



WORD

ISSN: 0043-7956 (Print) 2373-5112 (Online) Journal homepage: <https://www.tandfonline.com/loi/rwrd20>

The Relative Frequency of Phonemes in General-American English

Rebecca E. Hayden

To cite this article: Rebecca E. Hayden (1950) The Relative Frequency of Phonemes in General-American English, WORD, 6:3, 217-223, DOI: [10.1080/00437956.1950.11659381](https://doi.org/10.1080/00437956.1950.11659381)

To link to this article: <https://doi.org/10.1080/00437956.1950.11659381>



Published online: 04 Dec 2015.



Submit your article to this journal [↗](#)



Article views: 20773



View related articles [↗](#)



Citing articles: 11 View citing articles [↗](#)

THE RELATIVE FREQUENCY OF PHONEMES IN GENERAL-AMERICAN ENGLISH

REBECCA E. HAYDEN

1. *Introduction.* The purpose of this study is to analyze the relative frequency of occurrence of phonemes in a representative sample of General-American English. Bloomfield in *Language* has said, "If we take a large body of speech, we can count out the relative frequencies of phonemes and of combinations of phonemes. This task has been neglected by linguists and very imperfectly performed by amateurs, who confuse phonemes with printed letters."¹ The importance of this type of linguistic analysis has recently been stated by Reed:

Linguistic analysis . . . tends to be largely qualitative: the descriptive techniques of phonetics, phonemics, morphology, and syntax aim generally to analyze the nature and variety of linguistic phenomena rather than the magnitude and frequency of such phenomena. . . . Qualitative study of this sort is a prerequisite to successful quantitative analysis, so that the present results of qualitative linguistic analysis furnish the best basis upon which quantitative study may rest. . . . Qualitative analysis has been pursued for so long and with such thoroughness that it will now be fruitful for linguists to turn some of their attention to quantitative analysis.²

The present study was prompted by the writer's interest in problems of pronunciation faced in the teaching of English as a foreign language to students from abroad. It seemed that there might be a pedagogically more efficient and linguistically sounder approach to pronunciation problems if the relative frequency of occurrence of the various American English phonemes were determined. It is further believed that the data will be of value in other areas of the speech and linguistics fields.

2. *Previous studies.* To this writer's knowledge, previous analyses of the phonemes of American English have not been made with a quantitative statistical approach. All of the previous studies have been in the nature of what has been called "quantitative pre-analyses;"³ that is, raw totals have been counted and percentages calculated. Also, as a matter of fact, the previous analyses have actually been phonetic rather than phonemic. Furthermore, it is believed that there has been no study of the particular sort of speech material selected for this study. (See Materials and Procedures below.)

2.1. *Previous Studies Based on Written Matter.* Two earlier studies were based upon printed materials. Both of these studies used dictionary listings as the basis for the pronunciation of the words in the reading matter analyzed. In 1923, Godfrey Dewey's *Relativ Frequency of English Speech Sounds* was pub-

¹ Leonard Bloomfield, *Language*, (New York, 1933), Henry Holt, 136-7.

² David W. Reed, "A Statistical Approach to Quantitative Linguistic Analysis," *Word*, V, no. 3 (1949), 235.

³ *Ibid*, p. 236.

lished.⁴ In 1926, Ruth E. Atkins made a count of the speech sounds in Thorndike's word list.⁵ Dewey's was a commendable and monumental study of 100,000 words (372,729 sounds) occurring in a wide range of connected written matter.⁶ G. K. Zipf has said of this study: "It would, indeed, be unkind to think of a possible control for these figures, which are so extensive and so accurate in analysis. Yet even these cannot be said to represent any actually spoken language; nor was that the author's intent. . . . Let us repeat again, that the present law of frequency can only be exactly demonstrated in an actually spoken dialect."⁷

2.2. *Previous Studies Based upon Spoken Materials.* Two known past counts of American English sounds based upon spoken materials have been made. In 1930, French, Carter, and Koenig presented a study of the occurrences of speech sounds in 500 telephone conversations recorded in New York City;⁸ in this study, articles were omitted and thus the figures cannot be considered as completely representative of all speech sounds. (Although for their "purposes connected with the design of telephone circuits"⁹ this omission was reasonable.) In 1935, C. H. Voelker made a count of the speech sounds in some 5,946 radio announcements.¹⁰ This study comes closest to being similar to that made by the present writer. However, it must be pointed out that, aside from the difference in speech materials selected for analysis, in the Voelker study all types and dialects were represented together in the figures. Since phonemic analyses necessarily differ with dialects, a given count should be based on a fairly uniform dialect; therefore, it is difficult to judge *what* was counted in the former study. Nonetheless, in terms of raw percentages, it is interesting to note the closeness of figures, in the Voelker study and in that of this writer, for total vowels and consonants.

