bryden, 1968). Letters and strings of letters are commonly used stimuli in experimental psychology; hence, it is important to anticipate potential characteristics of letters that may affect results in experiments using them as stimuli.

High-frequency letters show an advantage over low-frequency letters in letter naming (Cosky, 1976; Jones, 2001), same–different matching (Egeth & Blecker, 1971; Krueger, 1973b, 1973c), and visual search (Latimer, 1972). In addition, bigram and trigram frequency affects anagram solutions (Dominowski & Duncan, 1964) and word recognition (Broadbent & Gregory, 1968; Rice & Robinson, 1975).

Subjects are sensitive to frequency relationships among letters even with stimuli that have never been seen before. Pseudowords that vary in order of approximation to English are a prime example. Zero-order approximations are created by taking letters at random; first-order approximations are created by taking letters at random from a text, thus preserving the frequency of individual letters. Higher orders of approximation preserve the frequency of larger units: bigrams for second-order approximations, trigrams for third-order approximations, and so forth. Subjects report more letters from higher order than from lower order pseudowords (Mewhort & Campbell, 1981; Miller, Bruner, & Postman, 1954).

Because frequency is a potential influence in any experiment using letters or strings of letters, it is important to have an adequate measure of letter frequency. Several methods have been tried. A quick and easy method with which to estimate letter frequency is to conduct *initial-letter counts*—that is, to count the occurrence of initial letters in words from a dictionary or word-frequency norm (e.g., Latimer, 1972; Tinker, 1928, Table 5).

Accurate tabulations of frequency require large samples of representative text. Unfortunately, the tabulations currently available have ignored letter case (e.g., Baddeley, Conrad, & Thompson, 1960; Gaines, 1939; Mayzner & Tresselt, 1965; Solso & King, 1976). As a result, existing counts of letter frequency are based on a mix of upper- and lowercase letters, dominated by the lowercase letters, even though most experiments use uppercase letters as stimuli.

As part of a project on letter matching, we wanted data to compare the relative recognizability of letters and turned to existing confusion matrices: three uppercase matrices (Gilmore, Hersh, Caramazza, & Griffin, 1979; Townsend, 1971; van der Heijden, Malhas, & van den Roovaart, 1984) and three lowercase matrices (Bouma, 1971, distance and eccentricity; Geyer, 1977). We examined the confusion matrices' diagonal, which records the relative recognizability of the letters.

The mean correlation across the diagonals of the upper- and lowercase confusion matrices was only −.076. The near-zero correlation indicates very large differences in recognizability as a function of case. Part of the difference reflects differences in shape for upper- and lowercase letters, but part of the difference may also reflect differences in the relative frequency of upper- and lowercase letters. When we checked frequency for upper- and lowercase letters, we were unable to find suitable case-sensitive frequency counts.

This research was supported by a grant to D.J.K.M. from the Natural Sciences and Engineering Research Council of Canada (NSERC) and by an AEG grant from Sun Microsystems of Canada. M.N.J. was supported by a postgraduate research scholarship from NSERC. Correspondence concerning this article should be addressed to M. N. Jones, Department of Psychology, Queen's University, Kingston, ON, K7L 3N6 Canada (e-mail: mike@minerva.psyc.queensu.ca or mewhortd@post.queensu.ca).

Do Upper- and Lowercase Letters Have the Same Relative Frequencies?

Tinker (1928) is the only published study to conduct a case-sensitive letter count. Tinker ranked the frequency of occurrence for upper- and lowercase letters from a type foundry's font-replacement table and confirmed the estimate by counting the upper- and lowercase letters in "several pages of two journals" (p. 490). The rank order correlation between Tinker's upper- and lowercase counts was impressively high (r_s = .93). Tinker's evidence has been accepted at face value. In their examination of the letter-frequency effect, for example, Appelman and Mayzner (1981) used Tinker's counts to conclude that "the order of letter frequency is quite similar for upper- and lowercase letters" (p. 437).

Tinker's (1928) evidence, however, is open to question. His counts were based on a limited sample of the English language. Furthermore, to compute a rank correlation, Tinker reduced frequency to an ordinal scale. His transformation limited the variability among letters because he assigned a large number of tied ranks: 23 uppercase letters and 5 lowercase letters were assigned a tied rank with at least one other letter (and in some instances, three or four other letters). Because we did not trust a conclusion based on such a limited sample of text, we reexamined case-sensitive frequency using several large and representative corpora of English text.

EXAMINING FREQUENCY ACROSS CASE

Word Corpus Selection

In selecting a word corpus, our goal was to find one that was representative of text that humans regularly experience. We examined (1) full-text articles from the *New York Times* (NYT) from January to March 1992 (~14 million words), (2) a subset of the Brown word corpus (Kučera & Francis, 1967; ~1 million words), (3) a commonly used online encyclopedia (~7 million words), (4) text extracted from about 100,000 randomly selected Web pages[1] (~61 million words), and (5) newsgroup text extracted from 400 different Internet discussion groups (~100 million words). We were careful to remove formatting symbols that would not appear in printed versions of the corpora.

To examine the consistency of letter frequency in general among the corpora, we computed Pearson product moment correlation coefficients between case-insensitive single-letter counts made on each of the corpora. Mean intercorpus correlation of case-insensitive letter counts was .9963. Case-sensitive letter counts were next calculated for each of the corpora. The mean correlation between uppercase frequency counts from each corpus was .9065; for lowercase counts, the mean intercorpus correlation was .9965. Hence, the corpora are in high agreement both on the frequency of letters in general and on the frequency of letters appearing in each case.

The correlation for the uppercase counts was lower than that for the lowercase counts because the newsgroup corpus agreed with the others less on uppercase frequency. This is probably because informal writing, such as that found in discussion group text, often exhibits improper capitalization in comparison with text that has been thoroughly edited, such as that of the NYT. With the newsgroup corpus ignored, the mean intercorrelation of uppercase counts from the remaining corpora went up to .9268, with highest consistency among the edited corpora (NYT, Brown, and encyclopedia).

In addition to the single-letter frequencies, we calculated bigram frequencies and checked them for consistency across the corpora. For the calculation, we defined a bigram as *directly adjacent characters within a word*. To examine consistency of case-insensitive bigram frequency among the corpora, we computed Pearson product–moment correlation coefficients between case-insensitive bigram counts made on each of the corpora. Mean intercorpus correlation of case-insensitive bigram counts was .9858, indicating that all the corpora were in high agreement on bigram frequency in general. Case-sensitive bigram counts were computed next. There are four categories for case-sensitive bigrams: lowercase paired with lowercase (aa), lowercase followed by uppercase (aA), uppercase followed by lowercase (Aa), and uppercase followed by uppercase (AA). The correlations for case-sensitive bigrams were .9844 for aa, .7042 aA, .8948 for Aa, and .7602 for AA.

