DOCUMENT RESUME

ED 037 704 AL 002 316

AUTHOR Bailey, Richard W.
TITLE Statistics and the Sounds of Poetry.
PUB DATE Mar 70
NOTE 42p.

EDRS PRICE EDRS Price MF-$0.25 HC Not Available from EDRS.
DESCRIPTORS Computational Linguistics, *Information Theory, *Literary Analysis, Literary Criticism, Literary Styles, *Mathematical Models, Phonemes, *Poetry, *Statistical Analysis

ABSTRACT

The author suggests that mathematical formulas could provide some direct and easily understandable frameworks for analogies in literary criticism. Most studies of textual problems, he points out, either have failed to use full mathematical models or have been reckless with the inherent limits of these techniques. Enumeration of linguistic traits is very common to literary analysis, and discussions of genre and text style based on linguistic data can be very informative. There are, however, serious dangers to be considered in the choice and size of text samples, and oversimplification of qualitative features should be avoided. He suggests that the most successful studies of this type involve a consideration of syllable and word structure rather than frequency of a literary device. In the text of this paper, the author examines various treatises on language usage in literary works, and comments on their value. His discussion of the applicability of mathematical formulas to literary analysis is quite technical. [Not available in hard copy due to marginal legibility of original document.] (FB)

STATISTICS AND THE SOUNDS OF POETRY

Richard W. Bailey

So distinguished a humanist as Richard J. Schoeck has recently called mathematics to the attention of literary critics as a fertile source for metaphors to illuminate their craft. Using examples from such fields as topology and vector analysis, he shows that much current criticism fumbles for words to express relations like those between author, work, and audience that might be clarified by an apt analogy from one of these disciplines. At the same time he cautions against the pitfalls that result from taking these metaphors too literally and mingling the exactness and precision that they would seem to offer with matters properly belonging to taste and judgment. "There is more than one literary parish," he warns, "that possesses its eager spirit who proclaims the new computer gospel with insufficient inquiry into its limitations. We need careful analysis, not hasty advertising" (Schoeck, 1968: 375). Some results from 'humanistic computing' clearly justify his skepticism; with the exception of two or three cases of disputed authorship concerning texts of minor literary importance, most studies of textual problems either have failed to exploit the full power of mathematical models or have run roughshod over the inherent limits of these techniques.

While critics like Schoeck tentatively explore the possible impact of mathematics on their work, linguists have been bolder in exploiting mathematical models, and the use of quasi-algebraic notation is now commonplace in accounts of language system. Statistical methods are also acknowledged to have application in descriptions of language use. William Labov, among others, has emphasized that the stigmata of social dialects are seldom so much a matter of all-or-nothing as the Biblical shibboleth; the various dialects of English differ much less in their underlying system of rules than the prophets of 'bidialectalism' and other forms of linguistic engineering would readily acknowledge. Linguistic differentiation -- regional, social, literary, and so on -- is almost always a question of the typical uses of a commonly shared system. A full treatment of the varieties of language used in a community must resort to a statistical account of favored paths through a network of linguistic rules.

Though scorned by many (for example, Ullmann, 1964: 118-21), the enumeration of linguistic traits is actually much less foreign to the practice of literary critics than essays like Schoeck's would seem to imply. Harry Levin comments that "we need make no word-count to be sure that [Hemingway's] literary vocabulary, with foreign and technical exceptions, consists of relatively few and short words" (1951: 596). In so saying, he at least confronts the kind of question that statistical description of a text might pose for itself, but literary critics seldom verify such assertions by statistical means. With a few notable exceptions (e.g., Leaska 1970), the concern that critics show for particulars is incompatible with the generalizing power of statistics. In bringing the insights of the Prague School formalists to the attention of American critics, René Wellek in his contribution to Theory of Literature drew attention to various interesting questions concerning genre and text style that might be illuminated by such means. Nevertheless, this influential book drew critics to consider literary minutiae and their function in the texture of a literary work, a tradition that emphasized the techniques of close reading initiated by explication de texte rather than the broader generalization involved in studying literary types in the context of aesthetic uses of language. Even as the influence of the 'new criticism' espoused by Theory of Literature has begun to wane, critics still denigrate the value of so useful a term as "period style" in a continuing concern for the idiosyncratic and personal in a literary work (see Chatman, 1966).

A full scale theory for discussing matters of style has recently been put forward by Lubomír Doležel in a paper entitled "A Framework for the Statistical Analysis of Style." In this essay, Doležel attempts to acknowledge the competing influences on a text of the author's personality, the generic constraints on his choices enforced by custom and tradition, and the shaping inherent in the linguistic medium in which he writes. The realization of this scheme requires a thorough integration of linguistics, cultural history, literary judgment, and statistics. Here in fact is an explicit and mathematical metaphor for the forces impinging on a literary work; Doležel, perhaps more thoroughly than Schoeck would find comfortable, has laid out the tasks typically fused in the broadest literary scholar in a program that may at first seem more palatable to the mathematician than to the belle-lettristic critic. With such a variety of potential influences on the work to consider, the narrowly trained academic may shy away from the standards of explicitness that Doležel's framework calls for or take refuge in the pervasive mathematical ludditism typical of literary men.

As I have argued elsewhere (Bailey, 1969), some of the questions of greatest concern to critics are amenable to mathematical treatment. Yet work of this kind is historically troubled by literary fatuity or statistical ineptness; those who undertake such research must confront two audiences and only seldom have they been able to satisfy both. Spurious exactness is as dangerous as the solipsism of hasty impressionism. A remark by E. E. Stoll, made a generation ago, might well serve as a rubric for all studies of this kind: "Error, which in criticism doth so easily beset us, is, when in the guise of science and armed with statistics, particularly insidious and dangerous. It seems to, but does not, put other error to flight: it is therefore in special need of detection" (Stoll, 1940: 390).

