Type-based bigram frequencies for five-letter words

LAURA R. NOVICK

Vanderbilt University, Nashville, Tennessee

and

STEVEN J. SHERMAN

Indiana University, Bloomington, Indiana

Researchers often require subjects to make judgments that call upon their knowledge of the orthographic structure of English words. Such knowledge is relevant in experiments on, for example, reading, lexical decision, and anagram solution. One common measure of orthographic structure is the sum of the frequencies of consecutive bigrams in the word. Traditionally, researchers have relied on *token-based* norms of bigram frequencies. These norms confound bigram frequency with word frequency because each instance (i.e., token) of a particular word in a corpus of running text increments the frequencies of the bigrams that it contains. In this article, the authors report a set of *type-based* bigram frequencies in which each word (i.e., type) contributes only once, thereby unconfounding bigram frequency from word frequency. The authors show that type-based bigram frequency is a better predictor of the difficulty of anagram solution than is token-based frequency. These norms can be downloaded from www.psychonomic.org/archive/.

Researchers in many different fields require subjects to make judgments that call upon their knowledge of the orthographic structure of English words—that is, their knowledge of what are common and uncommon sequences of letters. Such knowledge is relevant in experiments on natural language processing using reading-based tasks such as word identification and lexical decision, as well as in reading itself. For example, words that fit common (therefore, highly familiar) spelling patterns (e.g., *frame*) might be processed more quickly than words whose spellings are less regular (e.g., *judge*). Orthographic knowledge is also relevant for problem-solving tasks such as unscrambling anagrams and deciphering crossword puzzle clues. For example, given the anagram *nadts*, it would make sense to try *st* before *sn* as the first two letters of the solution, because more five-letter words begin with *st* than with *sn*. And it is only a short step from solving anagrams and crossword clues to completing word fragments for a test in a memory experiment or to identifying (or perhaps failing to identify) typographical errors (e.g., *stadn*).

In many studies across a variety of fields over at least the past 40 years, it has been necessary to either equate or manipulate (or perhaps just to measure) orthographic characteristics of the stimuli (e.g., bigram frequencies) in order to generate words (or nonwords) whose fits to the

constraints of English spelling are known (e.g., Dorfman, 1999; Mayzner & Tresselt, 1963; Mendelsohn & O'Brien, 1974; Rice & Robinson, 1975; Seidenberg, Waters, Barnes, & Tanenhaus, 1984; Srinivas, Roediger, & Rajaram, 1992; Westbury & Buchanan, 2002). A convenient measure of how well a letter string conforms to the common spelling patterns of English, although by no means the only such measure, is its summed bigram frequency (SBF). The SBF of a letter string is the sum of the frequencies of its successive bigrams. A five-letter word, for example, has four successive bigrams, in positions one and two, two and three, three and four, and four and five.

Bigram frequencies can be computed in a variety of different ways. The frequencies can be sensitive to letter position or not. Position-sensitive norms take into account the fact that *ck*, for example, is reasonably common in the middle or final positions of a word but never occurs in the first two positions. In particular, there are multiple *ck* counters, one for each pair of positions. Non-position-sensitive norms, in contrast, increment a single *ck* counter whenever that bigram is observed, regardless of the position in which it occurs. Because the characterization of position depends on word length (e.g., positions three and four are at the end of a four-letter word but in the middle of a six-letter word), norms that take into account letter position also conditionally on word length. Even when position is conceptually independent of word length (e.g., positions one and two are at the beginning of words of all lengths), bigram frequencies may depend on word length. For example, *gr* starts 19 different five-letter words in English but no three-letter words.

Finally, bigram frequencies can be computed from word tokens or word types. Word types are specific

We thank Erin Winchester for her help in compiling the lexicon on which the bigram norms are based. Correspondence concerning this article should be addressed to L. R. Novick, Department of Psychology and Human Development, Peabody College #512, 230 Appleton Place, Vanderbilt University, Nashville, TN 37203-5721 (e-mail: laura.novick@vanderbilt.edu).

words (e.g., entries in a dictionary). Word tokens are individual instances of the types found in a corpus of running text. For example, in the abstract of this article, the type *example* is represented by a single token (i.e., that word occurs only once) whereas the type *bigram* is represented by five tokens (i.e., that word occurs five times). Initially, computations of the orthographic structure of a word relied on what we will refer to as *token-based* norms of bigram frequency. These norms were computed by counting the frequency of each bigram in a sample of words from running text. Thus, multiple instances (i.e., tokens) of a particular word were counted separately. Underwood and Schulz's (1960) token-based norms did not take word length and letter position into account. Mayzner and Tresselt's (1965) token-based norms did.

