

SYNTACTIC PATTERNS IN A SAMPLE
OF TECHNICAL ENGLISH

The Importance of the Concept of Homogeneity

A fundamental assumption of statistical linguistics is that there are differences worthy of note in the frequency of various units in certain texts. At the same time, there are differences in frequencies which would not be considered important. The question is, how is an "important" difference to be determined?

The resolution of this problem has been made more important by the increasing popularity of statistical approaches to questions of style and authorship. Definitions of style from this point of view are based on notions of distinctiveness and consistency in literary performance. While distinctiveness appears to be the more important component of style, it is recognized that some consistency is necessary to lend significance to whatever feature might be distinctive.

The Determination of Homogeneity

For this discussion we define homogeneity as the similarity of parts of the whole with respect to certain features. For some features it may be perfectly clear, even without counting, that parts of a text or texts from a genre are not alike. This seems more likely to occur for some features and for some genres than for others, for

example, syntactic or phonological constructions in poetry, as opposed to parts of speech in technical writing.

Few would be satisfied to rely solely on subjective impression for the estimation of the similarity of text samples. For statistical linguists the decision to count is the foundation of their science. For literary scholars the decision to count stems from a desire to give quantitative verification of existing theories and interpretations, and to gain greater insight into the structure of literary works for the purpose of proposing new theories and interpretations. Both groups are faced with the problem of evaluating the results of the counting.

The Nature of Statistical Tests

The techniques of statistical description are, of course, uniquely suited to the statement of the raw, uninterpreted results. Measures of location such as means, modes, medians are commonly used for this purpose.

In examining the raw results it may be clear at once that there is a meaningful difference among the counts or scores. If samples of 100 sentences were taken at random from each of two texts, and the mean lengths for the two samples were 20 words and 40 words, no one would hesitate to conclude that one text revealed a "significantly" greater sentence length than the other. But if the figures were closer, say 27 and 33, more exact methods are needed.

It is a law of nature that a sample taken from a population will not always yield exactly the statistics of the population, that on occasion even a large discrepancy will be found. The extent to which sample values may be expected to vary from population values through chance alone is a subject of mathematical statistics, as is the extent to which two or more sample values from the same population will differ.

Language Statistics and Homogeneity

There is considerable data that demonstrates overall similarities in the frequencies of various units between samples from the same writer, from different writers, and even from different languages.¹ The problem for statistical linguistics and stylistics is the ordering of degrees of similarity into groups according to some notion of homogeneity. If the sample values differ no more than could reasonably be attributed to chance, we see no reason why the populations from which the samples were taken could not be called one homogeneous population.

Whether text samples pass a statistical test for homogeneity depends on the nature of the text, the chosen

¹See, for example, Herdan, The Advanced Theory of Language as Choice and Chance, pp. 17-27, and M. Renský, "The Noun-Verb Quotient in English and Czech," Philologia Pragensia, VIII (1965), pp. 289-302.

significance level, and the power of the test as determined by characteristics of the test itself in conjunction with the size of the sample. It is possible to imagine a perfectly uniform text, for example, one composed of nothing more than repetitions of the same identical sentence. In this case, a statistical test will reveal this homogeneity for any significance level or sample size. For real texts, though, the selection of the s.l. and s.s. poses a problem of practical and theoretical interest. The danger is that an investigator will be tempted to make a flat statement concerning the homogeneity of a feature for a text or a genre, when a slight change in s.l. or s.s. could have led to a reversal of that finding. Homogeneity, then, as a product of statistical hypothesis testing, should not be regarded as a function of the text alone, but rather as a function of the text and the significance level and power associated with the test and the sample size. If the samples represent different populations even if different only in some minimal way, it is only a question of increasing sufficiently the sample size to cause the hypothesis of homogeneity to be rejected.

In discussing the size of samples to be taken, Herdan states that "for statistical investigations in general, it is usually a question of how small the sample should be--for reasons of economy--without becoming unrepresentative of the universe, and without the errors acquiring such dimensions as to make significance testing illusory."²

²Ibid., p. 170

It is clear that hard information is needed on the extent to which parts of a single text will differ with respect to the frequency of various measured units. It is also clear that different units may occur with varying degrees of consistency throughout a text. The question of the homogeneity of a text is complex. But until the nature of variation within texts is understood, statements about variation between texts cannot be made with great authority.

The Design of the Study

A suitable model for the study of quantitative change in linguistic behavior is one which views change as taking place along dimensions, such that if two texts vary significantly in the proportion or distribution of one or more units, this difference would be attributed to the two texts occupying different positions in a context space. The examination of other texts of varying similarity to each of the original two texts should lead to the description of factors (dimensions) responsible for the original observed difference. The proposed dimensions can then be tested by predicting the behavior of texts not yet examined.