As before, case-sensitive bigram frequencies were more consistent among the edited text corpora than among the informal corpora. The lower correlations for aA and AA categories primarily reflect improper capitalization in the unedited corpora.

Because of the greater consistency among the edited corpora, we judged the NYT corpus to be the best and most representative source, and data in the tables presented here were derived from it. Case-sensitive counts from the other corpora are available at the Psychonomic Society's Web archive.

Single-Letter Frequency

To test the case-equivalence assumption, we correlated upper- and lowercase counts. The correlation was a startlingly low .6337, quite different from Tinker's (1928) rank–order correlation of .93. Contrary to Tinker's data, our counts indicate that upper- and lowercase letters do not have equivalent frequencies in print. Case-sensitive single-letter frequency counts from the NYT corpus are provided in Table 1. Experimenters interested in case-sensitive single-letter frequency can use these tables to help construct frequency-balanced stimulus sets.

Bigram Frequency

Table 2 presents the correlation matrix of case-sensitive bigram counts from the NYT corpus. There is moderate agreement between uppercase/uppercase and lowercase/lowercase bigram frequencies, but poor consistency between other case-sensitive bigram counts. Hence, bigrams do not have the same relative frequency in all case combinations. The Appendix contains predecessor-by-successor case-sensitive bigram counts calculated from the NYT cor-

Table 1
Raw Case-Sensitive Single-Letter Counts from the NYT Corpus

Letter	Uppercase *f*	Lowercase *f*	Uppercase Rank	Lowercase Rank
A	280,937	5,263,779	3	3
B	169,474	866,156	8	20
C	229,363	1,960,412	5	12
D	129,632	2,369,820	12	11
E	138,443	7,741,842	11	1
F	100,751	1,296,925	17	15
G	93,212	1,206,747	19	17
H	123,632	2,955,858	13	9
I	223,312	4,527,332	6	6
J	78,706	65,856	20	25
K	46,580	460,788	22	22
L	106,984	2,553,152	15	10
M	259,474	1,467,376	4	14
N	205,409	4,535,545	7	5
O	105,700	4,729,266	16	4
P	144,239	1,255,579	10	16
Q	11,659	54,221	24	26
R	146,448	4,137,949	9	8
S	304,971	4,186,210	2	7
T	325,462	5,507,692	1	2
U	57,488	1,613,323	21	13
V	31,053	653,370	23	21
W	107,195	1,015,656	14	19
X	7,578	123,577	25	23
Y	94,297	1,062,040	18	18
Z	5,610	66,423	26	24

pus. Experimenters interested in controlling case-sensitive bigram frequency can use these tables to help construct frequency-balanced stimuli.

Frequency of Other Alphanumeric Characters

Numerals and other nonalphabetic characters (e.g., &, $,%, #, @) are often used as distractors in attention experiments—as noise stimuli, for example, when one is studying the interference of flanking characters on the identification of a target letter (e.g., Eriksen & Hoffman, 1972; Estes, 1972; Krueger, 1970, 1973a). Because flanker interference depends on the relationship of the flanking characters and the targets (i.e., letters produce more interference than do nonalphabetic characters, and high-frequency bigrams produce more interference than do low-frequency bigrams), it is important to know the naturally occurring frequency characteristics of distractor characters, as well as the bigram frequency of the target letter and noise characters.

Table 2
Correlation Matrix of Case-Sensitive Bigram Combinations Computed from the NYT Corpus

	aa	aA	Aa	AA
aa	–	.012	.587*	.665*
aA		–	−.005	.008
Aa			–	.372*
AA				–

Note—Number of observations for each bigram type: aa, 48,301,606; aA, 10,007; Aa, 2,446,369; and AA, 712,561. *$p < .001$.

We computed single-unit and bigram frequency counts for 32 nonalphabetic characters including the 10 digits (ASCII 33–64; ! to @). Characters commonly used as distractors varied quite widely in their frequencies and were more commonly found as successors to letters (on the right) than as predecessors. Table 3 shows the raw frequency[2] with which each distractor character flanked a letter as predecessor and successor in the NYT corpus.

APPLYING CASE-SENSITIVE FREQUENCY TO LETTER IDENTIFICATION

We have demonstrated that upper- and lowercase letters do not have equivalent relative frequencies in print. As mentioned previously, experimenters often use uppercase letters as stimuli but compute frequency using counts dominated by lowercase letters. We next tested the ability of the case-sensitive frequency counts to predict naming time when subjects are identifying isolated uppercase letters. Our goal was to determine whether uppercase counts provide a better prediction of reaction time than lowercase counts when uppercase letters are used as stimuli.

In this experiment, the subject named a letter aloud as quickly as possible. Response latency was measured as the time between stimulus onset and initiation of the vocal response. The naming task includes both a decision and motor production component. To factor out the motor component, we subtracted motor production time from overall naming time on a per-subject (and per-letter) basis. To estimate the motor component, we asked subjects to name a letter on a cue presented well after the decision process had been completed. Naming on cue contains the same motor process as does the standard naming task but with the decision process complete. By comparing the name-on-cue task and the standard naming task on a per-subject basis, we were able to subtract the motor component from the decision component for each letter.

Method

Subjects. Eight graduate students from Queen's University participated in the experiment. They were each paid $10, and all had normal or corrected-to-normal vision.

Apparatus. Stimuli were presented on a 19-in. CRT monitor at a viewing distance of 75 cm; we used a chin bar to keep viewing distance constant. Subjects wore a Sony Dynamic Headset (HS-65A-1) equipped with a voice–key microphone. Microphone threshold was calibrated to each subject during the practice phase of the experiment. At a viewing distance of 75 cm, characters subtended a maximum visual angle of about 0.38º vertically and 0.29º horizontally.

Procedure. All subjects were given 5 min of dark adaptation before beginning the experiment. All subjects completed both the cued-production task and the standard letter-naming task so that an estimate of production time for a letter could be subtracted from overall naming time for the same letter. Each trial (for either task) began with a fixation cross at the center of the display. The subject was instructed to focus on the cross and, when ready to proceed, to press the start button with the index finger of the dominant hand.