	Vowels	Consonants
Voelker:	38.2%	61.8%
Hayden:	37.4%	62.6%

3. *Materials and Procedures.*

3.1. *Materials.* The sampling of General-American dialect selected for this study was a series of six lectures given in the Orientation and English Language Program for Foreign Students at the University of California. This selection was made because of the diversity of topics¹¹ discussed and because the material

⁴ Godfrey Dewey, *Relative Frequency of English Speech Sounds*, (Cambridge, 1923), Harvard University Press.

⁵ Ruth E. Atkins, "An Analysis of the Phonetic Elements in a Basal Reading Vocabulary," *The Elementary School Journal*, XXVI, no. 8 (1926), 595-606.

⁶ Godfrey Dewey, op. cit., pp. 8-9, 123-132.

⁷ George K. Zipf, "Relative Frequency as a Determinant of Phonetic Change," *Harvard Studies in Classical Philology*, XL (1929), 52.

⁸ N. R. French, C. W. Carter, Jr., and W. Koenig, Jr. "The Words and Sounds of Telephone Conversations," *The Bell System Technical Journal*, IX (1930), 290-324.

⁹ *Ibid.*, p. 324.

¹⁰ Charles H. Voelker, "A Sound Count for the Oral Curriculum," *The Volta Review* XXV, no. 3 (1935), 55-6.

¹¹ The subjects for the lectures included "Problems of the Foreign Student in the United States," "Current Issues in Education in the United States," "Labor in the United

seemed relevant to the writer's interest in the language difficulties of foreign students attending a university. The speakers were members of the administration or faculty of the University. The language of the six speakers showed so little divergence that their speech could be analyzed according to a single system of phonemes.¹²

3.2. *Procedures.*

3.21. *How the Material Was Gathered.* The lectures were wire-recorded and then transcribed phonemically.¹³ From the transcriptions, the phonemes occurring in each lecture were counted, and the raw totals of the phonemes in all the lectures were calculated in terms of their percentages of the grand total of 65,122 phonemes. This breakdown is shown on Table One.

3.22. *How the Sampling was Tested for Reliability.* In order to test the reliability of the sampling and to determine the significance of the results obtained, the data were further analyzed in accordance with statistical theory. The theory, including formulae, was that set forth by Reed for analyzing linguistic material statistically.¹⁴ First, the Standard Error of a Proportion was determined for the total of each phoneme counted, according to the formula: $SE = \sqrt{\frac{pq}{N}}$ (p represents frequency percentage of a proportion; q represents the frequency percentage of the remaining proportions; N represents the total number of occurrences).

Second, the Standard Error of Difference between two proportions was calculated for the totals of each successive group of two phonemes (arranged according to relative frequency of occurrence in vowel and consonant groups); the formula: $SE_{dif} = \sqrt{SE_1^2 + SE_2^2}$. Third, the Standard Error Deviations for these groups were made by dividing their frequency percentage differences by their SE_{dif} . Fourth, and finally, the significance or non-significance of the SE Deviations was determined by checking a probability chart¹⁵ and following the theory that "a percentage of probability above 5.00 is usually taken to indicate a mere chance deviation; a percentage of probability below 5.00 is usually taken to indicate a strong likelihood that some factor other than chance is responsible for the deviation; and a percentage of probability below 1.00 is

States," "Politics and Government in the United States," "Agriculture and Forestry in the United States," and "The United States in International Trade."

¹² The speakers had the general characteristics of speech described by phoneticians (e.g. John S. Kenyon, *American Pronunciation*), such as pronunciation of r's, use of /æ/ in such words as *class*, etc. of the general American dialect.

¹³ The phonemic analysis, in general, followed the system of Kenneth L. Pike (*Phonemics*, University of Michigan, 1947, p. 45), including tabulation of /aⁱ/, /a^u/, and /oⁱ/ separately as if they were single units in line with Pike's "close-knit sequences of vowel units," and in using the unit symbols /č/, /j/, and /w/ instead of the diagraphs /tʃ/, /dʒ/, and /hw/. The following exceptions to the Pike system were made: (1) syllabicity was considered as a suprasegmental factor and thus Pike's separate syllabic phonemes /m/, /n/, /l/, and /r/ were disregarded; (2) the consonant /y/ was analyzed as two phonemes /h/ and /y/ because its infrequency (only one time in half of the material) made it seem insignificant as a unit.

¹⁴ David W. Reed, *op. cit.*, pp. 237-247.