In this study we propose to examine some aspects of the statistical behavior of certain syntactic units in a sample of technical English. In this as in any other study we must carefully set our goals and gather an appropriate

amount of data to carry them out.

The major focus of this study will be on the variation in frequency of syntactic units within the writing of two individuals. A primary hypothesis to be tested is that the distributions of units will remain reasonably the same throughout a single text written by one person. If the distributions are not uniform, several explanations could be offered. For example, the varying content could influence the frequencies; that is, even in a single text there might be contextual variations. A comparison of the individual chapters should reveal such variations since the chapters represent the way in which the content has been divided in the text. For this reason the chapters will be compared with each other in each of the two texts.

There may be other causes for internal differences in a text. During the time that the text was written various circumstances could have arisen to influence the frequencies. This study does not attempt, however, to account for such influences except as they may be correlated with chapter content and position.

The other primary hypothesis to be tested is that the two sample texts will reveal essentially the same distributions. Several studies have compared samples of technical writing as a whole with samples of non-technical writing, but no one seems to have reported on the variation in linguistic performance among individual American

technical writers.

In order to be sure that differences between the texts would be attributable as much as possible to the writers themselves it was decided to select the sample texts from the same discipline. In other words, if a history text differed in average sentence length from a biology text this could be due either to the different writers or the subject areas or both. While it may seem unreasonable to believe that biology and history writings could exhibit distinctive patterns, there is also no inherent reason why technical and non-technical should vary.

The texts selected for this study are both from linguistics. They are:

1. Emmon Bach's Introduction to Transformational Grammars (New York, 1964), all but exercises at the end of chapters.

2. Kenneth Pike's Language in Relation to a Unified Theory of the Structure of Human Behavior (The Hague, 1967), pp. 25-82, excluding bibliographical sections.

The choice of linguistics as the technical field was arbitrary. These samples of technical writing cannot be regarded as random samples of technical writing as a whole, or even of linguistic writing, or even of Bach's or Pike's writing. The requirement of this study for large amounts of data from single texts precluded the possibility of gaining representativeness through the use of many smaller samples. Factors leading to the selection of the particular text by Bach were its relative shortness as a complete

book, its recent publication date, and the varied material covered. The three chapters by Pike may be regarded as a smaller control sample to be available to confirm any major conclusions for the Bach sample. Moreover, it was felt that Pike exhibited a rather different approach to sentence construction from Bach, and that this difference, when demonstrated quantitatively, would dispel any notion that technical writers could not show individual styles. For convenience the samples from Bach and Pike will be referred to hereafter as simply Bach and Pike.

Before conducting a statistical investigation of texts various parameters or units must be selected which later will be counted and used as the basis for determining the similarity of the samples to be compared. The parameters discussed here represent 2 syntactic levels, that of clause and sentence. Table 1 depicts the basic clause level units.

TABLE 1
 CLAUSE LEVEL CLASSIFICATION

| Type | Name | Examples |
|------|-----------------|---|
| 3 | "Be" Clause | This theorem is true. The description has not been useful. |
| 4 | "Active" Clause | This description has many parts. Ideas flourish. Progress gives men hope. Linguists study language. We consider this false. |
| 5 | Passive Clause | This was realized by others. |
| C | "There" Clause | There are few days left. There seems to be no way to do this. |
| E | "It" Clause | It is not easy to estimate this quantity. It seems futile to try this. |

Sentence types are defined through constituent clause types. A sentence is assumed to consist of a sequence of clauses, each of which is either a main clause or a subordinate clause. In the coded text symbols for main clauses are preceded by an "M". Further, some clauses will be embedded within another clause. Embedded clauses appear in parentheses following the clause in which they are embedded. Thus, those sentences which are composed of the same clauses in the same order are considered to belong to the same sentence type. The following examples should clarify the clause and sentence type classifications:

1. Numerous examples and problems are presented throughout this introduction. Bach, page 2. One main passive clause: M5.

2. These are works that embody in the medium of language the esthetic values of the individual or the community. Bach, page 1. A main be clause followed by a subordinate transitive clause: M34.

3. The particular way of stating a theory of a language with which we shall be concerned has taken inspiration from modern logic. Bach, page 9. A main transitive clause with an embedded be clause: M4(3).

4. It is doubtful whether there are any natural languages conforming to any of these types. Bach, page 105. A main it clause followed by subordinate there and transitive clauses: MEC4.