During the production (cued-naming) task, the subjects were informed that the letters would be presented in alphabetical order. On

Table 3
Single-Unit Frequency of Various Nonalphabetic Characters (ASCII 33–64), and Bigram Frequency as Predecessor (#A) or Successor (A#) to an Alphabetic Character

Character	Single Unit	As Predecessor (#A)	As Successor (A#)
!	2,178	58	1,866
"	284,671	142,168	26,827
#	10	0	0
$	51,572	427	61
%	1,993	13	9
&	6,523	438	350
'	204,497	187,914	185,857
(	53,398	43,473	55
)	53,735	11	37,506
*	20,716	882	530
+	309	8	112
,	984,969	111	810,376
-	252,302	160,049	138,556
.	946,136	41,636	847,611
/	8,161	3,948	4,207
0	546,233	2,006	38
1	460,946	959	5,792
2	333,499	1,065	2,435
3	187,606	1,335	1,945
4	192,528	880	1,820
5	374,413	999	1,514
6	153,865	1,576	1,491
7	120,094	840	1,074
8	182,627	828	1,021
9	282,364	1,697	481
:	54,036	13	48,354
;	36,727	58	28,301
<	82	74	18
=	22	1	1
>	83	52	70
?	12,357	10	11,938
@	1	1	1

each trial, after subjects fixated the cross and pressed the start button, a letter immediately replaced the cross and remained visible for a random delay period of 1–3.5 sec. Following the delay, the letter was removed, and the screen flashed to cue the subject to respond. The subjects' task was to name the letter aloud as quickly as possible following this cue. Latency was defined as the *elapsed time between cue and verbal response*. The screen remained blank for 1–3 sec before the next trial.

During the naming task, letters were presented in random order. On each trial, after subjects fixated on the cross and pressed the ready button, a letter immediately replaced the cross and remained visible until a response was initiated. The subjects' task was to name the letter aloud as quickly as they were able to identify it. Latency was defined as the elapsed time between stimulus onset and response initiation. Following the response, the screen went blank while the experimenter recorded the subjects' accuracy. The screen remained blank for a random interval of 1–3 sec between trials, and the fixation cross then reappeared to signify a new trial.

To balance practice effects, an *ABBABAAB* design was used. For each subject, the two tasks were randomly assigned to *A* or *B* . For example, if *A* was assigned to the naming task and *B* to the production task, each *A* contained 3 blocks of 26 letter-naming trials in random order (each letter equally represented within a block), and each *B* contained 3 blocks of 26 cued-production trials. Thus, in total the data presented here represent 312 letter-naming and 312 cued-production trials per subject.

Results

Decision latency was calculated by subtracting mean production latency for a letter (on a per-subject basis) from each of the 12 naming trials for the same letter.The Pearson correlation between uppercase frequency and decision latency to uppercase letters was $r(25) = -.602$, $p < .01$. By contrast, the correlation between lowercase frequency and decision latency to uppercase letters was $r(25) = -.328$, n.s.

The pattern supports our presumption that uppercase decision latency is predicted better by upper- than by lowercase counts.[3] We tested the difference between these correlations with Williams' (1959) ratio for nonindependent correlations (see also Steiger, 1980). Uppercase decision latency was predicted significantly better by the uppercase frequency counts than by the lowercase ones [$t(23) = -1.92, p < .05$].

The advantage for same-case prediction illustrates the importance of using case-sensitive frequency counts in single-letter identification. In the Discussion section, we consider some examples of case-sensitive bigram confounds in a classic psychological experiment.

DISCUSSION

We have provided case-sensitive frequency counts for single letters and bigrams from large samples of carefully selected corpora. In addition, we have demonstrated that identification time for uppercase letters is predicted better by frequency of uppercase letters than by frequency of lowercase letters. The effect of case-sensitive frequency is, however, by no means limited to single-letter identification. A salient illustration of case-sensitivity from experience is the *proper name effect*. In a series of lexical decision experiments, Peressotti, Cubelli, and Job (2003) found a consistent reaction time advantage for proper names that had the first letter capitalized, compared with common nouns and proper names with the first letter in lowercase. Since proper names are experienced with the first letter capitalized, the proper name effect provides further support for the importance of accounting for case-sensitive frequency in experimental stimuli.

Our main point is that unless case is taken into account, letter and bigram frequency may be an unexpected determinant of performance. Consider the classic Posner matching task (e.g., Posner & Mitchell, 1967) still used in many experiments. Posner and Mitchell simultaneously presented two letters and asked subjects to respond "same" or "different." Under physical matching instructions, subjects were required to respond "same" only if the two letters were *physically* identical (AA, aa) and "different" if not (AE, Aa). Under name matching instructions, subjects were asked to respond "same" if the two letters had the *same name* (AA, Aa) and "different" if not (AE, Ah).

A general finding is that *same* responses are faster than *different* responses in physical matching[4] (the *fast-same* effect). This effect is troubling to theory because

one would expect a *different* response to be made as soon as one featural mismatch has been detected between the letters, whereas a *same* response requires an exhaustive search of all features to verify that the two stimuli are identical.

The fast-same effect has been explained as an artifact of more possible *different* combinations than *same* combinations within an experiment (Nickerson, 1973). Alternatively, the advantage for *same* responses has been attributed to internal noise in the perceptual system (e.g., Krueger, 1978), or to an encoding bias for the second letter primed by the first (e.g., Proctor, 1981). Before accepting any possibility, however, we note that the fast-same effect is confounded with case-sensitive bigram frequency. Considering the pairs that Posner and Mitchell (1967) used as stimuli, the mean bigram log frequency[5] of physically *same* pairs calculated from the NYT corpus is 8.03, whereas the corresponding value for physically *different* pairs is only 7.35. Hence, the *fast-same* effect may be contaminated by a frequency difference between physically *same* and *different* letter pairs. We know that letter and bigram frequency affects processing; thus, the difference in frequency for *same* versus *different* pairs may partially explain the advantage of faster *same* responses in physical matching.

A common finding under name-matching instructions is that *same* responses are faster to pairs that are physically the same (AA, aa) than to those that are physically different (Aa). This finding was used to argue for a temporal hierarchy of processing stages, the first involving physical stimulus examination, and the second requiring translation from a physical code to a name code and memory access (Posner, 1978; Posner & Mitchell, 1967). A *same* response to the stimulus pair Aa presumably takes longer because it requires translation from physical to name code, whereas a response to AA is faster because it can be accomplished by raw stimulus examination without the need for identification.

However, the temporal hierarchy explanation is also confounded with case-sensitive bigram frequency. Mean bigram log frequency of physically *same* pairs (AA) computed from the NYT corpus is 8.03, whereas for physically *different* same-name pairs (Aa) it is 3.35. Faster *same* responses to physically *same* over physically *different* responses in name matching may well be influenced by a frequency difference.

We selected the classic Posner matching task as an example because there is such a wealth of letter-matching data that we are comfortable that case-sensitive bigram frequency can be ruled out as an explanation of any of the effects. For example, the bigram confound cannot account for more recent data (e.g., Proctor, 1981). The classic Posner and Mitchell (1967) experiment is an illustration of how, if not properly controlled when designing stimuli, case-sensitive frequency could confound interpretation of experimental results.