For type-based norms, in contrast, every word (i.e., type) contributes only once to the relevant bigram frequencies, regardless of the frequency of that word in the language (i.e., how often it occurs in running text). With token-based norms, certain bigrams appear to be quite common, but only because they occur in a few high-frequency words. For example, according to Mayzner and Tresselt's (1965) token-based norms, 2.6 times as many five-letter words in running text begin with *wo* as with *sh* (163 vs. 62 occurrences out of 3,422 word tokens). The lexicon from which our own type-based frequency norms were computed, however, shows that *wo* begins only 17 different five-letter words (i.e., types), whereas *sh* begins 56 different five-letter words, a 3.3 to 1 difference in the opposite direction. The discrepancy occurs because in the Mayzner and Tresselt norms, the frequency of *wo* is augmented by the high frequencies in the language of *would*, *world*, *woman*, and *women*.

Knowledge of token- and type-based frequencies, as well as of frequencies that are sensitive to word length and letter position or not, may be called upon in different situations, depending on the type of judgment one needs to make. For example, Mayzner and Tresselt (1963) showed that it is important to take into account word length and letter position to understand the difficulty of anagram solution. Westbury and Buchanan (2002) argued that these variables should be ignored in using bigram frequencies to predict lexical decision time, however, so that bigram frequency can be unconfounded from orthographic neighborhood size. Similarly, token-based information seems crucial for deciphering coded text (e.g., Pratt, 1942) and understanding typing skill (e.g., Salthouse, 1984); but type-based information may be more helpful for completing word fragments and solving anagrams, because in those tasks decisions about the placements of individual letters depend primarily on the identities of the other letters presumed or known to be in the word. An advantage of using type-based bigram frequency norms in psychological research is that the contributions to performance of bigram frequency and word frequency can be evaluated separately.

Solso and Juel (1980) previously tabulated type-based norms of bigram frequencies, which they referred to as

bigram *versatilities*. Their norms were compiled separately for two- through nine-letter words taken from Kučera and Francis (1967; exclusion criteria were hyphenated words, words containing apostrophes, and numbers). These norms have several limitations. First, the Kučera and Francis word frequency book contains only a subset of the words in the English language: those that occurred in the texts they used. Importantly, this source excludes many ordinary words (e.g., *chime*, *daisy*, *fudge*, *whale*) that undoubtedly contribute to people's knowledge of bigram frequencies. It is unclear, therefore, how well Solso and Juel's norms capture the actual type-based orthographic structure of English.

A second limitation of Solso and Juel's (1980) norms derives from the fact that only a printed version of them exists: Determining how well the orthographic structure of a particular letter string fits the constraints of English spelling is a time-consuming and potentially error-prone process, especially if such information is desired for more than a few words. For example, to determine the SBF of *judge*, one must look up the frequency of *ju* in the first two positions of a five-letter word (*ju*12 = 18), the frequency of *ud* in positions two and three (*ud*23 = 13), the frequency of *dg* in positions three and four (*dg*34 = 12), and the frequency of *ge* in positions four and five (*ge*45 = 40). Summing these frequencies, one finds that the SBF of *judge* is 83. Repeating this process, one discovers that *frame* is much more orthographically regular than *judge* (SBF = 174), although these two words have very similar frequencies in the language according to the Kučera and Francis (1967) norms (word frequencies of 77 and 74, respectively, out of approximately one million tokens of varying lengths). The difficulty and tediousness of computing SBFs by hand led one researcher to publish bigram statistics for a set of five-letter words of potential use in studies of anagram solution (Gilhooly, 1978). That article reports SBF values based only on Mayzner and Tresselt's (1965) token-based norms, and of course it can be used only when the word in which one is interested occurs in the list.

The fact that only a printed version of Solso and Juel's (1980) norms exists leads to another limitation as well: The only kinds of bigram statistics that it is feasible to generate for a letter string are those that can be looked up in the table about the given letter string—for example, the SBF or the frequency of the most or least frequent bigram in the word. It is impossible, from a practical standpoint, to generate bigram statistics that compare one ordering of a set of letters to another ordering of those same letters. For example, given the letters *u*, *d*, *e*, *g*, and *j*, does the word *judge* represent a likely or unlikely permutation? What about the word *beach*, given the letters *e*, *a*, *h*, *c*, and *b*?