In attempting to carry out the tasks set by such theorists as Doležel and Labov, one soon finds that even simple problems raise vexing questions about samples and sample sizes, about the treatment of qualitative features of texts by statistical means, and about the choice of appropriate statistical tests to employ in evaluating the linguistic attributes extracted from the text. All of these issues played a part in the study to be discussed in this paper, a project that emerged from a seminar in statistical stylistics held at the State University of New York at Buffalo in the summer of 1969.[1] No startling [illegible]

[1] Participants in the seminar were Miss Heide Marie Miller, Mr. Fredrick L. Eyer, and Professor John M. Coetzee. Mr. Herbert B. Sanford III provided invaluable programming assistance in the early stages of the work, and advice on statistical methodology was freely given by Professor C. B. Bell. My interest in the field of statistical stylistics resulted

results of great literary or linguistic consequence will emerge in this essay, but it is my hope that the strategy outlined here can be usefully applied to questions of greater moment.

Our work began in an examination of the relation between statistical prominence and perceptual significance. "The analyst may forget," we are reminded by Theory of Literature, "that artistic effect and emphasis are not identical with the mere frequency of a device" (Wellek and Warren, 1956: 171). In examining this issue, we turned to a study of foregrounding of segmental phonemes in highly orchestrated texts. To what extent could this phenomenon be attributed to the numerical deployment of the sound resources of the language? Most handbook accounts of sound patterning in poems are unsatisfying because of their failure to specify the whole range of poetic sound effects characteristic of verse. A more careful taxonomy, like that outlined by David I. Masson, involves a consideration of syllable and word structure, line and syntactic patterning. In the belief that we could cut away these apparently unnecessary entities, we restricted our study to the stream of sounds in a text, hoping that significant patterns would emerge without acknowledging any other elements of linguistic or literary form. In doing so, we were encouraged by Jiří Levý's claim -- derived from studies in several languages -- that "verse makes more use of the

from the stimulating teaching of Professor Lubomír Doležel of the University of Toronto. Any defects of design or execution in this study are solely the responsibility of the author.

typical sounds of a given language and suppresses the rare ones" (Levý, 1967: 99). Even more heady claims concerning the behavior of sound segments tempted us to pursue the work; Marcello Boldrini, for example, asserts that "the use of a certain speech-sound is constant, or slightly variable in one author, but shows significant differences between author and author. This conclusively proves the originality in the use of speech-sounds on the part of the poets" (Boldrini, 1948: 64). Both Levý and Boldrini appeal to extensive experiments to support their views; both of them, our study suggests, were wrong. The view that phonemic foregrounding is a matter of the frequency distribution of sound segments cannot stand.

In casting about for a statistical technique that might yield insight into the problem, we found that information theory seemed to hold the greatest promise. Originally applied to problems in thermodynamics, information theory has had widespread application in designing communications systems (see Jackson, 1953, and Cherry, 1956); it is particularly attractive as a mathematical metaphor for criticism since it provides the means to define the patterning hidden in a great variety of complex and apparently random phenomena, the figure concealed in the aesthetic carpet. As a simile for the hidden structure in a chaotic set of events, information theory has already found use in Thomas Pynchon's novel, The Crying of Lot 49, and in several attempts to account for artistic organization including Abraham Moles' treatise, Information Theory and Esthetic Perception. Furthermore, studies of acknowledged worth have profited from the mathematics of this field: for example, A. M. Kondratov's work on rhythmical patterning in prose and poetry, and studies of vocabulary richness and grammatical organization made by Henry Kučera, Robert S. Wachal, and others.

The texts chosen for our study were transcribed in Trager-Smith phonemic notation, a system chosen mainly to allow comparisons with A. Hood Roberts' massive compilation of phonemic behavior, A Statistical Linguistic Analysis of American English. Vachel Lindsay's heavily alliterative poem, "The Congo," was transcribed in a form consonant with the dialect represented in Roberts' work, and two texts already available in phonemic transcription were included to provide further comparisons: Dylan Thomas' assonantal poem, "Fern Hill," and a prose passage often used in dialect analysis, "Grip the Rat."[2] These works, coupled with Roberts' materials, permit an examination of Jan Mukařovský's belief that "the standard language is the background against which is reflected the esthetically intentional distortion of the linguistic components of the work" (1933: 18).

Levý's claim that poets eschew the rare sounds of a language was derived from work by Doležel on grapheme distribution in Czech (Doležel 1963). The poem that Doležel examined -- Kundera's "Monology" -- does support Levý's assertion, but further studies on Czech phonemes by Ludviková and Kraus present a rather different picture. Their

[2] The transcription of "Fern Hill" was taken, with minor modifications, from Loesch, 278-83; the transcription of "Grip the Rat" from Francis, 159-60.

results, derived from the application of information theory, are reproduced in Table 1. The entropy values (H_1) shown there are calculated by the following formula in which the sample probabilities ($\hat{p}_i$) are taken to represent the probabilities for the population:

$$H_1 = - \sum_{1}^{n} p_i \log_2 p_i .$$

A low value for H_1, first-order entropy, reflects the tendency of some phonemes to occur with markedly greater frequency than others. In the case in which all the events measured occur with equal probability, H_1 takes on a maximum value, the diadic logarithm of n:

$$H_{max} = \log_2 n .$$

In the case of our English texts, a phonemic alphabet of thirty-two symbols was used. If English sounds all occurred with equal frequency -- as of course they do not -- the entropy value for this set of events would be $\log_2 32$ or 5.0. On the other hand, as the frequency of particular phonemes increases at the expense of others, a successively lower value for H_1 will result. To facilitate the comparison of texts with alphabets of varying sizes, it is also useful to calculate the relative entropy,

$$H_{rel} = \frac{H_1}{H_{max}},$$

as well as its complement, the 'redundancy':

$$R = 1.0 - H_{rel} .$$

Further details concerning both the mathematical and the linguistic grounding for these calculations can be found in Gleason's *Introduction to Descriptive Linguistics* and in E. V. Paducheva's "Information Theory and the Study of Language."