We described *initial-letter counts* as a quick and easy method to estimate uppercase letter frequency (i.e., counting the occurrence of initial letters in words from a dictionary or word-frequency norm—e.g., Latimer, 1972; Tinker, 1928, Table 5). However, we caution against the use of such a technique, which is based on the logic that the more frequently a letter occurs at the beginning of a word, the more often it will be capitalized. It further assumes that all words are equally likely to begin a sentence—an assumption that is false. We counted the number of pages dedicated to each letter from several dictionaries and used this count as a predictor of uppercase frequency from the NYT corpus, and the mean correlation was .6825.

Obtaining the Case-Sensitive Frequency Counts

Raw case-sensitive frequency counts of single characters and all possible bigrams (including the nonalphabetic characters) for the corpora listed here can be found at the Psychonomic Society's Web archive. We encourage experimenters to consult these tables to balance frequency when constructing stimuli composed of letters.

Learning is a product of repetition, and too often frequency goes ignored in the design of experiments. Whether consciously or not, humans take advantage of frequency information every day, and an experiment is certainly not immune to a frequency bias. We advise experimenters to use frequency counts that are sensitive to letter case when they construct experiments using letter stimuli.

REFERENCES

Appelman, I. B., & Mayzner, M. S. (1981). The letter-frequency effect and the generality of familiarity effects on perception. *Perception & Psychophysics*, **30**, 436-446.

Baddeley, A. D., Conrad, R., & Thompson, W. E. (1960). Letter structure in the English language. *Nature*, **186**, 414-416.

Bouma, H. (1971). Visual recognition of isolated lowercase letters. *Vision Research*, **11**, 459-474.

Broadbent, D. E. (1967). Word-frequency effect and response bias. *Psychological Review*, **74**, 1-15.

Broadbent, D. E., & Gregory, M. (1968). Visual perception of words differing in letter digram frequency. *Journal of Verbal Learning & Verbal Behavior*, **7**, 569-571.

Bryden, M. P. (1968). Symmetry of letters as a factor in tachistoscopic recognition. *American Journal of Psychology*, **81**, 513-524.

Cosky, M. J. (1976). The role of letter recognition in word recognition. *Memory & Cognition*, **4**, 207-214.

Dominowski, R. L., & Duncan, C. P. (1964). Anagram solving as a function of bigram frequency. *Journal of Verbal Learning & Verbal Behavior*, **3**, 321-325.

Egeth, H., & Blecker, D. (1971). Differential effects of familiarity on judgments of sameness and difference. *Perception & Psychophysics*, **9**(4), 321-326.

Eriksen, C. W., & Hoffman, J. E. (1972). Some characteristics of selective attention in visual perception determined by vocal reaction time. *Perception & Psychophysics*, **11**, 169-171.

Estes, W. K. (1972). Interactions of signal and background variables in visual processing. *Perception & Psychophysics*, **12**, 278-286.

Gaines, H. F. (1939). *Cryptanalysis: A study of ciphers and their solutions.* New York: Dover.

Geyer, L. H. (1977). Recognition and confusion of the lowercase alphabet. *Perception & Psychophysics*, **22**, 487-490.

Gilmore, G. C., Hersh, H., Caramazza, A., & Griffin, J. (1979). Multidimensional letter similarity derived from recognition errors. *Perception & Psychophysics*, **25**, 425-431.

Howes, D. H., & Solomon, R. L. (1951). Visual duration thresholds as a function of word-probability. *Journal of Experimental Psychology*, **41**, 401-410.

Jones, M. N. (2001). *T-REX: A template-resonance excitation model of single letter classification.* Master's thesis, Queen's University at Kingston.

Krueger, L. E. (1970). Effect of bracketing lines on speed of "same"–"different" judgments of two adjacent letters. *Journal of Experimental Psychology*, **84**, 324-330.

Krueger, L. E. (1973a). Effect of irrelevant surrounding material on speed of "same–different" judgments of two adjacent letters. *Journal of Experimental Psychology*, **98**, 252-259.

Krueger, L. E. (1973b). Effect of letter-pair frequency and orientation of speed of "same"–"different" judgments by children and adults. *Bulletin of the Psychonomic Society*, **2**, 431-433.

Krueger, L. E. (1973c). Effect of stimulus frequency on speed of "same"–"different" judgments. In S. Kornblum (Ed.), *Attention and performance IV* (pp. 497-506). New York: Academic Press.

Krueger, L. E. (1975). Familiarity effects in visual information processing. *Psychological Bulletin*, **82**, 949-974.

Krueger, L. E. (1978). A theory of perceptual matching. *Psychological Review*, **85**, 278-304.

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

Latimer, C. R. (1972). Search time as a function of context letter frequency. *Perception*, **1**, 57-71.

Mayzner, M. S., & Tresselt, M. E. (1965). Tables of single-letter and digram frequency counts for various word-length and letter-position combinations. *Psychonomic Monograph Supplements*, **1** (Whole No. 2), 13-32.

Mewhort, D. J. K., & Campbell, A. J. (1981). Toward a model of skilled reading: An analysis of performance in tachistoscopic tasks. In G. E. MacKinnon & T. G. Waller (Eds.), *Reading research: Advances in theory and practice* (Vol. 3, pp. 39-118). New York: Academic Press.

Miller, G. A., Bruner, J. S., & Postman, L. (1954). Familiarity of letter sequences and tachistoscopic identification. *Journal of General Psychology*, **50**, 129-139.

Nickerson, R. S. (1965). Response times for same–different judgments. *Perceptual & Motor Skills*, **20**, 15-18.

Nickerson, R. S. (1973). Frequency, recency, and repetition effects on same and different response times. *Journal of Experimental Psychology*, **101**, 330-336.

Peressotti, F., Cubelli, R., & Job, R. (2003). On recognizing proper names: The orthographic cue hypothesis. *Cognitive Psychology*, **47**, 87-116.

Posner, M. I. (1978). *Chronometric explorations of mind.* Hillsdale, NJ: Erlbaum.

Posner, M. I., & Mitchell, R. F. (1967). Chronometric analysis of classification. *Psychological Review*, **74**, 392-409.

Proctor, R. W. (1981). A unified theory for matching-task phenomena. *Psychological Review*, **88**, 291-326.

Rice, G. A., & Robinson, D. O. (1975). The role of bigram frequency in the perception of words and nonwords. *Memory & Cognition*, **3**, 513-518.

Solso, R. L., & King, J. F. (1976). Frequency and versatility of letters in the English language. *Behavior Research Methods & Instrumentation*, **8**, 283-286.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, **87**, 245-251.