In this article, we report a new set of type-based norms of bigram frequency, as well as a computer program that accesses these norms to calculate both absolute (SBF) and relative (rank order) bigram statistics for a list of submitted letter strings. In solving an anagram, the rank

order of the solution in terms of SBF may be at least as important as the actual SBF, because the task facing the problem solver is to decide which of the possible arrangements of the given letters is the correct one (e.g., see Mendelsohn & O'Brien, 1974). Thus, for example, assuming appropriately equated anagrams, it should be easier to solve an anagram for *juice* than for *polar*, because *juice* has an SBF rank of 1 (according to our norms), whereas *polar* has an SBF rank of 27 (these two words have very similar word frequencies and SBFs).

Our type-based frequency norms are based on 2,550 different words selected from the list of five-letter words in Olson and Schwartz (1967). By comparison, Mayzner and Tresselt's (1965) token-based norms were based on only 856 different (five-letter) words (which occurred with varying frequencies). The Olson and Schwartz list contains most of the five-letter entries from Webster's Third International Dictionary (hyphenated words and combining forms were excluded by Olson and Schwartz). By relying on a dictionary source for our words, we should include the ordinary words that Solso and Juel (1980) missed. We excluded from the Olson and Schwartz list words that fit into any of the following five categories: (1) a four-letter root word to which an *s* has been appended (e.g., *cards*, *looks*), (2) a proper noun (e.g., *Cupid*), (3) a foreign word or British spelling (e.g., *adieu*, *metre*), (4) a word containing an apostrophe (e.g., *didn't*), or (5) a letter string that neither author recognized as being a word (e.g., *adoxa*, *compo*, *scaup*, *zymin*). The last criterion was invoked because we wished to tabulate a set of bigram frequencies that would be likely to describe college-educated adults' knowledge of the orthographic structure of American English. Such norms would seem to be most useful for researchers interested in studying natural language processing among native speakers of that language. Because we were interested in orthographic structure, and not productive vocabulary, the final criterion excluded words that were not recognized, but it did not exclude words that were recognized but could not be defined.

Five files (compressed as a single .zip archive) can be downloaded from the Psychonomic Society's Norms, Stimuli, and Data archive Web site, <http://www.psychonomic.org/archive/>: (1) a file containing the lexicon from which the norms were computed; (2) a file containing position-sensitive type-based frequencies for single letters in five-letter words; (3) a file containing position-sensitive type-based frequencies for bigrams in five-letter words; (4) a file containing the source code for a computer program (written in THINK Pascal for the Macintosh) that will access the bigram norms database and calculate several bigram statistics for any string of five letters; and (5) an instructional file describing the contents of these five files and how to use the program.

The computer program provides the following information, as is illustrated in Tables 1 and 2: First, summary information is printed, which includes the letter string entered (e.g., *JUDGE* in Table 1), its rank order in terms of SBF among the 120 possible orders of the five letters in the letter string, and its SBF. For letter strings with an

SBF rank of 1 (e.g., *BEACH* in Table 2), the program also computes two measures of the distance between the Rank 1 letter order and its close neighbors: (1) the difference between the SBFs of the Rank 1 and Rank 2 letter orders and (2) the difference between the SBF of the Rank 1 order and the average SBF for the rank 2–5 orders. After printing the summary information about the letter string, the program prints the SBFs of the 12 arrangements of the given letters that have the highest SBFs (i.e., the arrangements ranked 1–12).

We have suggested that position-sensitive type-based bigram frequencies may be useful for understanding the difficulty of anagram solution. To test this hypothesis, we compared type-based and token-based measures of SBF as predictors of mean solution time and accuracy for anagrams of 108 solution words used in two experiments.¹ Experiment 1 used 60 solution words that were selected to vary in terms of whether the solution had the top-ranked SBF according to our new norms (some words

Table 1
Sample Output for "Judge" From the Computer Program That Computes the Summed Bigram Frequency (SBF) of Any Five-Letter Stimulus, as Well as the SBFs of the Top 12 Letter Strings That Can Be Made From the Given Letters

Solution Word: JUDGE Rank: 7 SBF: 0.0242			
Top 12 Letter Orders:			
Rank	SBF	Letter String	Letter Order
1	0.0823	JUGED	12453
2	0.0757	UJGED	21453
3	0.0745	GUJED	42153
4	0.0709	JGUED	14253
5	0.0709	GJUED	41253
6	0.0682	UGJED	24153
7	0.0242	JUDGE	12345
8	0.0220	GUDEJ	42351
9	0.0192	JUDEG	12354
10	0.0180	JUGDE	12435
11	0.0180	DEGUJ	35421
12	0.0177	GUJDE	42135