Had Levý been able to examine the results obtained by Ludviková and Kraus, he might have modified his claim concerning the behavior of sound segments in poems. For extremely different text styles -- drama and technical prose -- correspondingly different entropy values emerged. Yet poetry seems to be quite unremarkable in its deployment of sounds. Doležel's calculation of a low H_1 value for grapheme distribution in poetry (4.5722) appears to be anomalous in light of the work of Ludviková and Kraus, though a similar study by Štukovský does suggest that graphic and phonemic segments vary according to genre. Nevertheless, the application of entropy measures does not seem to be very productive in contrasting sound distribution in poems with that in other styles.

An application of these measures to three Rumanian poems has recently been carried out by Solomon Marcus.[3] Though the poems are quite short, the relation of the entropy values calculated from them would seem to confirm an aesthetic judgment, for the low value in Table 2 for "La Mijloc de Codru" reflects the piling up of particular

[3] Most of the slight differences between the values shown in Table 2 and those published by Marcus are owing to misprints in his essay; I gratefully acknowledge Dr. Marcus' correspondence concerning these matters.

sounds to produce a light, popular effect. The greater sonority of the third poem, "Se Bate Miezul Nopţii," contributes to a meditative and philosophical tone in which phonemic foregrounding plays a less significant role, a trait clearly represented in the higher H_1 value.

In addition to these values for entropy and redundancy, Marcus introduces an important measure of central tendency in qualitative variables, the repeat rate or 'informational energy.' As Table 2 shows, this value (E) varies inversely with the entropy, and it evidently follows that the text with the greatest repetition of phonemes will have the highest E characteristic. Octav Onicescu has shown that this easily calculated value,

$$E = \sum_{1}^{n} p_i^{\,2} ,$$

is, like entropy, a useful indicator of the structuring of events in a pattern (see also Herdan, 1966, 271-73).

The results of Marcus' study would seem to indicate that information theory provides a reliable correlate of the aesthetic effects that readers recognize in the sound patterning in these three poems. Nevertheless, the task still remains of testing this intuitive indication of significance against the possibility that differences in H_1 result from purely chance causes. Although the variables we are considering are not independent, a testing procedure devised by G. P. Basharin provides the mathematical grounds for exploring the approximate significance of such differences. Both the estimate of the population entropy (H_{est}) and the variance (s^2) can be calculated as follows, with $\underline{N}$ equal to the sample size in phonemes and $\underline{n}$ equal to the size of the alphabet of symbols:

$$H_{est} = H - \frac{n - 1}{2 N} \log_2 \underline{e} + O\left(\frac{1}{N^2}\right)$$

$$s^2 = \frac{1}{N}\left[\sum_{1}^{n} p_i \log_2{}^2 p_i - H^2\right] + O\left(\frac{1}{N^2}\right)$$

With the help of these formulae, it is possible to establish confidence bounds for the dispersion of $\underline{H_1}$ values. Taking a limit of .05 (1.96 $\underline{s}$), we can estimate the upper and lower values within which 95 of every 100 samples with distributions like those in the poems will vary on the basis of chance. The results of these calculations are given in Table 3. Since the observed value for $\underline{H_1}$ for "La Mijloc de Codru" falls below the lower bounds for the other two poems, we can take this value to reflect a significant difference in the sound patterning of this poem. The observed values for "Somnoroase Păsărele" and "Se Bate Miezul Nopții" both fall within the same range, and no significance can be ascribed to the difference between them at the .05 level of discrimination.

In a parallel study of the difference between prose and poetry in Tamil, Gift Siromoney found a significant contrast of the $\underline{H_1}$ values for graphemes in large samples, though he did suggest that other testing procedures than the one proposed by Basharin provide a more powerful discriminator of the two types. In a subsequent study of $\underline{H_1}$

in Telugu prose, P. Balasubrahmanyam and Siromoney are much more pessimistic about the utility of this measure as an indication of stylistic differences. A recomputation of their published results (given in Tables 4 and 5) suggests a significant difference in grapheme deployment between novels and short stories, while the results for the other varieties of prose fall into the same range. Since the two linguists say little about the aesthetic properties of the various Telugu styles, it remains to be seen whether this measure has any correlation with what readers perceive about sound patterning in these texts.

Our own work with these measures in examining phonemic distributions in English texts presents an even more mixed picture as Tables 6 and 7 show. The observed $\underline{H_1}$ values for the three texts studied are all higher than the 4.4947 entropy figure calculated by Roberts for a huge sample of American English. Splitting "The Congo" into three segments of approximately equal length, we note that $\underline{H_1}$ is fairly stable for this text. But Levý's hypothesis that verse will, in general, have a lower $\underline{H_1}$ than other styles is clearly rejected by these data, for "The Congo" has a higher value than either "Fern Hill" or "Grip the Rat." "The Congo" does differ significantly from the other two texts, though not in the direction that Levý would lead us to expect. Furthermore, the vivid perceptual contrast between the highly orchestrated "Fern Hill" and the thoroughly mundane "Grip the Rat" is not at all clearly reflected in the results of these calculations.