Tinker, M. A. (1928). The relative legibility of the letters, the digits, and of certain mathematical signs. *Journal of General Psychology*, **1**, 472-495.

Townsend, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics*, **9**, 40-50.

van der Heijden, A. H. C., Malhas, M. S. M., & van den Roovaart, B. P. (1984). An empirical interletter confusion matrix for continuous-line capitals. *Perception & Psychophysics*, **35**, 85-88.

Williams, E. J. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society*: *Series B*, **21**, 396-399.

NOTES

1. An automated URL spider was sent out to collect text from 100,000 Web pages with domain extensions .edu, .org, or .ca. The spider was restricted to collecting text a maximum of three directories deep within a URL.

2. Mean intercorpus correlation was $r(21) = .9373$ for the symbols and $r(9) = .9482$ for the digits; hence, frequency of occurrence was consistent across the corpora.

3. Interestingly, logarithmic transformations on frequency did not predict latency as well as raw frequency (the opposite is usually the case).

4. Posner and Mitchell found a *fast-same* effect in their original study, but it was not reliable. The effect had been reported earlier (Nickerson, 1965) and has been demonstrated many times since then.

5. The bigram frequency distributions were positively skewed, thus, the log-transformed data are described here.

ARCHIVED MATERIALS

The following materials associated with this article may be accessed through the Psychonomic Society's Norms, Stimuli, and Data archive, http://www.psychonomic.org/archive/.

To access these files or links, search the archive for this article using the journal (*Behavior Research Methods, Instruments, & Computers*), the first author's name (Jones), and the publication year (2004).

File: Jones_BRMIC_2004.zip

Description: The 590K compressed archive file contains the same two files (single letter and bigram counts) in four formats (tab-delimited text, space-delimited ASCII, Microsoft Excel 2000, and Adobe PDF).

Jones2004_Single.* contains single-character counts (ASCII 33–126; ! to ~) separately for each of the five corpora. Each row represents one character; each column, one of the corpora (NYT, encyclopedia, Brown, Web page, newsgroup).

Jones2004_Bigram.* contains bigram counts for each combination of characters (ASCII 33–126; ! to ~) separately for each of the five corpora. Each row represents a predecessor-by-successor bigram; each column, one of the corpora (NYT, Encyclopedia, Brown, Web page, newsgroup).

Jones2004_README.txt contains a full description of the content of Jones_BRMIC_2004.zip, including extended definitions of the variables (a 3K plain text file).

Author's e-mail Address: mike@psyc.queensu.ca.

APPENDIX
Predecessor (A_) by Successor (_A) Case-Sensitive Bigram Counts From the NYT Corpus

These tables contain log-transformed predecessor-by-successor case-sensitive bigram frequency counts from the NYT corpus. The integer frequency count of a given bigram (f_i) was transformed to its table entry (x_i) by

$$x_i = \ln(f_i),$$

and an approximation may be unpacked by

$$f_i \approx \text{round}[\exp(x_i)].$$

Note that due to the logarithmic transformation, cells with ".00" had so few observations that the value was rounded to zero (an exponential transformation will return an integer value of 1), whereas cells with missing values had observed frequencies of zero (since the logarithm of zero is undefined).

Table A1
Uppercase-by-Uppercase Bigram Frequency

	Successor																									
Predecessor	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	5.81	7.50	8.08	7.52	8.21	6.69	7.40	7.90	8.16	6.68	6.24	8.53	8.17	8.34	6.05	8.79	1.39	8.57	7.92	8.49	6.07	6.60	7.97	5.21	5.77	6.13
B	7.40	6.12	7.12	4.33	5.93	3.09	3.53	4.60	6.08	2.83	2.48	6.05	6.11	6.73	6.42	4.96	1.79	6.02	5.54	5.65	6.63	1.10	3.58	.00	2.83	3.14
C	7.70	7.34	6.01	5.51	7.76	5.53	3.53	3.22	8.46	1.79	3.22	5.04	5.88	7.71	6.88	5.91	.	6.71	7.33	7.46	6.75	4.49	4.03	3.61	6.32	.00
D	7.48	4.95	6.25	6.12	8.02	4.49	3.00	4.28	8.11	.00	3.04	7.02	4.44	8.94	6.69	2.71	.00	7.34	5.16	3.53	6.59	1.61	3.18	2.08	3.81	1.95
E	6.06	7.52	7.95	8.42	7.26	7.06	7.98	8.47	7.85	6.40	7.38	8.59	8.37	9.15	5.58	7.66	3.14	9.39	8.41	9.01	6.75	7.92	7.07	5.06	5.97	5.73
F	6.27	1.10	5.2	4.09	6.60	6.49	3.47	3.50	6.52	3.22	3.69	5.37	3.97	5.96	7.70	2.48	.00	5.52	4.56	4.64	4.83	4.34	3.04	.69	2.20	.00
G	7.16	2.71	3.64	6.26	6.84	3.33	4.55	1.95	7.21	.	1.95	5.18	5.23	8.61	6.36	5.51	.	6.92	4.38	3.66	6.47	2.08	3.37	.	2.40	1.79
H	5.56	3.14	8.14	3.74	4.96	2.48	6.68	3.40	3.18	.69	4.65	3.89	3.04	5.34	6.17	6.87	.	4.65	8.00	8.63	3.30	3.33	6.34	2.77	1.39	3.22
I	8.18	7.03	7.99	7.81	6.50	7.13	6.84	8.28	7.50	4.90	6.75	8.32	7.94	8.11	6.01	6.82	2.56	8.48	7.93	8.96	6.21	7.84	7.02	8.39	4.47	4.79
J	3.99	3.66	.69	2.56	3.40	.00	.69	.	4.67	.00	.	1.39	2.08	5.69	3.26	.00	.	3.99	.69	3.09	3.00	.	.	.69	.	.
K	6.33	1.79	7.24	3.71	5.29	3.47	2.89	3.69	5.68	.	3.30	5.31	1.61	6.15	6.65	2.30	.00	7.70	6.00	3.61	5.38	3.33	3.66	.	2.64	2.20
L	8.81	6.88	6.50	5.85	8.34	5.93	5.61	5.36	8.07	1.10	5.83	8.01	3.26	4.39	8.01	6.97	.00	6.67	6.77	6.35	6.72	3.00	4.28	1.79	5.16	3.50
M	8.01	5.18	3.81	5.02	7.28	4.89	5.21	4.73	7.33	2.94	3.99	5.54	6.52	5.98	8.78	4.57	1.39	7.07	5.84	5.49	7.05	2.40	3.93	1.39	5.46	2.56
N	9.74	3.50	5.34	5.49	8.83	2.20	6.15	6.16	9.39	.	4.76	3.61	6.20	7.14	9.47	4.49	.69	7.36	5.18	5.56	8.08	1.10	6.01	.00	5.74	.10
O	3.83	7.50	9.10	7.23	6.62	7.60	7.30	7.71	8.62	6.63	5.75	7.75	7.71	7.40	7.12	8.69	.	8.29	7.61	8.55	4.16	5.78	6.66	3.58	7.58	4.22
P	7.98	3.95	4.25	3.30	8.61	2.56	3.00	3.37	6.25	1.39	3.50	5.93	8.56	3.50	6.97	5.80	.00	5.66	6.84	3.85	5.95	2.20	4.26	4.66	3.66	1.95
Q	4.96	.	3.64	1.95	3.43	.	2.77	2.30	3.74	.	.69	.00	.	3.40	1.79	.00	.	2.40	3.71	.	1.79	.00	2.20	.	2.48	.69
R	8.76	7.12	6.91	6.61	9.20	7.08	7.07	5.69	7.38	5.20	5.13	3.74	4.86	3.83	9.39	.51	.00	6.59	6.42	7.94	7.69	2.08	4.50	2.08	5.37	1.10
S	8.50	7.30	6.55	7.98	9.11	4.62	5.83	4.62	9.08	1.61	6.36	7.05	6.37	8.27	7.62	6.15	2.40	7.82	7.79	8.79	8.50	3.69	5.68	1.39	7.40	1.95
T	9.09	3.81	7.39	4.36	7.91	5.65	7.40	6.20	8.71	.00	3.89	6.95	5.30	8.46	7.95	5.82	4.09	8.78	8.92	7.30	7.37	2.30	4.25	4.56	10.2	3.93
U	6.95	7.26	6.69	6.50	5.56	5.56	6.30	6.18	4.63	5.68	4.37	6.88	6.49	5.24	7.88	6.53	6.22	7.13	7.04	6.67	1.10	2.08	2.71	3.50	5.33	3.85
V	6.92	.69	1.61	4.51	6.85	.00	1.39	2.83	6.91	3.40	2.20	5.47	1.79	5.68	7.17	1.10	3.56	5.62	4.44	7.32	3.00	.69	2.30	5.20	2.20	1.39
W	6.21	4.09	1.79	4.88	7.87	1.39	2.64	3.69	2.30	2.71	2.71	4.58	4.19	5.83	7.22	2.83	3.30	4.89	5.28	5.02	3.40	1.39	3.81	1.39	3.89	1.39
X	5.76	2.94	3.22	3.14	8.57	.00	1.10	.	5.11	2.94	.69	2.40	2.64	4.62	4.60	1.61	2.20	1.95	4.75	3.83	3.71	2.71	1.79	5.06	1.39	.
Y	7.13	5.49	5.27	5.18	6.79	2.71	5.11	5.10	3.26	.69	5.78	6.68	5.58	10.4	5.63	3.53	.69	7.14	5.56	7.60	3.89	3.04	3.87	1.95	1.10	2.56
Z	5.70	1.79	3.58	2.77	4.71	.00	.69	.	5.96	.	.	2.20	.69	4.58	4.32	1.95	1.95	3.53	1.79	4.83	3.89	.	.00	.	2.56	4.11

