Predicting Initial Cluster Frequencies by Phonemic Difference

James E. Cutting*

ABSTRACT

The frequency of occurrence for stop-liquid and stop-semivowel clusters can be predicted on the basis of the number of distinctive features that separate the member phonemes: the greater the phonemic difference, the more frequent the cluster. Predictions made in this manner are generally much better than those made from chance co-occurrence of successive phonemes. Assessments are made on four extensive corpora. Each of six distinctive features is examined individually.

Why does the phonology of a given language permit certain phoneme clusters to occur, but not others? Why do some clusters occur frequently, others rarely? Saporta (1955), among others, suggested that the answer to both questions may lie in an application of Zipf's law (Zipf, 1949): "The relative frequency of consonant clusters will reveal a tendency on the part of any language system to produce speech in such a way as to consider the effort of both the speaker and the listener" (Saporta, 1955:25). Zipf (1935) had applied this principle to descriptions of phonemes, but Saporta (1955; Keller and Saporta, 1957) first applied it to phoneme clusters. Unique to Saporta's approach was the conjoining of a phonetic feature system (Jakobson, Fant, and Halle, 1951) and the principle of least effort. He suggested that phoneme cluster frequency correlated with the number of distinctive features _not_ shared between the member phonemes of a given cluster. Altmann (1969) termed this measure _phonemic difference_. Saporta purposefully excluded clusters with liquids (/l/ and /r/) and semivowels (/w/ and /y/) and suggested that these combinations needed further study. The present paper is concerned with these clusters and with predicting their frequency of occurrence by the principle of phonemic difference.

Carroll (1958) criticized Saporta's analyses on the grounds that inadequate consideration was given to the possible chance nature of the results. After performing the proper analyses, Carroll concluded that phonemic difference has merit as a measure for predicting cluster frequency, but that a much more extensive data base (cluster count) should be used in further research. Following this advice, the present investigation uses data available from several extensive corpora gathered since the publication of the Saporta and Carroll articles. In addition, the principle of phonemic difference is matched against a control

*Also Wesleyan University, Middletown, Conn.

[HASKINS LABORATORIES: Status Report on Speech Research SR-42/43 (1975)]

(or chance-factor) principle, where cluster frequency is predicted from the frequency of occurrence of the member phonemes.

The present research, however, differs in several ways from previous studies. First, it is concerned with only a selected number of clusters where phonemic difference is not extensive. The first member of all clusters is a stop consonant /p,b,t,d,k,g/, and the second member is either a liquid or a semivowel /l,r,w,y/. This limitation removes from consideration the more widely differing clusters used by Saporta and Carroll, which may follow a different principle than those investigated here. Second, this limitation alters the main hypothesis. Saporta suggested that the most frequently occurring clusters should be those with intermediate phonemic difference. Here, I hope to demonstrate that for certain clusters maximal difference in the number of distinctive features shared by successive phonemes serves to predict cluster frequency. It should be remembered, however, that Saporta's intermediate differences and the maximal differences presented in this paper are nearly the same: interphoneme dissimilarities along five features. Third, the present interpretation of phonemic differences is not linked to the principle of least effort. Saporta (1955) noted that for the speaker the situation of least effort should be one in which successive phonemes are most similar, but for the listener that situation is one in which the phonemes are least similar. This led him to predict that intermediate phonemic difference should be important, serving both speaker and listener. However, Wang (1959) found that the least effort principle did not apply to the perceptions of final clusters, thus questioning the role of the listener in this application of Zipf's law. Least efforts for the speaker may be difficult to assess in an unbiased fashion. Therefore, it seems unwise to shackle a principle of phonemic difference to the broader principle of least effort; instead, it should stand alone. Fourth, the present study is concerned only with initial clusters rather than with both initial and final. The major reason for this limitation is simply that there are very few liquid-stop and semivowel-stop clusters in English.

Before presenting evidence supporting a revised phonemic difference principle, a few matters concerning methodological approach must be mentioned. First, the particular distinctive features system used is that proposed by Halle (1964), elaborated from earlier versions (Jakobson, Fant, and Halle, 1951; Cherry, Halle, and Jakobson, 1953). The subsequent feature system of Chomsky and Halle (1968) is rejected on the grounds that many of the additional features are redundant for the particular phonemes selected here. The values for each stop, liquid, and semivowel,are made explicit by Wickelgren (1966) using Halle's definitions, and they are shown in Table 1. Second, each feature is considered equally important. Such an assumption is dangerous. Carrol (1958), for example, suggested that each feature should be considered separately, since voicing alone contributed extensively to Saporta's main finding. Separate analyses can confirm whether phonemic difference is a general principle for phoneme clusters, or whether it is limited only to certain contrasts. To demonstrate the generality of this principle, proper analyses will be performed. Third, the present paper considers all stop-liquid and stop-semivowel combinations to be true clusters. This assertion goes against the view that stop + /y/ combinations, for example, may not be legitimate clusters since they only occur before the vowel /u/ (Hofmann, 1967; Chomsky and Halle, 1968), or that /kl, kr, kw/ combinations might be more parsimoniously described as single consonants (Hofmann, 1967; Menyuk and Klatt, 1968; Menyuk, 1972). For a review of the phoneme-versus-cluster controversy, see Devine (1971). Fourth, in order to predict cluster frequencies and

TABLE 1: Distinctive feature representation from Halle (1964; see also Wickelgren, 1966) of certain phonemes in English that can form initial clusters.