A decade ago, Roman Jakobson made an eloquent plea for cooperation between information theory and poetics, and the idea that significant results would emerge from such an alliance has been repeated several times since. The study of sound patterning in poems, however, should apparently turn elsewhere -- either to other statistical procedures or to a more complex hypothesis in which suprasegmentals, syllable boundaries, or syntactic structure (if not all three) are taken into account. The history of scholarship in this area reflects all too clearly Stoll's warning against error "in the guise of science and armed with statistics." Boldrini's claim that sound distribution is a possible authorship indicator is clearly refuted by the results we have examined; Levý's view that phoneme frequencies in poems contrast with other text styles does not hold for English and Czech and finds only partial verification in Tamil and Rumanian. Entropy measurement proves to be a poor measure of the aesthetic significance of poetic sound patterning, however successful its other applications to linguistics or communication science.

Another strategy for approaching this problem without going beyond the level of segmental phonemes was suggested by H. Spang-Hanssen's paper, "The Study of Gaps between Repetitions." In highly alliterative verse, we surmised, like phonemes will tend to cluster closer together than in a prose text. In other words, while the overall frequency of sounds in an orchestrated text might not be markedly different from what can be expected in prose, the gap distance from one instance of a given phoneme to the next might well be shorter in the poem. By calculating the size of the intervals between phonemes (from 0, adjoining instances, to a distance $\geqq$ 100), we hoped to find a clearer measure of sound patterning than emerged from the application of entropy measurement.

To clarify the matter still further, two random sets of phonemes were generated in which each phoneme in the inventory was selected independently at a chance of 1/32. The entropy values for the phoneme distribution in these artificial texts approached the maximum $\underline{H}_1$ value of 5.0, as expected, since in both random texts of 2,000 phonemes, sounds occur with approximate equiprobability. When entropy measures were applied to the distribution of gap intervals in the texts, no pattern of the expected kind emerged (as Table 8 shows). All the $\underline{H}_1$ values for this feature in the English texts cluster close together, though the relationship of the three texts was for the first time in the order we expected. The alliterative text, "The Congo," did reveal a greater exploitation of some gap intervals as reflected in its relatively low $\underline{H_1(gaps)}$ figure, 5.5948, while "Fern Hill" and "Grip the Rat" came successively closer to the even distribution of gaps found in the random texts. Wide-ranging values for the three subsets of "The Congo," however, suggest that this feature is not particularly stable in the poem. Yet the low value for the middle segment of the poem does reflect a significant feature of the text, the concentration of a repetitive refrain in this section -- a trait that piles up a succession of like intervals and reduces the $\underline{H}_1$ value for gaps. When these refrains are removed from the text, however, the $\underline{H_1(gaps)}$ value approaches the figure achieved by the other two texts, despite the clearly alliterative nature of the remaining lines of the poem.

A further test of the behavior of gaps in the three texts was carried out with the help of the Kolmogorov-Smirnov two-sample test (see Siegel, 1956: 127-36). Once again the condition of independence is not met, but the results of this test can at least be taken as indicative of the relations between the distributions. The greatest difference between the cumulative frequency distributions of the gap intervals, $\underline{D}$, is given in Table 9 for each pair of texts. If the observed value of $\underline{D}$ exceeds the value calculated by the following formula (where $\underline{n}$ = the number of gaps in the text), the difference may be taken as significant at a 95% level of confidence:

$$1.36 \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

These calculated values appear in parenthesis in Table 9. Only the randomly constructed texts contrast significantly with the English texts in their distribution of gap intervals, though "The Congo" and "Fern Hill" are slightly different in their deployment of gaps. Nevertheless, this study of the sequential ordering of phonemes is no more revealing of the aesthetic differences between the texts than the probabilistic data subsumed in the entropy calculations turned out to be.

In a third attempt to characterize the behavior of phonemic segments in these texts, we examined the rank-ordering of the sounds. Could the relatively high rank of /b/ in "The Congo" (the fourteenth most frequent phoneme) be considered significant when compared to its position of twenty-first in Roberts' list, twenty-second in "Fern

Hill," and twenty-third in "Grip the Rat"? A measure of rank-correlation, the Spearman rho test, yielded the results given in Table 11 (see Siegel, 1956: 202-13). All of the coefficients reveal a strong relationship between the samples. "Grip the Rat" most resembles Roberts' ordering of phonemes, closely followed by "The Congo" and "Fern Hill." Of particular significance is the very high correlation between the subsets of "The Congo," a result that suggests that phonemic foregrounding exploits the same phonemic segments throughout the poem. Nevertheless, the rho values are much the same in comparisons between Roberts' ranking and the ranks in the texts. The striking differences between "Grip the Rat" and "The Congo" are still not vividly apparent from these calculations.

As a result of our study of entropy, gap distribution, and rank-ordering, we are forced to conclude that the foregrounding of segmental phonemes cannot be specified by an examination of phonemic distribution alone. Stress placement, syntactic structure, and thematic organization must be acknowledged in the formation of a hypothesis describing sound patterns in texts. A conclusion of this sort might easily have been anticipated at the outset of the study, of course, but the careful examination of the phenomenon in the simplest terms seemed plausible enough to make it worth testing. Furthermore, the statistical techniques illustrated here will have value in testing more complicated hypotheses, for they illustrate the procedures that must be involved in a thorough-going treatment of language use.