Table A2
Uppercase-by-Lowercase Bigram Frequency

	Successor																								
Predecessor	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
A	5.55	9.51	8.44	9.08	6.20	8.99	8.06	5.74	8.03	3.40	5.99	9.83	10.2	10.2	3.22	8.52	5.44	9.60	9.67	9.25	8.67	8.27	6.67	4.49	5.71	6.61
B	9.95	4.54	4.69	4.91	9.87	4.69	4.55	4.73	9.04	4.39	3.69	8.60	.	3.66	9.79	.	.00	10.1	.00	1.61	10.8	.	.00	1.10	8.12	1.10
C	10.4	.	.00	.	8.88	.	.	10.2	9.30	.	.	9.43	.	.00	11.3	6.79	.	8.94	3.89	3.22	8.71	.	2.56	.00	6.41	6.53
D	9.89	.	1.39	.	10.5	.00	.	3.99	9.43	3.74	.	2.71	4.19	2.40	9.33	.	.	9.39	3.00	.	8.59	4.39	5.76	.	6.12	3.97
E	9.01	4.78	7.19	8.76	2.56	5.65	6.57	4.77	7.13	2.08	4.33	8.69	7.68	8.94	1.79	6.09	6.17	7.51	7.47	6.40	8.51	8.58	6.28	8.53	5.44	4.16
F	8.89	.	.	1.10	9.88	1.10	.00	2.40	9.29	4.30	.	8.80	.	.00	9.81	.	.	9.89	.00	3.00	8.01	.	.	.	2.71	.
G	9.28	.00	.00	3.26	9.72	1.10	.	6.01	8.40	2.08	.	7.66	1.95	2.48	9.94	4.29	.	9.53	1.39	.	8.47	.	4.58	.	4.88	.
H	10.1	.	.	.	10.5	.	.	.	9.46	1.39	.00	4.56	2.64	.	10.0	.	.	4.13	3.18	.00	8.64	1.79	3.18	.	6.38	1.95
I	6.18	5.12	6.05	6.34	2.71	8.90	5.34	1.10	2.08	.69	3.95	7.71	7.34	10.9	6.76	3.30	3.00	8.62	9.43	10.1	1.95	6.38	3.18	1.39	.69	4.52
J	10.2	.	.	.	9.33	.	.	3.58	7.57	.	.	.	.	.	9.76	.00	.	7.67	.00	.	9.25	.	.00	.	1.95	.
K	8.84	2.94	3.37	3.71	9.10	4.28	4.42	6.70	8.66	3.33	.	7.49	2.30	7.80	8.41	.	.	7.74	1.10	1.10	7.74	2.56	4.36	.	5.90	.69
L	9.82	.	.	.	9.71	.	.	1.10	9.52	3.22	.	6.02	.	.69	9.64	.00	.	.00	2.64	6.35	8.10	3.00	.00	.	7.36	.
M	11.0	2.20	8.68	6.05	9.76	3.00	3.26	2.20	10.1	.	3.83	1.95	3.37	2.64	10.1	.00	.	11.2	9.03	3.22	8.91	.	1.95	.	7.92	.00
N	9.60	4.41	5.06	5.30	10.9	5.11	5.09	3.97	8.57	1.61	.00	.00	1.39	.	10.0	.	.	2.20	.	2.48	7.04	1.61	2.20	.	5.68	.00
O	6.25	6.52	7.85	5.83	3.83	8.34	5.26	7.04	6.05	3.09	6.45	8.54	5.63	9.34	3.83	7.94	1.10	8.57	7.09	7.87	7.67	7.14	6.30	5.61	4.98	5.16
P	10.2	.	.00	.69	9.68	5.23	2.56	9.47	8.61	.	.69	8.54	.	1.79	9.34	1.95	.	10.2	5.52	3.64	8.28	1.10	4.39	.69	5.29	.
Q	5.25	4.14	4.68	4.82	5.00	4.82	4.57	4.60	4.30	.	.	.	.	.	.00	.	.	.	.	8.27	8.47	.	2.40	.	.	.
R	9.19	3.95	4.16	4.93	10.3	4.73	3.99	6.70	9.26	.	.	1.10	.00	.	9.97	.	.	.00	1.10	1.79	9.13	.	1.39	1.79	6.74	2.56
S	10.0	2.71	9.36	.	10.8	1.39	5.33	10.1	9.21	.69	7.05	7.47	7.96	6.40	10.1	8.98	7.06	5.06	.69	10.8	9.61	5.19	7.85	.00	8.11	4.74
T	8.94	4.23	4.81	.	9.32	.	.	12.0	9.57	.00	2.48	1.79	.	.	9.71	.	.	9.24	7.98	.	8.75	1.95	7.88	1.61	7.54	2.64
U	1.39	3.87	3.71	4.17	3.33	2.83	4.13	4.04	1.39	2.20	7.04	6.09	4.61	10.2	1.10	7.21	.	6.53	6.70	6.44	.69	.	3.87	.00	.00	4.99
V	8.57	.	.	.	8.21	.	.	.	9.16	.69	.	5.70	.	.69	7.61	.	.	4.62	1.95	4.89	4.79	.	.	.	4.14	.
W	9.86	.	.	.	10.2	.	.	9.92	9.70	.	3.30	2.71	.69	.	9.09	.00	.	6.80	2.77	5.12	5.23	.	.00	.	6.38	.
X	4.93	.	.	.	4.83	.	.	1.39	4.68	.	.	.	.	.	2.40	.00	.	.	.69	.69	2.20	.	.	.69	2.30	.
Y	8.09	1.61	.	.	8.59	.	.	.	6.83	.	.	2.40	.	1.95	10.4	3.40	.	7.75	3.00	1.10	7.25	5.09	.	.	.	3.33
Z	6.95	2.48	.00	3.33	6.86	.	.00	5.56	6.75	.00	.	.69	2.40	.	5.92	.	.	.69	2.71	.	5.80	4.38	4.22	.	3.43	.