5. We set up terminally discontinuous constructions as continuous ones and then separate them. Bach, page 120. Two main transitive clauses: M4M4.

The coding of the original texts was carried out "manually," that is, no computer program was written to convert

the source text to coded text. For each chapter (8 in Bach, 3 in Pike) the occurrences or tokens of each of the clause and sentence types were counted and compared. The chi-square test was employed to determine the validity of the assumption that the chapters in each text can be regarded as random samples from one population.

The counting and statistical analysis was carried out through the facilities of the Michigan Terminal System at the University of Michigan Computing Center. This time-sharing system is presently driven by two IBM System /360-67 processors. The clause level unit analysis programs were written in assembly language and FORTRAN IV. The sentence type counting was programmed in SNOBOL4.

Results for Bach

Table 2 depicts the frequency counts of the five clause types in Bach. Here considerable variation is apparent, especially in the be clause and the passive clause. The there and it clause frequencies appear to be relatively constrained. The assumption that the chapters may be regarded as random samples from one population must be rejected.

The frequency of the most common sentence types in Bach is illustrated in Table 3. The percentages given in the table represent the proportion of a sentence type among the five sentence types listed. It was expected that a few sentence types would occur quite often, and that many types would be found only once. It was disappointing,

TABLE 2
 CLAUSE DISTRIBUTION FOR BACH

| Chapter | Clause Type | | | | |
|---------|--------------|---------------|---------------|--------------|-------------|
| | <u>be</u> | active | passive | <u>there</u> | <u>it</u> |
| 1 | 54 24.3% | 114 51.4% | 41 18.5% | 4 1.8% | 9 4.1% |
| 2 | 90 19.8% | 191 42.0% | 135 29.7% | 18 4.0% | 21 4.6% |
| 3 | 95 21.6% | 179 40.8% | 139 31.7% | 8 1.8% | 18 4.1% |
| 4 | 118 21.1% | 238 42.5% | 171 30.5% | 12 2.1% | 21 3.7% |
| 5 | 185 22.8% | 405 50.0% | 167 20.6% | 35 4.3% | 18 2.2% |
| 6 | 55 15.8% | 174 49.9% | 98 28.1% | 8 2.3% | 14 4.0% |
| 7 | 238 30.5% | 301 38.6% | 171 21.9% | 31 4.0% | 39 5.0% |
| 8 | 86 19.4% | 208 46.8% | 94 21.2% | 28 6.3% | 28 6.3% |
| Total | 921 22.7% | 1805 44.5% | 1016 25.1% | 144 3.6% | 168 4.1% |

Chi-square value: 123.99.
 Probability: less than .001.

TABLE 3
DISTRIBUTION OF MOST FREQUENT SENTENCE TYPES

| Chapter | Type | | | | |
|---------|--------------|--------------|--------------|-------------|-------------|
| | M3 | M4 | M5 | M44 | M45 |
| 1 | 21 40.4% | 19 36.5% | 4 7.7% | 5 9.6% | 3 5.8% |
| 2 | 17 18.5% | 31 33.7% | 25 27.2% | 7 7.6% | 12 13.0% |
| 3 | 15 18.8% | 33 41.2% | 22 27.5% | 2 2.5% | 8 10.0% |
| 4 | 14 14.6% | 27 28.1% | 37 38.5% | 14 14.6% | 4 4.2% |
| 5 | 35 24.5% | 52 36.4% | 27 18.9% | 22 15.4% | 7 4.9% |
| 6 | 9 14.1% | 32 50.0% | 13 20.3% | 7 10.9% | 3 4.7% |
| 7 | 28 24.3% | 43 37.4% | 33 28.7% | 4 3.5% | 7 6.1% |
| 8 | 14 15.6% | 42 46.7% | 15 16.7% | 8 8.9% | 11 12.2% |
| Total | 153 21.0% | 279 38.1% | 176 24.0% | 69 9.4% | 55 7.5% |

Chi-square value: 71.32.
Probability: less than .001.

however, to find that only five types occurred with sufficient frequency for statistical testing.

There is clearly little consistency in the frequency of these sentence types, and the chi-square test is able to reject strongly the hypothesis of homogeneity of the chapters. A cursory inspection of the table reveals little overall pattern. The main passive type (M5) occurs least in chapters 1 and 8, the introduction and the conclusion. This is consistent with the notion of the passive clause being highly correlated with technical material. Of course, the main passive type is not the only source of passive clauses. The active plus subordinate passive type (M45) listed in the table also provides one passive clause per sentence. We find that this type has its lowest frequencies in chapters 4 and 7. There is, then, no strong correlation between sentence types on the basis that they both contain passive clauses.