While the strategies discussed so far have not yielded the results anticipated by our intuitive hypothesis, certain differences in the phoneme distributions in the three texts still call for explanation. As an interim measure for discussing these differences, we propose a 'prominence index' (*P*) to account for significant variations in rank-order in the author's deployment of sound segments in texts. This index is based primarily on the rank difference between a given phoneme in the text and the rank calculated for that phoneme in a representative body of American English by Roberts. The phoneme /ž/, for example, has a rank of twenty-three in Roberts' results and a rank of fourteen in "Grip the Rat," a rank difference of +9. The phoneme /a/, on the other hand, is eighth most frequent in Roberts' study, but twenty-first in "Grip the Rat," a rank difference of -13. When phonemes have the same frequency in a sample text, such ties are treated as they are in the Spearman rho test. In "The Congo," for example, /č/ and /θ/ occur in the same number; in calculating *P*, the ranks are averaged to 29.5.

Positive rank-differences would seem to have greater perceptual value than negative ones. That is, readers will hardly be aware of the relative absence of a particular phoneme, but greater exploitation of a phoneme than might normally be expected will usually be noticed. Furthermore, the promotion of rare phonemes will have more impact on the sound orchestration of a text than the same rank difference for a common sound. Therefore, the perceptual index should take into account the normal expectation of a phoneme, as well as its rank difference in the text. From these considerations, the following formula has been devised for the prominence index. R_i denotes the rank in Roberts' count; $\underline{r_i}$ the rank in the text; $\underline{p_i}$ the probability for the phoneme given by Roberts (values reproduced here in the first column of Table 10). A factor of 100 has been introduced to avoid decimals.

$$P = 100\ (R_i - r_i)\ (1.0 - p_i)$$

The prominence indices for the phonemes in our three texts are given in Table 12.

What conclusions can be drawn from the values calculated for the prominence index? In "Grip the Rat," most of the index values are low, a result that reflects the commonplace nature of the text when set against the expectations of American English. Only two phonemes differ markedly in prominence from Roberts' count: /a/, with an index of -1240, and /ɔ/ with +1292. The choice of these two phonemes for particular prominence is no accident since, it will be recalled, this passage was designed to elicit dialect differences from its readers. Such a test passage would necessarily provide frequent opportunities for informants to demonstrate their characteristic use of low and back vowels since these phonemes are significant markers of regional differences in American speech. High index values for /æ/ and /o/ in this passage can be attributed to the same cause, and the realization of these four phonemic segments in speech will have considerable value for the dialectologist.

The relatively low rho correlations between "Fern Hill" and the other samples shown in Table 11 is easily explained by reference to the index values calculated for this text. Unlike "The Congo," "Grip the Rat," and Roberts' study, the transcription of "Fern Hill" used here is not based on the American English of the Upper Midwest but on Thomas' own reading of his poem. The low index value for /r/, -1214, reflects the postvocalic r-lessness typical of British speech, while a high value for /h/ -- +779 -- shows the convention of the "centering glide" that a Trager-Smith notation offers for characterizing this r-lessness. Other dialect differences between Thomas' speech and American English are also highlighted by the index values for low and back vowels, particularly /æ/ and /ɔ/. But not all the large index values are owing to dialect differences of this sort. Once this measure has been coordinated with thematic and syntactic structure, the high value for /ŋ/ can be traced to Thomas' preference for participial constructions ("lilting house," "spinning place") and for the thematic repetition of the word 'young'. Further aesthetic consequences may also emerge from the apparent preference for voiced segments: /z/ over /s/, /d/ over /t/, /g/ over /k/, /b/ over /p/, /ð/ over /θ/. Without integration into a full scale analysis of the poem, these facts are without consequence. But clearly they are a part of the aesthetic impact of the text and must be considered in an examination of the structuring of language for artistic ends.

Indications of phonemic foregrounding like those found in "Fern Hill" also emerge from an examination of the prominence indices for "The Congo." As a phonemic counterpart to the drum-beating intended to accompany performances of this poem, Lindsay de-emphasizes front vowels -- /i/, -454, and /e/, -191 -- for the more resonant back vowels -- /u/, +392, and /o/, +640. In addition, the exploitation of these back vowels is also reflected in the high index value for the so-called back glide in Trager-Smith phonemics: /w/, +668. A preference for voiced consonants, even more extreme than that noted for "Fern Hill," is also noticeable in the index values for "The Congo," with high ranks assigned to /b/, /g/, /z/, /ǰ/, and /ð/, and relatively low ones to /p/, /k/, /s/, /č/, and /θ/. Once again, it would be futile to infer sense from sound or to suggest that the effects shown by the index have a value independent of the semantic and syntactic organization of the poem. Yet, as already noted in Table 11, the promotion in rank shown by the index values is highly consistent in the poem, and the phonemic structuring reflected in the P values vividly furthers Lindsay's aesthetic ends.

Our study has revealed that the distribution of phonemes in languages operates within highly restrictive limits. Certain high frequency phonemes will occur about ten percent of the time, while less frequent phonemes occupy quite stable positions in a decreasing series. Table 13 shows the redundancy values for the various languages discussed in this paper; the rather small variation between languages suggests that the frequency series occupied by phonemes is relatively fixed, even in the deformation of normal language in aphasia. The approximate similarity of redundancy values thus suggests that a linguistic universal constrains human speech in the deployment of phonemes. But we have also found that the particular phonemes chosen to occupy a given frequency can vary considerably between texts. The prominence index proposed here reflects the promotion or demotion of particular phonemes in the series, a characteristic that proves to be related not only to dialect and language, but also to the aesthetic organization of a text.