¹⁵ *Ibid.*, p. 244.

TABLE ONE
Statistical Analysis of Phoneme Count

Phoneme	Number of Occurrences	Frequency Percentage	SE	SE _{adj}	SE Deviation	Percentage of Probability
		<i>per cent</i>	<i>per cent</i>	<i>per cent</i>		<i>per cent</i>
ə	6488	9.96	.1225			
I	6350	9.75	.1163	.1689	7.16	Below 1.00
æ	2011	3.09	.0678	.1339	20.07	Below 1.00
ɛ	1324	2.03	.0552	.0875	12.11	Below 1.00
e	1265	1.94	.0541	.0773	1.16	24.60
a	1174	1.80	.0521	.0750	1.86	6.28
i	1083	1.66	.0500	.0722	1.93	5.36
u	987	1.52	.0479	.0693	2.02	4.34
O	971	1.49	.0474	.0674	0.44	66.00
a ⁱ	954	1.46	.0470	.0667	0.44	66.00
ɔ	667	1.02	.0394	.0612	7.18	Below 1.00
U	648	0.99	.0388	.0552	0.54	58.92
a ^u	414	0.64	.0312	.0497	7.04	Below 1.00
o ⁱ	42	0.06	.0095	.0340	17.05	Below 1.00
	24378	37.4				
n	5179	7.95	.1060	.1485	2.42	1.56
t	4945	7.59	.1038	.1440	3.40	Below 1.00
r	4625	7.10	.0999	.1308	16.89	Below 1.00
s	3186	4.89	.0845	.1120	11.07	Below 1.00
l	2377	3.65	.0735	.1018	2.94	Below 1.00
ɖ	2180	3.35	.0705	.0987	1.41	15.86
d	2091	3.21	.0691	.0959	2.39	1.68
k	1942	2.98	.0666	.0933	1.17	24.20
m	1870	2.87	.0654	.0884	5.76	Below 1.00

TABLE ONE—*Continued*

Phoneme	Number of Occurrences	Frequency Percentage	SE	SE _{diff}	SE Deviation	Percentage of Probability
		<i>per cent</i>	<i>per cent</i>	<i>per cent</i>		<i>per cent</i>
z	1539	2.36	.0594			
v	1518	2.33	.0591	.0838	0.35	72.64
p	1468	2.25	.0581	.0828	0.96	33.70
w	1151	1.77	.0516	.0777	6.17	Below 1.00
b	1075	1.65	.0499	.0718	1.67	9.50
f	1051	1.61	.0493	.0702	0.56	57.54
y	780	1.20	.0425	.0651	6.30	Below 1.00
g	745	1.14	.0416	.0595	1.00	31.74
h	721	1.11	.0410	.0584	0.50	61.70
š	565	0.87	.0362	.0545	4.40	Below 1.00
ŋ	522	0.80	.0348	.0501	1.39	16.46
č	343	0.53	.0282	.0451	5.98	Below 1.00
j	325	0.50	.0273	.0391	0.76	44.72
θ	286	0.44	.0256	.0373	1.60	10.96
ɰ	241	0.37	.0237	.0350	2.00	4.56
ž	19	0.03	.0063	.0245	13.87	Below 1.00
	<u>40744</u>	<u>62.6</u>				
	<u>65122</u>	<u>100.0</u>				

Key Words: /ɒ/ cup, about; /ɪ/ fit; /æ/ cat; /ɛ/ led; /e/ date; /a/ calm; /i/ feet; /u/ noon; /o/ note; /aː/ time; /ɔ/ all; /ʊ/ foot; /aʊ/ house; /oʊ/ oil.

/n/ no; /t/ tea; /r/ red, car; /s/ see; /l/ let; /ð/ the; /d/ day; /k/ kill; /m/ may; /z/ zoo; /v/ voice; /p/ pay; /w/ wall; /b/ bay; /f/ fly; /y/ yes; /g/ go; /h/ hot; /ʃ/ she; /ŋ/ sing; /č/ check; /j/ jar; /θ/ thing; /ɰ/ when; /ž/ pleasure.

usually taken to indicate almost the certainty that some factor other than chance is responsible for the deviation.”¹⁶ The calculations outlined here are shown on Table One.

4. Results (See Table One).

4.1 The Relative Frequency of Occurrence of Phonemes in a Sample of General-

¹⁶ Ibid., p. 243-4.