Bach and Pike Compared

Table 4 depicts the distribution of clauses in Pike. As for Bach, the assumption that the chapters represent random samples from one population must be rejected. As in Bach, the passive varies considerably from chapter to chapter. Bach's first chapter, the introduction, has the smallest proportion of passives but Pike's first chapter has the most passives. Bach's be clauses range from 15.8

per cent to 30.5 per cent, but Pike's be clauses are more stable, ranging from 16.1 per cent to 22.4 per cent. Pike's active and passive clauses are also more consistent, but with eight chapters it must be taken into account that Bach has a greater opportunity to reveal inconsistency. Bach appears to use slightly more be clauses, many fewer active clauses, and somewhat more passive and it clauses. The difference in the frequency of there clauses does not seem substantial. A chi-square test comparing Bach's and Pike's clause totals yields a probability far less than .001.

TABLE 4
CLAUSE DISTRIBUTION FOR PIKE

| Chapter | Clause Type | | | | |
|---------|--------------|--------------|--------------|--------------|------------|
| | <u>Be</u> | Active | Passive | <u>There</u> | <u>It</u> |
| 1 | 51 17.8% | 143 49.8% | 76 26.5% | 11 3.8% | 6 2.1% |
| 2 | 145 22.4% | 338 52.2% | 132 20.4% | 15 2.3% | 17 2.6% |
| 3 | 60 16.1% | 227 60.9% | 65 17.4% | 17 4.6% | 4 1.1% |
| Total | 256 19.6% | 708 54.2% | 273 20.9% | 43 3.3% | 27 2.1% |

Chi-square value: 23.19
Probability: between .001 and .005.

We recall that in examining Bach's sentence types only a handful occurred with sufficient frequency in each chapter to allow statistical testing, in spite of a sample of almost 2000 sentences. There are far fewer sentences in the Pike sample of 446 sentences, and in addition Pike appears to use proportionally more sentence types due to his preference for sentences with four or more clauses, copies of which are not likely to be found again. It is not surprising, then, that just two or three sentence types occur often enough for testing. Rather than attempt any judgment on the consistency of Pike's sentence types on such meager evidence, we proceed to a summary of the most frequent sentence types in Pike and Bach.

The results, given in Table 5, clearly indicate the authors' different preferences, but at the same time there are marked similarities in their frequency of usage of some types, for example the M34 and M43 types. We must remember that Bach's most common sentence types were shown to be strongly non-homogeneous, and thus the data in Table 5 cannot be regarded as highly predictive of the performance to be found in other Bach samples. Because of this great internal inconsistency a chi-square test was not carried out on the data in Table 5.

Conclusions

This study has produced, we believe, much useful and interesting data which leads to several major conclusions

TABLE 5

MOST FREQUENT SENTENCE TYPES IN BACH AND PIKE

| Type | Rank in Bach | Propor. in Bach | Propor. in Pike | Type | Rank in Pike | Propor. in Pike | Propor. in Bach |
|------|--------------------|-----------------------|-----------------------|-------|--------------------|-----------------------|-----------------------|
| m4 | 1 | 14.3% | 11.2% | m4 | 1 | 11.2% | 14.3% |
| m5 | 2 | 8.9% | 3.6% | m3 | 2 | 6.3% | 7.8% |
| m3 | 3 | 7.8% | 6.3% | m44 | 3 | 5.4% | 3.6% |
| m44 | 4 | 3.6% | 5.4% | m5 | 4 | 3.6% | 8.9% |
| m45 | 5 | 2.8% | 2.2% | mc | 5 | 2.5% | 1.5% |
| mf* | 6 | 2.2% | .45% | m45 | 6 | 2.2% | 2.8% |
| 4m4 | 7 | 2.0% | .90% | m43 | 7 | 2.0% | 1.7% |
| m54 | 8 | 1.8% | .90% | m4m4 | 8 | 1.8% | 1.2% |
| m34 | 9 | 1.8% | 1.6% | m34 | 9 | 1.6% | 1.8% |
| m43 | 10 | 1.7% | 2.0% | m4(4) | 10 | 1.3% | .66% |
| mc | 11 | 1.5% | 2.5% | m344 | 11 | 1.3% | .45% |
| m35 | 12 | 1.3% | 1.1% | m455 | 12 | 1.1% | .50% |
| m4m4 | 13 | 1.2% | 1.8% | m35 | 13 | 1.1% | 1.3% |
| me4 | 14 | 1.1% | .67% | m54 | 14 | .90% | 1.8% |
| m33 | 15 | .86% | . . | m444 | 15 | .90% | . . |

*f represents an imperative clause.

about the nature of language performance.