Table A3
Lowercase-by-Lowercase Bigram Frequency

	Successor																								
Predecessor	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	7.38	11.5	12.2	12.1	9.25	10.6	11.6	9.03	12.4	9.03	11.0	13.1	11.9	13.8	7.79	11.4	7.95	13.3	13.0	13.5	11.1	11.6	10.5	9.64	12.0	9.23
b	11.4	9.00	6.57	6.50	12.4	3.85	3.56	5.33	10.9	8.14	3.56	11.5	6.93	5.61	11.5	5.25	.	10.6	10.2	8.52	11.3	7.05	6.57	.	11.4	2.71
c	12.4	4.08	10.4	5.76	12.7	3.00	.69	12.5	11.7	1.10	11.4	11.1	4.89	4.90	12.8	3.76	7.87	11.2	9.47	12.1	11.2	1.95	3.00	.69	9.93	5.82
d	11.6	7.33	7.73	10.1	12.8	7.46	9.83	7.55	12.2	7.54	5.46	9.76	9.44	9.16	11.4	6.09	6.83	10.7	11.0	7.30	11.0	9.31	8.73	.00	9.98	5.09
e	12.8	10.2	12.4	13.3	12.2	11.2	11.0	9.36	11.3	7.58	9.97	12.4	12.0	13.4	10.6	11.4	9.52	13.9	13.4	12.3	9.23	11.6	11.5	11.3	11.4	8.43
f	11.1	4.32	4.78	3.50	11.5	11.3	6.02	3.64	11.9	4.23	5.42	10.1	6.27	5.31	12.4	3.26	.	11.5	8.16	10.8	10.5	3.81	4.91	1.39	8.18	2.56
g	11.4	5.36	3.58	6.81	12.2	6.20	9.57	11.7	11.0	3.09	4.68	9.99	7.66	10.4	11.0	5.28	1.95	11.3	10.3	9.35	10.6	2.94	5.96	2.08	9.12	3.64
h	13.1	8.20	6.35	7.84	14.1	6.60	5.24	6.21	12.6	3.50	5.91	9.00	8.58	9.71	12.4	5.91	5.12	10.6	9.32	11.2	10.2	5.18	8.07	.	9.54	2.64
i	11.9	10.4	12.8	12.4	12.1	11.1	11.8	6.54	7.10	7.21	10.4	12.5	11.8	14.0	12.7	10.5	8.09	11.9	13.1	13.1	8.71	11.7	6.71	9.25	6.69	10.2
j	8.52	1.61	3.43	3.56	9.39	2.08	.	1.10	7.24	2.30	3.76	2.40	3.93	3.37	10.0	1.61	.	1.39	2.30	2.56	10.1	.69	.	.	1.39	1.79
k	9.21	6.78	4.55	6.64	11.9	6.73	6.35	7.68	11.0	3.85	6.42	9.22	6.53	9.80	8.42	6.30	.	8.04	10.5	6.95	7.46	5.42	7.05	.00	8.70	2.56
l	12.5	8.60	8.44	11.9	12.9	10.0	7.95	7.18	12.7	4.22	9.54	12.7	9.68	7.78	12.0	9.50	3.43	8.53	11.3	10.9	10.9	9.58	8.49	3.40	12.1	6.25
m	12.3	10.7	5.82	5.65	12.8	7.59	4.80	5.64	12.0	3.56	4.30	7.21	10.8	7.95	12.0	11.8	1.39	5.81	10.6	5.84	10.7	4.43	4.88	.	9.84	2.89
n	12.0	8.42	12.2	13.3	12.8	10.2	13.2	9.07	12.1	8.40	10.5	10.1	9.81	10.9	12.1	7.93	7.36	8.38	12.4	13.1	10.7	10.3	8.30	7.55	11.1	7.75

APPENDIX (Continued)

Table A3 (Continued)