	Stop consonants						Liquids		Semivowels	
	Labials		Alveolars		Velars					
	/p/	/b/	/t/	/d/	/k/	/g/	/l/	/r/	/w/	/y/
Vocalic	–	–	–	–	–	–	+	+	–	–
Consonantal	+	+	+	+	+	+	+	+	–	–
Grave	+	+	–	–	+	+	–	–	+	–
Diffuse	+	+	+	+	–	–	+	–	+	+
Voiced	–	+	–	+	–	+	+	+	+	+
Continuant	–	–	–	–	–	–	+	+	+	+

to confirm these estimates, one needs accurate assessments of actual cluster occurrence. Four extensive corpora were selected: Denes (1965), a corpus including 72,000 phonemes spoken in British English from "phonetic readers" used to teach English to foreign students; Hultzén, Allen, and Miron (1964), a corpus of 20,000 phonemes spoken in General American from selected material in eleven different dramatic plays; Roberts (1965), an analysis of a word list including 15 million entries phonemically transcribed in General American; and Trnka (1966), an analysis of a phoneme count from 100,000 words of connected material transcribed in British English. Fifth, to provide a prediction of cluster frequency based on chance co-occurrence of successive phonemes, accurate assessments of the frequency of each of the ten phonemes in question are needed. The four sources cited above provide these estimates. They agree fairly well with others (Hayden, 1950; Tobias, 1959; Delattre, 1966; Card and Eckler, 1975). See Gerber and Vertin (1969) and Wang and Crawford (1960) for comparisons.

Table 2 displays the phonemic difference between member phonemes for the 24 clusters, along with observed and predicted percentages. Cluster frequencies were determined separately for each corpus. Observed percentages were then calculated by dividing the number of occurrences of each cluster by the total occurrences of all 24 clusters. To compute predicted percentages, the percentage occurrence for each of the ten individual phonemes was first obtained; the product of the percentages for the two phonemes in each cluster was determined; this product was divided by the sum of the products for all 24 clusters, and multiplied by one hundred.

Notice the variation among the four corpora. In particular, variation is greater for observed frequencies than for predicted frequencies. For example, observed frequencies for /pr/, /kr/, and /tl/ differ by factors of 3:1 or greater, whereas predicted frequencies differ by much less.

Eight correlation coefficients were calculated, two for each corpus. First, predicted and observed cluster frequencies were correlated within the same corpus, then the phonemic difference scores were correlated with the obtained cluster frequencies. The results of these analyses are shown in Table 3, along with statistical assessments of each. In addition, the mean predicted and mean observed frequencies were calculated from the four corpora, and correlations

TABLE 2: Twenty-four initial clusters, their phonemic difference, and their observed and predicted frequencies from four different corpora.

	Phonemic difference	Denes (1965) observed	Predicted from Denes (1965)	Hultzén et al. (1964) observed	Predicted from Hultzén et al. (1964)	Roberts (1965) observed	Predicted from Roberts (1965)	Trnka (1966) observed	Predicted from Trnka (1966)
/pr/	5	11.1	2.3	13.3	1.9	25.7	3.0	9.0	5.6
/pl/	4	10.7	3.1	8.3	1.1	6.5	1.5	6.0	3.0
/pw/	3	.1	2.1	.0	2.1	.0	2.1	.0	1.7
/py/	4	.5	1.2	2.2	3.0	.7	3.1	.9	.5
/br/	4	4.6	2.7	6.6	2.4	7.8	3.1	10.4	5.0
/bl/	3	9.6	3.6	3.9	1.4	4.7	1.5	8.1	2.7
/bw/	2	.0	2.5	.0	2.6	.0	2.1	.0	1.5
/by/	3	.1	1.4	.5	3.8	.5	3.1	.4	.4
/tr/	4	11.9	10.8	11.6	10.0	12.4	13.1	11.6	19.8
/tl/	3	10.4	14.2	4.4	6.0	.0	6.5	.0	10.6
/tw/	4	1.5	10.0	1.1	10.8	1.9	9.0	2.7	6.0
/ty/	3	2.4	5.8	.0	16.0	.0	13.4	1.3	1.7
/dr/	3	2.8	5.4	6.6	4.1	6.6	5.7	5.9	12.0
/dl/	2	2.5	7.1	2.2	2.5	.0	2.9	.0	6.4
/dw/	3	.0	5.0	1.7	4.5	.2	3.9	.4	3.6
/dy/	2	4.0	2.9	.0	6.5	.0	5.9	1.0	1.1
/kr/	4	2.4	3.7	8.8	3.5	8.8	4.6	10.8	7.5
/kl/	5	7.1	4.9	6.1	2.1	7.3	2.3	9.2	4.1
/kw/	4	7.6	3.5	5.0	3.8	4.5	3.2	5.0	2.3
/ky/	5	2.4	2.0	2.2	5.6	.7	4.7	.9	4.1
/gr/	3	6.2	1.5	12.2	1.4	9.1	1.6	10.2	1.9
/gl/	4	1.4	2.0	1.7	.9	2.6	.8	5.6	1.1
/gw/	3	.1	1.4	.0	1.6	.0	1.1	.4	.6
/gy/	4	.5	.8	1.7	2.3	.0	1.7	.0	.2
Total		99.9	99.9	100.1	99.9	100.0	99.9	99.8	99.9