The procedures used in this study are not a result of positivistic yearnings for exactness in all questions of judgment. Instead they point toward a rigor of method and definition. Too much of literary commentary verges toward precision without confronting what it means to be precise. This effort to relate statistical techniques and literary matters, I hope, points toward the boundary between problems of fact and questions of opinion. Perhaps the words of George Steiner -- a man particularly attuned to the amoral possibilities of a criticism oblivious to the potency of literary truth -- can serve as a fitting close to this study:. "The shapes of reality and of our imaginative grasp are exceedingly difficult to forsee. Nevertheless, the student of literature now has access to and responsibility towards a very rich terrain, intermediate between the arts and sciences, a terrain bordering equally on poetry, on sociology, on psychology, on logic, and even on mathematics" (Steiner, 1965: 86).

The University of Michigan

Table 1

Entropy Values (H_1) for Various Czech Styles

	Drama		Narrative Prose		Poetry		Newspaper Prose		Technical Prose	
	Hrubín	Topol	Fried	Hrubín	Hrubín	Florian	Večerní Praha	Rudé právo	Cookbook	Histology Text
phoneme	4.5006	4.5112	4.5618	4.5701	4.5684	4.5984	4.6351	4.6382	4.5980	4.6823
grapheme				4.5919			4.6985			4.7044

N = 186,641 phonemes for an alphabet of 39 symbols
N = 140,000 graphemes for an alphabet of 42 symbols

Table ~~[illegible]~~ 2

Entropy Values for Three Poems by Mihail Eminescu

	n	N	H_{max}	H_1	H_{rel}	R	E
La Mijloc de Codru	28	232	4.8074	4.0562	.8612	.1388	.06963
Somnoroase Păsărele	28	267	4.8074	4.2274	.8794	.1206	.06471
Se Bate Miezul Nopții	28	213	4.8074	4.2542	.8849	.1151	.06218

Table ~~III~~ 3

Significance Tests: Three Poems by Mihail Eminescu

	Hest	s^2	s	lower bound	upper bound
La Mijloc de Codru	4.0562	.0051	.0714	4.0003	4.2801
Somnoroase Pásărele	4.1544	.0038	.0617	4.1065	4.3423
Se Bate Miezul Nopţii	4.1627	.0042	.0648	4.1272	4.3811

Table 4

Entropy Values for Various Styles of Telugu Prose

	n	N	H_{max}	H_1	H_{rel}	R	E
Novels	53	26,488	5.7279	4.5965	.8025	.1975	.05952
Short Stories	53	3,995	5.7279	4.5343	.7916	.2084	.06176
Plays	53	13,021	5.7279	4.5757	.7988	.2012	.06153
Other Forms	53	6,501	5.7279	4.5570	.7956	.2044	.06521
All Prose	53	49,845	5.7279	4.5785	.7993	.2007	.06041

Table ~~3~~ 5

Significance Tests: Telugu Prose

	H_{est}	s^2	s	lower bound	upper bound
Novels	4.5951	.00007 ~~.000[illegible]~~	.0085	4.5798	4.6132
Short Stories	4.5249	$.000\overset{4}{\wedge}7$	.0217	4.4917	4.5768
Plays	4.5728	$.000\overset{1}{\wedge}5$	.0124	4.5513	4.6000
Other Forms	4.5512	.00034	.0185	4.5207	4.5933
All Prose	4.5777	.00004	.0064	4.5659	4.5911

Table ~~VI~~ 6

Entropy Values for Some English Texts

	n	N	H_{max}	H_1	H_{rel}	R	E
"Grip the Rat"	32	1754	5.0	4.5407	.9081	.0919	.05184
"Fern Hill"	32	1714	5.0	4.5034	.9007	.0993	.05288
"The Congo"	32	4192	5.0	4.5691	.9138	.0862	.05069
"The Congo" (11. 1-76)	32	1965	5.0	4.5628	.9126	.0874	.05106
"The Congo" (11. 77-152)	32	2227	5.0	4.5558	.9112	.0888	.05143
"The Congo" (11. 38-114)	32	2192	5.0	4.5845	.9169	.0831	.04989

Table 7

Significance Tests: English Texts

	E_{est}	s^2	s	lower bound	upper bound
"Grip the Rat"	4.5279	.0005	.0232	4.4953	4.5861
"Fern Hill"	4.4904	.0005	.0233	4.4578	4.5491
"The Congo"	4.5637	.0002	.0149	4.5399	4.5982
"The Congo" (ll. 1-76)	4.5514	.0005	.0218	4.5201	4.6055
"The Congo" (ll. 77-152)	4.5457	.0004	.0208	4.5150	4.5966
"The Congo" (ll. 38-114)	4.5743	.0004	.0203	4.5448	4.6242

Table 8

Distribution of Gaps between Phonemes

	n_{gaps}	E	H_1	H_{max}	H_{rel}	R
"Grip the Rat"	1723	.02630	5.7138	6.6439	.8600	.1400
"Fern Hill"	1683	.02648	5.6914	6.6439	.8566	.1434
The Congo"	4160	.03000	5.5948	6.6539	.8421	.1579
"The Congo" (11. 1-76)	1953	.02813	5.6487	6.6439	.8502	.1498
"The Congo" (11. 77-152)	2195	.03239	5.4871	6.6539	.8259	.1741
"The Congo" (11. 5438-114)	2160	.03074	5.5730	6.6439	.8388	.1612
Random Text I	1968	.01808	6.1087	6.6439	.9195	.0805
Random Text II	1968	.01763	6.1316	6.6439	.9229	.0771

Table 9

Observed Values of D for Gap Distribution

(figures in parenthesis show the value D must exceed at .05)