Table 2
Sample Output for "Beach" From the Computer Program That Computes the Summed Bigram Frequency (SBF) of Any Five-Letter Stimulus, as Well as the SBFs of the Top 12 Letter Strings That Can Be Made From the Given Letters

Solution Word: BEACH Rank: 1 SBF: 0.0639			
Distance to #2 ranked letter order = 0.0180			
Average distance to 2nd–5th ranked letter orders = 0.0220			
Top 12 Letter Orders:			
Rank	SBF	Letter String	Letter Order
1	0.0639	BEACH	12345
2	0.0459	CHEAB	45231
3	0.0408	EBACH	21345
4	0.0408	BAECH	13245
5	0.0400	BHACE	15342
6	0.0357	HEACB	52341
7	0.0349	ABECH	31245
8	0.0349	EABCH	23145
9	0.0340	CHABE	45312
10	0.0313	CHAEB	45321
11	0.0294	CHEBA	45213
12	0.0283	BHEAC	15234

were top ranked; others had SBF ranks of 8–25). Subjects were given up to 30 sec to solve each anagram. Experiment 2 used 48 solution words, again selected to be top ranked in terms of SBF or not (rank of 1 vs. ranks of 8–26). Subjects were given up to 95 sec to solve each anagram. In each experiment, means were computed for each solution word across subjects (N s of 96 and 48, respectively, in Experiments 1 and 2). To put the data from the two experiments on the same scale, the Experiment 2 data were scored as if subjects received only 30 sec for solution. Mean times were computed on the basis of correct solutions and timeouts (i.e., cases in which the subject failed to solve the anagram within 30 sec, in which case a time of 30 sec was assigned); only incorrect solutions were omitted (e.g., responding *omlet* for the anagram of *motel*; errors were rare). Accuracy rates were computed as the proportion of subjects who gave the solution word within 30 sec.

We conducted a separate simultaneous multiple regression analysis for each dependent variable using four predictor variables: (1) the SBF of the solution word computed using Mayzner and Tresselt's (1965) token-based norms, (2) the SBF of the solution word computed using our new type-based norms, (3) whether the type-based SBF was top ranked among the 120 possible orders of the five letters (coded "1" if yes and "–1" if no), and (4) word frequency according to the Kučera and Francis (1967) norms (the actual predictor variable was the natural logarithm of the raw frequencies, which ranged from 0 to 365; 1 was added to each frequency before computing the natural logarithm). In both regression analyses, only TopRank was a reliable predictor of solution difficulty. For mean solution time, $R = .33$, $F(4,103) = 3.13$, $p < .02$. TopRank had $\beta = -.36$, $t(103) = -3.17$, $p < .01$ ($p > .17$ for each of the other three predictors). Subjects spent less time trying to solve anagrams whose solution words were top ranked in terms of type-based bigram frequency ($M = 15.52$ sec, based on 64 words) than anagrams whose solution words were relatively unlikely in terms of type-based bigram frequency in comparison with the other possible orderings of the given letters ($M = 18.58$ sec, based on 44 words). For mean accuracy, $R = .32$, $F(4,103) = 2.94$, $p < .03$. TopRank had $\beta = .34$, $t(103) = 3.04$, $p < .01$ ($p > .25$ for each of the other three predictors). Subjects were more likely to solve (within 30 sec) anagrams whose solutions were top ranked in terms of type-based SBF ($M = .66$) than those whose solutions were not top ranked ($M = .54$). As predicted, a measure of orthographic structure computed from type-based bigram frequencies was a better predictor of the difficulty of anagram solution than was a token-based indicator of orthographic structure.

People may call upon their knowledge of the orthographic structure of English when performing a variety of cognitive tasks, including reading, spelling, identifying words, making lexical decisions, completing word fragments, solving anagrams, deciphering crossword puzzle clues, and breaking codes. Various measures of ortho-

graphic structure have been investigated, including, for example, syllabification, orthographic neighborhood size, and bigram frequency. Researchers interested in the effects of bigram frequency on language processing have argued for and used a variety of different measures of bigram knowledge. As discussed earlier, bigram norms can be computed taking into account word length and letter position or not and weighting individual words by their frequency in the language (token-based norms) or not (type-based norms). In this article, we have presented a new set of position-sensitive type-based bigram frequency norms for five-letter words that are computer searchable, and we have shown that a measure of how well a word fits the constraints of English spelling that was derived from these norms better predicts the difficulty of anagram solution than does a measure derived from a comparable set of token-based norms. The new norms provide a tool for language researchers that allows orthographic structure to be decoupled from word frequency.