TABLE 3: Correlations between observed cluster frequencies and (a) those predicted from phoneme frequencies within the same corpus, and (b) number of distinctive features separating phonemes within a cluster. df = N-3 = 21 (see McNemar, 1969).

Corpus	Correlation with frequencies predicted by chance	Correlation with phonemic difference
Denes (1965)	.40	.34
Hultzén, Allen, and Miron (1964)	-.19	.45*
Roberts (1965)	.03	.52^{+}
Trnka (1966)	.45*	.46*
Mean of corpora	.20	.47*

*p<.05, two-tailed
^{+}p<.01, two-tailed

performed. Notice that for the American corpora the observed frequencies correlated more highly with the phonemic difference scores than did the control, or chance-factor, frequency estimates. (This is all the more impressive since, owing to the large number of ties in phonemic difference, maximum correlation coefficients calculated here can be only .90.) The phonemic difference correlations for the Hultzén, Allen, and Miron (1964) corpus and the Roberts (1965) corpus were significantly greater than the control correlations (t = 2.51, p<.025 and t = 2.33, p<.05, respectively; McNemar, 1969:157-158). It is interesting to note that these are the two corpora based on American English pronunciation, whereas the other two are based on British English.

Is this principle general and distributed across the various phonetic features? Or is it, as Carroll (1958) suggests, primarily a function of one feature?--voicing. The answer can be seen in Table 4. The percent occurrence of all clusters that do not share each of the six features is compared against chance, calculated by dividing the number of clusters involved by the total number of clusters (24). Phonemic difference along all but one feature, consonantal, fits into the general scheme, providing equal to or greater than chance prediction. Consider the features in more detail. Vocalic and consonantal features can be yoked since all clusters differ on one or other (but not both) of the features; the first separates /l,r/, the second /w,y/, from the stop consonants. The vocalic feature predicts cluster frequencies very nicely, but the consonantal feature does not. Notice that the observed average of these two features is exactly 50 percent, or chance. Following Carroll (1958), one can eliminate these two features, yoked together, since they do not provide greater-than-chance prediction. The continuant feature can also be dismissed since members of all 24 clusters differ along this dimension. Only three features remain: grave, diffuse, and voiced. Each of these three features appears to contribute nearly equally to the phonemic difference effect, providing from 11 to 18 percent better-than-chance prediction. Moreover, voicing is not the most potent feature, as Carroll suggested it might be.

TABLE 4: Analysis of the effect of phonemic difference for each of the six distinctive features in predicting cluster frequency. Data base is mean of four corpora.

Distinctive feature	Observed percent of clusters with this feature not shared		Chance	Clusters involved
Vocalic	86.3	100.0	50.0	stop + /l,r/
Consonantal	13.7		50.0	stop + /w,y/
Grave	69.9		58.3	/pl, pr, py, bl, br, by, tw, dw, kr, kl, ky, gr, gl, gy/
Diffuse	59.2		41.7	/pr, br, tr, dr, kl, kw, ky, gl, gw, gy/
Voiced	66.1		50.0	/p,t,k/ + liquid /p,t,k/ + semivowel
Continuant	100.0		100.0	all

If the phonemic-difference principle is tenable as a predictor of cluster frequency, one might expect that any phonological change within a cluster should be in the direction of increasing phonemic difference, or perhaps in the elimination of the cluster altogether. In American English an increase can be seen in the affrication of alveolar + /r/ clusters; /tr/ and /dr/ clusters tend to go to /tʃr/ and /dʒr/, as in TRY and DRY. Affrication, in effect, adds the additional contrast of strident-nonstrident to members of both clusters. According to the mean of the four corpora, these two clusters are the second and tenth most frequent of the 24, and the addition of the strident contrast increases their phonemic difference to 5 and 4, respectively. In American English the stops /t/ and /d/ are involved in cluster elimination as well. Often these are simplified to a single consonant: /ty/ and /dy/ go to /t/ and /d/ in TUBE and DUTY. Notice that the Hultzén et al. (1964) corpus and the Roberts (1965) corpus lack any /ty/ and /dy/ clusters, whereas the Denes (1965) and Trnka (1966) corpora contain a number of them, which illustrates one difference between American and British English.