American English. A total of 65,122 phonemic units, in six lectures given by university lecturers, was counted and analyzed in terms of thirty-nine phonemes, including fourteen vowels and twenty-five consonants. The raw total of vowels was 24,378, or 37.4%, and the raw total of consonants was 40,744, or 62.6%. The relatively most frequent phoneme was the vowel, /ə/, which occurred 6,488 times, or 9.96%; the second most frequent was also a vowel, /I/ at 6,350 times, or 9.75%. Next in highest relative frequencies were seven consonants, respectively /n/ at 5,179, or 7.95%; /t/ at 4,945, or 7.59%; /r/ at 4,625, or 7.10%; /s/ at 3,186, or 4.89; /l/ at 2,377, or 3.65%; /d/ at 2,180, or 3.35%; /d/ at 2,091 or 3.21%. The tenth was a vowel, /æ/, at 2,011, or 3.09%; The first ten most frequent phonemes, those with percentages from 9.96% to 3.09%, together amounted to 60.54% of the sample of speech analyzed. Thus approximately only one-fourth of the thirty-nine phonemes made up approximately three-fifths of the total phonemic units (65,122) in the material analyzed.

The relatively most infrequent phoneme analyzed was /ʒ/, with only 19 occurrences, or 0.03%. There were nine other phonemes with percentages under 1.00%. Together these ten low-ranking phonemes made up only 5.23%, or approximately one-twentieth of the material analyzed.

The remaining nineteen in-between phonemes, with percentages from 2.98% to 1.02%, made up 34.19%, or approximately one-third of the material analyzed.

4.2 *The Reliability of the Sampling and the Significance of the Results of the Phoneme Count.* Following the statistical procedure for determining the reliability of a sampling, the Standard Error of a Proportion formula was applied to the data of the sample selected. A glance at Table One shows that the standard error for the various proportions ranges from .0095% to .1225% for the vowel phonemes, and from .0063% to .1060% for the consonant phonemes. Since the possible range for standard error is between zero and 35.4%, with zero indicating the most reliable sample and 35.4% the least reliable, it would appear that the sample selected for this study has an extremely high degree of reliability.

Also, the Standard Error of the Difference between two proportions was determined for each successive group of two phonemes. The range here, only slightly higher than that of the standard errors for the individual proportions, from .0340% to .1689% for vowel phonemes, and from .0245% to .1485% for consonant phonemes, further bears out the very high degree of reliability of the data.

Furthermore, the Standard Error Deviations for the above groups were calculated and their significance or non-significance determined. (See Section 3.22 above). A study of Table One reveals that, among the ten most relatively frequent phonemes, there is a significance in the Standard Error Deviations between nine of the phonemes and their successive phonemes; the only non-significant deviation in the top ten phonemes is between /ḍ/ and /d/. In other words, bearing in mind that these ten phonemes represent over sixty percent of the total count, the study shows that there is almost certainty that some

factor other than chance is responsible for the Standard Error Deviations for most of these particular phonemes in the sample analyzed. For the remaining phonemes, which compose less than forty percent of the total count, the deviations are significant for somewhat over one-third of the groups of two successive phonemes, and non-significant for somewhat less than two-thirds of those groups. In other words, some of these deviations show that there is probability that some factor other than chance is responsible, but most of these deviations show that there is probability that mere chance is operating.

5. *Conclusions.* The foregoing analysis of the relative frequency of occurrence of phonemes in a sample of General-American English spoken material, leads to the following conclusions: (1) The study shows that in the case of the phonemes which occur relatively most frequently, that is from 9.96% to 3.09% each, including /ə/, /I/, /æ/, /n/, /t/, /r/, /s/, /l/, /d/, /d/, the standard error deviations, on the whole are significant, but in the case of the other phonemes, all under 3.00%, the standard error deviations tend to be less significant. (2) It seems to the writer that this study should be of practical value in the teaching of English as a foreign language, as well as in other areas of the speech and linguistics fields. For example, in the light of the above analysis, it would seem efficient and sound to approach the correction of phonemic difficulties of foreign students in an order that relates to the relative frequency of occurrence of the various phonemes; thus, /ə/ and /I/, among the vowels, should be the focus of attention in early correction situations. Furthermore, the knowledge that foreign students do have much difficulty with these particular phonemes, bears out the intent of this conclusion. (3) Further studies of this nature would not only add to an increase in raw totals, which would more likely "appropriate the conditions of the still larger body of evidence from which it is drawn,"¹⁷ but such might also lead to even more reliable and significant results than have been possible in this one study. This writer suggests that future studies might also be concerned with such other factors as the relative frequencies of points of articulation (e.g. alveolar) or kinds of articulation (e.g. plosive), consideration of "factors other than chance" operating in significant deviations, and so forth. Therefore, as a final word, it is hoped that others will undertake such quantitative linguistic analyses.

¹⁷ Ibid., p. 238.