REFERENCES

- DORFMAN, J. (1999). Unitization of sublexical components in implicit memory for novel words. *Psychological Science*, **10**, 387–392.
- GILHOOLY, K. J. (1978). Bigram statistics for 205 five-letter words having single-solution anagrams. *Behavior Research Methods & Instrumentation*, **10**, 389–392.
- KUČERA, H., & FRANCIS, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- MAYZNER, M. S., & TRESSELT, M. E. (1963). Anagram solution times: A function of word length and letter position variables. *Journal of Psychology*, **55**, 469–475.
- 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.
- MENDELSON, G. A., & O'BRIEN, A. T. (1974). The solution of anagrams: A reexamination of the effects of transition letter probabilities, letter moves, and word frequency on anagram difficulty. *Memory & Cognition*, **2**, 566–574.
- OLSON, R., & SCHWARTZ, R. (1967). Single and multiple solution five-letter words. *Psychonomic Monograph Supplements*, **2** (8, Whole No. 24), 105–152.
- PRATT, F. (1942). *Secret and urgent: The story of codes and ciphers*. Garden City, NY: Blue Ribbon Books.
- RICE, G. A., & ROBINSON, D. O. (1975). The role of bigram frequency in the perception of words and nonwords. *Memory & Cognition*, **3**, 513–518.
- SALTHOUSE, T. A. (1984). Effects of age and skill in typing. *Journal of Experimental Psychology: General*, **113**, 345–371.
- SEIDENBERG, M. S., WATERS, G. S., BARNES, M. A., & TANENHAUS, M. K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning & Verbal Behavior*, **23**, 383–404.
- SOLSO, R. L., & JUEL, C. L. (1980). Positional frequency and versatility of bigrams for two- through nine-letter English words. *Behavior Research Methods & Instrumentation*, **12**, 297–343.
- SRINIVAS, K., ROEDIGER, H. L., III, & RAJARAM, S. (1992). The role of syllabic and orthographic properties of letter cues in solving word fragments. *Memory & Cognition*, **20**, 219–230.
- UNDERWOOD, B. J., & SCHULZ, R. W. (1960). *Meaningfulness and verbal learning*. Chicago: Lippincott.
- WESTBURY, C., & BUCHANAN, L. (2002). The probability of the least likely non-length-controlled bigram affects lexical decision reaction times. *Brain & Language*, **81**, 66–78.

NOTE

1. A more complete analysis of these data considering a variety of anagram, word, and solver factors as predictors of the difficulty of anagram solution is currently in preparation.

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, search the archive for this article using the journal (*Behavior Research Methods, Instruments, & Computers*), the first author's name (Novick), and the publication year (2004).

FILE: Novick-BRMIC-2004.zip

DESCRIPTION: The compressed archive file contains five files:

lexicon.txt, containing the lexicon of words from which the bigram and single-letter frequencies were generated, as a 16k text file generated by Microsoft Word Version 10.1.1 for the Macintosh.

bigram frequencies.txt, containing the position-sensitive type-based frequencies for bigrams in five-letter words developed by Novick and Sherman (2004), as a 20K text file generated by Microsoft Word Version 10.1.1 for the Macintosh.

single letter frequencies.txt, containing the position-sensitive type-based frequencies for single letters in five-letter words developed by Novick and Sherman (2004), as a 4K text file generated by Microsoft Word Version 10.1.1 for the Macintosh.

program code.rtf, a 40K file containing the source code for a computer program that will access the bigram norms file and calculate several statistics for any string of five letters. This program was written using THINK Pascal Version 4.0.1 for the Macintosh. The source code is provided for users who wish to translate the program to run on other platforms. An executable version of the original program may be obtained upon request from Laura Novick.

archive description.rtf, an 8K file containing a description of the files included in the archive and instructions for running the computer program. &

AUTHORS' E-MAIL ADDRESSES: laura.novick@vanderbilt.edu, sherman@indiana.edu.

FIRST AUTHOR'S WEB SITE: <http://www.peabody.vanderbilt.edu/novick/>.

(Manuscript received December 29, 2003;
revision accepted for publication July 12, 2004.)