In conclusion, a revised version of the phonemic-difference principle postulated by Saporta (1955) serves to predict cluster frequency. Those clusters investigated here are stop-liquid and stop-semivowel combinations, which Saporta did not consider. As Carroll (1958) suggested, this principle serves to predict cluster frequency better than chance, particularly for the American English corpora. The effect is distributed across the different distinctive features and is particularly strong for grave, diffuse, and voicing features.

REFERENCES

Altmann,G. (1969) Differences between phonemes. Phonetica 19, 118-132.

Card, L. E. and A. R. Eckler. (1975) A survey of letter-frequencies. Word Ways: Journal of Recreational Linguistics 8, 81-85.

Carroll, J. B. (1958) The assessment of phoneme cluster frequency. *Language* *34*, 267-278.

Cherry, E. C., M. Halle, and R. Jakobson. (1953). Toward the logical description of languages in their phonemic aspect. *Language* *29*, 34-46.

Chomsky, N. and M. Halle. (1968) *The Sound Pattern of English*. (New York: Harper & Row).

Delattre, P. (1966) *Comparing the Phonetic Features of English, French, German, and Spanish*. (Heidelberg: Julius Groos Verlag).

Denes, P. B. (1965) On the statistics of spoken English. *J. Acoust. Soc. Amer.* *35*, 892-904.

Devine, A. M. (1971) Phoneme or cluster: A critical review. *Phonetica* *24*, 65-85.

Gerber, S. E. and S. Vertin. (1969) Comparative frequency counts of English phonemes. *Phonetica* *19*, 133-141.

Halle, M. (1964) On the bases of phonology. In *The Structure of Language*, ed. by J. A. Fodor and J. J. Katz. (Englewood Cliffs, N. J.: Prentice-Hall), pp. 324-333.

Hayden, R. E. (1950) The relative frequency of phonemes in general American English. *Word* *6*, 217-223.

Hofmann, T. R. (1967) Initial clusters in English. *Quarterly Progress Report* (Research Laboratory of Electronics, MIT) *84*, 263-274.

Hultzén, L., J. Allen, and M. Miron. (1964) *Tables of Transitional Frequencies of English Phonemes*. (Urbana: University of Illinois Press).

Jakobson, R., C. G. M. Fant, and M. Halle. (1951) *Preliminaries to Speech Analysis*. (Cambridge, Mass.: MIT Press, 1963).

Keller, K. C. and S. Saporta. (1957) The frequency of consonant clusters in Chontal. *Intl. J. Amer. Ling.* *23*, 28-38.

McNemar, Q. (1969) *Psychological Statistics*, 4th ed. (New York: John Wiley & Sons).

Menyuk, P. (1972) Clusters as single underlying consonants: Evidence from children's productions. In *Proceedings of the Seventh International Congress of Phonetic Sciences*, ed. by A. Rigault and R. Charbonneau. (The Hague: Mouton), pp. 1161-1165.

Menyuk, P. and D. Klatt. (1968) Child's production of initial consonant clusters. *Quarterly Progress Report* (Research Laboratory of Electronics, MIT) *81*, 205-213.

Roberts, A. H. (1965) *A Statistical Analysis of American English*. (The Hague: Mouton).

Saporta, S. (1955) Frequency of consonant clusters. *Language* *31*, 25-30.

Tobias, J. V. (1959) Relative occurrence of phonemes in American English. *J. Acoust. Soc. Amer.* *31*, 631.

Trnka, B. (1966) *A Phonological Analysis of Present-Day Standard English*. (University: University of Alabama Press).

Wang, W. S-Y. (1959) Transition and release as perceptual cues for final plosives. *J. Speech Hearing Res.* *2*, 66-73.

Wang, W. S-Y. and J. Crawford. (1960) Frequency studies of English consonants. *Lang. Speech* *3*, 131-139.

Wickelgren, W. A. (1966) Distinctive features and errors in short-term memory for English consonants. *J. Acoust. Soc. Amer.* *39*, 388-398.

Zipf, G. K. (1935) *The Psycho-biology of Language*. (Cambridge, Mass.: MIT Press, 1965).

Zipf, G. K. (1949) *Human Behavior and the Principle of Least Effort*. (Cambridge, Mass.: Addison-Wesley).