	"The Congo"	"Fern Hill"	"Grip the Rat"	"Random Text II
Random Text I	.1572	.1578	.1297	.0147
	(.0360)	(.0431)	(.0431)	(.0431)
Random Text II	.1577	.1573	.1317	
	(.0360)	(.0431)	(.0431)	
"Grip the Rat"	.0365	.0337		
	(.0385)	(.0471)		
"Fern Hill"	.0489			
	(.0385)			

Table ~~2~~10

Absolute and Relative Frequencies of Phonemes in English Samples

	Roberts' Count	"Grip the Rat"		"Fern Hill"		"The Congo"	
ə	.11819	205	.11688	195	.11377	462	.11021
i	.09292	99	.05644	126	.07351	225	.05367
t	.06945	117	.06670	64	.03734	170	.04055
y	.06772	119	.06784	131	.07643	264	.06298
r	.06582	96	.05473	42	.02450	226	.05391
n	.06288	110	.06271	119	.06943	244	.05821
e	.04735	73	.04162	48	.02800	177	.04222
a	.04634	31	.01767	84	.04901	109	.02600
w	.04519	110	.06271	87	.05076	353	.08421
s	.03905	74	.04219	55	.03209	113	.02696
l	.03308	57	.03250	79	.04609	233	.05558
d	.03041	73	.04162	88	.05134	172	.04103
h	.02628	71	.04048	89	.05193	80	.01908
m	.02613	42	.02395	51	.02975	144	.03435
k	.02449	48	.02737	25	.01459	107	.02552
ð	.02247	69	.03934	70	.04084	182	.04342
z	.02008	29	.01653	74	.04317	78	.01861
u	.01908	35	.01995	19	.01109	138	.03292

[Table ~~6~~ 10 (continued)]

v	.01879	16	.00912	22	.01284	55	.01312
f	.016$\overset{9}{\wedge}$8	37	.02109	28	.01634	53	.01264
b	.016~~9~~$\overset{2}{\wedge}$8	20	.01140	24	.01400	135	.03220
p	.01610	12	.00684	18	.01050	45	.01073
ð	.01540	53	.03022	50	.02917	99	.02362
o	.01537	42	.02395	8	.00467	109	.02600
ŋ	.00919	19	.01083	34	.01984	50	.01193
ɛ	.00868	17	.00969	21	.01225	57	.01360
š	.00717	7	.00399	6	.00350	15	.00358
ɔ	.00643	50	.02851	45	.02625	36	.00859
č	.00460	9	.00513	5	.00292	16	.00382
θ	.00420	7	.00399	6	.00350	16	.00382
ǰ	.00359	7	.00399	1	.00058	26	.00620
ž	.00030	0	.0	0	.0	3	.00072
Totals		1754	1.00000	1714	1.00000	4192	1.00000

Table 11

Rho Coefficients for Phonemic Ranking

	Roberts' Count	"The Congo" (11. 38-114)	"The Congo" (11. 77-152)	"The Congo" (11. 1-76)	"The Congo" (11. 1-152)	"Fern Hill"
"Grip the Rat"	.87483	.87462	.85125	.86736	.87306	.84079
"Fern Hill"	.83714	.75809	.75820	.78768	.77651	
"The Congo" (11. 1-152)	.87828	.99322	.99404	.97827		
"The Congo" (11. 1-76)	.85461	.98249	.96122			
"The Congo" (11. 77-152)	.87343	.98522				
"The Congo" (11. 38-114)	.87039					

Table 12

Prominence Indices for English Texts

	"Grip the Rat"	"Fern Hill"	"The Congo"
ə	0	0	0
i	-363	-91	-454
t	0	-838	-744
y	+187	+187	+93
r	-187	-1214	-93
n	+141	+188	+188
e	-238	-857	-191
a	-1240	0	-763
w	+430	+286	+668
s	+192	-288	-481
l	-193	+193	+580
d	+242	+483	+193
h	+195	+779	-779
m	-341	0	+195
k	-96	-585	-244
ð	+391	+489	+782
z	-490	+686	+392
u	-196	-687	+491
v	-687	-393	-393
f	+98	0	-393
b	-197	-93	+689
p	-492	-394	-394

Table 12 (continued)

æ	+886	+783	+394
o	+640	-295	+640
ŋ	+99	+595	0
g	+99	+198	+397
š	-298	-199	-397
ɔ	+1292	+1093	+99
č	+100	0	-450
θ	+150	+100	+50
ǰ	+100	0	+299
ž	0	0	0

Table 13

Redundancy Values in Several Languages

	R	Sample Size
English Texts (Bailey)	.0913	7,660 phonemes
English (Roberts, 1965)	.1011	15,465,010 phonemes
Modern Tamil Prose (Siromoney, 1963)	.1155	22,855 graphemes
Rumanian, average for Eminescu poems (Marcus, 1967)	.1248	712 phonemes
Czech (Ludviková and Kraus, 1966)	.1271	186,641 phonemes
Rumanian, normal speech (Voinescu *et al.*, 1967)	.1653	100,000 phonemes
Rumanian, aphasic speech (Voinescu *et al.*, 1967)	.1786	100,000 phonemes
Telugu Prose (Balasubrahmanyam and Siromoney, 1968)	.2007	49,345 graphemes

REFERENCES

Bailey, Richard W.

1969. "Statistics and Style: A Historical Survey," in Doležel and Bailey, 217-36.

Balasubrahmanyam, P., and Gift Siromoney.

1968. "A Note on Entropy of Telugu Prose," *Information and Control* 13,281-85.

Basharin, G. P.

1959. "On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables," tr. Newcomb Greenleaf, *Theory of Probability and Its Applications* 4,333-36.