	Successor																									
Predecessor	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
o	10.5	10.6	11.2	11.3	9.82	13.0	10.6	9.43	10.8	8.47	10.6	12.1	12.6	13.6	11.7	11.7	6.27	13.4	11.9	12.2	12.9	11.5	12.0	8.62	9.85	8.30
p	12.0	6.81	5.74	5.81	12.3	6.47	7.80	10.4	11.0	4.91	6.47	11.8	8.67	5.60	12.0	11.1	2.08	12.2	10.2	10.4	10.9	1.10	5.85	.00	8.97	3.58
q	3.30	1.95	.	.00	.69	.69	.00	.	6.64	.	.	1.10	.	.00	.00	.	1.10	.	1.10	1.95	10.9	1.79	.69	.	.	.
r	12.8	9.63	11.2	11.6	13.7	9.61	11.0	8.83	12.7	5.58	11.4	10.8	11.3	11.5	12.8	10.1	6.18	11.0	12.5	12.2	11.0	11.0	8.79	5.47	11.6	6.92
s	12.1	8.59	11.1	9.10	12.9	8.62	6.66	12.0	12.5	4.16	10.2	10.4	10.1	9.18	12.0	11.3	8.25	8.72	12.2	13.2	11.7	6.44	9.15	.00	9.75	5.97
t	12.5	8.43	9.81	7.32	13.3	8.08	7.71	14.2	13.2	4.17	6.14	10.7	9.83	8.95	13.2	7.33	1.95	12.2	12.2	11.4	11.4	6.49	10.5	2.83	11.5	8.49
u	10.9	10.8	11.3	10.9	11.2	9.19	11.1	6.29	10.8	5.98	8.08	11.9	11.0	12.1	8.53	11.2	5.73	12.3	12.3	12.4	5.42	7.31	6.43	7.56	8.94	7.43
v	10.8	.00	6.05	4.20	12.9	.00	4.52	1.39	12.0	.69	3.18	5.15	2.40	5.34	10.4	.	.	6.23	6.43	3.71	6.76	4.74	.	.	8.11	1.10
w	12.1	6.89	6.09	7.88	11.8	6.22	3.71	12.0	12.1	.	7.07	8.85	7.31	10.7	11.7	5.53	1.95	9.53	10.0	7.87	4.69	.69	2.94	.00	8.43	2.20
x	9.11	3.78	9.11	.00	9.61	5.59	1.10	7.76	9.11	.	.	5.06	4.39	2.94	7.47	10.3	4.76	1.39	4.55	9.75	7.96	4.62	6.11	5.06	6.00	.00
y	9.16	8.22	8.48	7.77	11.4	6.35	6.34	6.24	9.92	3.09	5.96	9.07	9.49	9.11	10.6	8.63	1.61	8.97	10.7	8.86	6.57	5.64	8.04	4.42	3.43	6.56
z	9.41	5.86	3.69	4.73	10.1	2.56	5.00	6.32	8.93	.69	5.29	6.95	6.00	4.67	8.13	4.98	3.14	4.75	4.98	4.84	6.85	4.84	5.21	.	7.23	8.10

Table A4
Lowercase-by-Uppercase Bigram Frequency

	Successor																									
Predecessor	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
a	.00	4.09	3.40	.00	.	4.83	3.81	.00	.00	.00	.	.69	3.89	.	.00	3.93	1.39	3.40	3.04	2.20	.00	3.50	.69	.	.00	1.95
b	.	.	.	.	.	.	.	1.39	.00	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
c	5.27	3.76	7.47	7.00	5.89	3.74	6.55	4.77	4.91	.	6.24	5.93	5.62	5.86	.00	4.65	3.40	4.38	3.93	2.94	.00	4.78	2.83	.	.	.
d	.	3.30	1.10	.	2.71	.	.	.00	1.39	.00	.00	.	.	.	.	2.94	.	.00	2.77	.	.	.	.00	.	.	.
e	1.79	5.43	5.12	2.30	.	4.06	4.48	2.64	.00	2.56	3.18	5.24	4.62	2.71	.	4.63	.00	3.66	3.95	1.61	.00	4.33	4.25	.	.00	1.79
f	.00	.	.00	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.00	.	.	.	.	.	.	.
g	1.79	.	.	.	.	.	1.10	.	.	.	.	.00	.	.	.	.	.	.	1.79	.	.	.	.	.	.	.
h	.	.	.69	.00	1.39	.00	.	.	.	.	3.18	.	.	.	.00	.	.	.	.	.	.	.	.	.	.	.
i	.00	4.36	3.18	.	.	4.53	4.06	.	1.10	.	.	2.30	3.83	2.08	.00	3.30	.	1.39	3.04	1.10	.00	2.30	.	.00	.	.
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
k	4.92	.	.	.	.	.	.	.	.00	.	.	.	1.10	.	.	.00	.	.	.	.00	.	.	.	.	.	.
l	.	2.20	1.79	.	.	1.10	1.39	.	.00	.	.	.	.	.	.	.	.	.	4.29	.	.	.69	.	.	.	.
m	.00	.	1.39	.	.	.00	.	.	.00	.	.	.00	.	.	.	.00	.	.	.00	.00	.	.	.	.	.	.
n	2.48	1.79	1.39	1.10	.00	4.29	.00	1.10	.69	.	.	.	.	.00	1.79	.	.	1.61	.69	.69	.	.00	1.61	.00	.00	.
o	3.00	1.61	2.08	.00	1.10	.00	1.79	5.43	.	1.61	1.61	.	2.30	.00	.	.	.	.69	3.33	1.61	.	.	1.10	.	.00	.
p	2.30	.	.	.	.	.	.	3.30	.00	.	.00	.	.	.	.	.	.	2.20	1.10	.00	.00	.	1.39	.	.	.
q	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
r	.	.69	4.42	.	.	.00	.69	.	.	.	.	.00	1.39	.69	.	.69	.	.	4.01	.00	.	.00	.	.	.	.00
s	1.95	2.89	2.89	.	.00	1.10	.	.69	.00	.	.00	.69	3.14	.00	.69	.69	.	.69	2.48	1.61	.	.	2.89	.	.	.
t	3.33	.00	.69	.	.	3.53	.	.	.00	.	.	1.95	.69	3.26	.	.	.	.	.00	.	.	.	3.78	.	.	.
u	.	3.93	2.89	1.10	.00	.69	.	.	.	.	.	.	.	.	.	3.50	.	.	.	.69	.	.69	.	.	.	.00
v	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.00	.	.	.	.	.	.
w	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.00	.	.	.	.
x	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
y	.00	.	2.20	.	.	.	.69	.	.69	.	.	.00	.69	.	.	.	.	.	.00	.00	.	.	.	.	.	.
z	.	.	.	.	.	.	1.79	.	.	.	1.79	.	.	.	.	1.39	.	.	.	.	.	.	.	.	.	.

(Manuscript received May 6, 2003;
revision accepted for publication June 21, 2004.)