Boldrini, Marcello.

1948. "Tendenze e individualità nella poesia italiana moderna sotto l'aspetto fonetico-statistico," *Contributi del Laboratorio di Statistica*, serie sesta (= Milano: Edizioni dell'Università Cattolica del Sacro Cuore, vol. 21), 16-65.

Chatman, Seymour.

1966. "On the Theory of Literary Style," *Linguistics* 27,13-25.

Cherry, Colin, ed.

1956. *Information Theory* (New York: Academic Press).

Doležel, Lubomír.

1963. "Předběžný odhad entropie a redundance psané češtiny," *Slovo a slovesnost* 3,165-75.

1969. "A Framework for the Statistical Analysis of Style," in Doležel and Bailey, 10-25.

———, and Richard W. Bailey.

1969. *Statistics and Style* (New York: American Elsevier).

Francis, W. Nelson.

1958. *The Structure of American English* (New York: Ronald Press).

Gleason, H. A., Jr.

1961. *An Introduction to Descriptive Linguistics* (New York: Holt, Rinehart and Winston), 373-90.

Herdan, Gustav.

1966. *The Advanced Theory of Language as Choice and Chance* (New York: Springer-Verlag).

Jackson, Willis, ed.

1953. *Communication Theory* (New York: Academic Press).

Jakobson, Roman.

1961. "Linguistics and Communication Theory," *Studies of Language in Its Mathematical Aspects* (= *Proceedings of Symposia in Applied Mathematics*) (Providence, R. I.: American Mathematical Society), 245-52.

Kondratov, A. M.

1963. "Information Theory and Poetics: The Entropy of Russian Speech Rhythm," in Doležel and Bailey, 113-21.

Kučera, Henry.

1968. "Some Quantitative Lexical Analyses of Russian, Czech and English," *American Contributions to the Sixth International Congress of Slavists* (The Hague: Mouton), 155-98.

Leaska, Mitchell A.

1970. *Virginia Woolf's Lighthouse: A Study in Critical Method.* (New York: Columbia University Press).

Levin, Harry.

1951. "Observations on the Style of Ernest Hemingway," *Kenyon Review* 13,581-609.

Levý, Jiří.

1967. "Mathematical Aspects of the Theory of Verse," in Doležel and Bailey, 95-112.

Loesch, Katharine Taylor.

1961. "Prosodic Patterns in the Poetry of Dylan Thomas." (Evanston, Ill.: unpublished Ph.D. dissertation presented to Northwestern University).

Ludvíková, Marie, and Jiří Kraus.

1966. "Kvantitativní vlastnosti soustavy českých fonémů," *Slovo a slovesnost* 4,334-44.

Marcus, Solomon.

1967. "Entropie et énergie poétique (avec application à trois poésies de Eminescu)," *Cahiers de Linguistique Théorique et Appliquée* 4,171-80.

Masson, David I.

1960. "Sound-Repetition Terms," *Poetics: International Conference on Work-in-Progress Devoted to Problems of Poetics* (The Hague: Mouton), 189-99.

Moles, Abraham.

1966. Information Theory and Esthetic Perception, tr. Joel E. Cohen (Urbana: University of Illinois Press).

Mukařovský, Jan.

1932. "Standard Language and Poetic Language," A Prague School Reader on Esthetics, Literary Structure, and Style, ed. and tr. Paul L. Garvin (Washington, D. C.: Georgetown University Press, 1964), 17-30.

Onicescu, Octav.

1966. "Énergie informationnelle," Comptes Rendus de l'Académie des Sciences (Paris), 263, ser. A, 841-42.

Paducheva, E. V.

1963. "Information Theory and the Study of Language," Exact Methods in Linguistic Research, tr. David G. Hays and Dolores V. Mohr (Berkeley and Los Angeles: University of California Press), 119-79.

Pynchon, Thomas.

1966. The Crying of Lot 49 (Philadelphia: J. B. Lippincot).

Roberts, A Hood.

1965. A Statistical Linguistic Analysis of American English (The Hague: Mouton).

Schoeck, Richard J.

1968. "Mathematics and the Languages of Literary Criticism," Journal of Aesthetics and Art Criticism 26,367-76.

Siegel, Sidney.

1956. *Nonparametric Statistics for the Behavioral Sciences* (New York: McGraw-Hill).

Siromoney, Gift.

1963. "Entropy of Tamil Prose," *Information and Control* 6,297-300.

Spang-Hanssen, H.

1956. "The Study of Gaps between Repetitions," *For Roman Jakobson*, ed. Morris Halle *et al.* (The Hague: Mouton), 492-502.

Steiner, George.

1965. "To Civilize Our Gentlemen," *Language and Silence* (Harmondsworth: Penguin Books, 1969).

Stoll, Elmer Edgar.

1940. "Poetic Alliteration," *Modern Language Notes* 55,388-90.

Štukovský, Róbert.

1964. "Letter Entropy in a Special Vocabulary," *Jazykovedný časopis* 15,27-34.

Ullmann, Stephen.

1964. *Language and Style* (New York: Barnes & Noble).

Voinescu, I., A. Fradis, and Lucreţia Mihăilescu.

1967. "The First Degree Entropy of Phonemes in Aphasics," *Revue Roumaine de Neurologie* 4,67-79.

Wachal, Robert S.

1966. "Linguistic Evidence, Statistical Inference, and Disputed Authorship" (Madison, Wis.: unpublished Ph.D. dissertation presented to the University of Wisconsin).

Wellek, René, and Austin Warren.

1956. *Theory of Literature* (New York: Harvest Books).