The first conclusion is that the model of a writer producing language by drawing samples of linguistic units at random from a specific and unchanging population is untenable. The evidence given here is strongly against such a model, but it is not certain whether the difficulty with such a model is to be traced to non-random sampling from a constant population, or random sampling from a changing population, or non-random sampling from a changing population. Moreover, it is not clear how any one of the three alternative models could be demonstrated superior to any of the others, since there seems to be no way to distinguish empirically between the effects of non-random sampling and a changing population.

The random sample-uniform population (RSUP) model for a single writer appears to be the foundation for many studies in statistical stylistics and linguistics, although this is often not expressed in any explicit way. These studies are designed as follows. The hypothesis is that two or more writers or genres differ substantially in the use of one or more linguistic units such as sentence type, sentence length, adjective-verb ratio, etc. Brinegar, for example, has stated this hypothesis in this way:

The use of this method assumes that every author unconsciously uses words that, at least in the long run, could be considered as random drawings from a fixed frequency distribution of word lengths. This should

be true at least for writings of a related type over a reasonable span of years.³

A null hypothesis of no significant difference (homogeneity) is tested through the drawing of random samples from the writers or genres and the selection of a statistical test and a significance level. The sample sizes are chosen on the basis of their being large enough to be representative of the writers or genres. The test is applied to the sample data and the null hypothesis is either maintained or rejected. Thus homogeneity is a black or white proposition in this approach, a function of the vagueness of sample size and significance level.

The RSUP model interprets the discovery of a statistically "significant" difference between writers or genres as something unexpected and hence worthy of note, meaning that the writers or genres represent distinct populations. Yet, the demonstration of statistically significant differences is something to be expected with a sufficiently large sample size. What is needed is an approach which takes sample size out of immediate consideration and relativizes the concept of homogeneity.

³Claude S. Brinegar, "Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship," Journal of the American Statistical Association, LVIII (March, 1963), p. 87.

There is really no intuitive support for the notion of the homogeneity of linguistic units as something absolute. There is nothing intuitively objectionable about a statement that one feature is more or less consistent than another.

This study is not the first, of course, to use large enough sample sizes within a genre to demonstrate very significant internal differences. A recent analysis of some aspects of the Brown University corpus reveals such differences within fifteen genres for parts of the sentence length distribution in words, using the chi-square test.⁴ There was no attempt to relativize the results of the chi-square test, but sentence length was described for each genre in terms of the mean, standard deviation, and coefficient of variation, the latter being the standard deviation divided by the mean, a measure better suited than the standard deviation alone to indicate the extent of dispersion in the distribution.

An investigation of the homogeneity of individual vocabulary items between genres revealed once again very significant differences.⁵ Rather than settle for a state-

⁴Henry Kučera and W. Nelson Francis, Computational Analysis of Present-Day American English (Providence, R.I.: Brown University Press, 1967), pp. 378-379.

⁵Ibid., pp. 277-293.

ment that certain words were found to be non-homogeneous in the corpus at certain sample sizes, Mosteller proposes an "index of contextuality" as a measure of relative consistency for a specific word.⁶ This index is computed by dividing the chi-square value by the sample size and then multiplying by 1000. The effect is to treat each sample as if it consisted of exactly 1000 units, and the resulting index can be used to rank the homogeneity of individual words. In this fashion Mosteller computes indexes of 6.2, 6.9, and 9.6 for to, and, and the, which were the least contextual, or least influenced in frequency by context, in the Brown corpus.

The second major conclusion of this study is that an index like Mosteller's is an appropriate way to treat homogeneity in a corpus. Using this approach to study different writers one would segment each writer's works into n segments, take random samples, preferably of equal size, from each of the segments and compute an index of contextuality for each feature measured. If there is a central tendency in a feature for the writers, an index of contextuality for pairs of writers considered together may be computed as a measure of their variance.

⁶F. Mosteller, "Association and Estimation in Contingency Tables," Journal of the American Statistical Association, LXIII (March, 1968), pp. 1-28.

Such a comparison of indexes of contextuality presupposes a consistent number of segments and data divisions (parts of speech, sentence length groupings, etc.), since these lead to the degrees of freedom of the contingency table, and for a greater number of degrees of freedom a greater chi-square value is expected for a given deviation from randomness. Nevertheless, for purposes of rough comparison one may wish to examine the homogeneity of features with different underlying degrees of freedom. Another index should be of interest in this regard: the sample size necessary to reject the null hypothesis at a level of .001. This index is in a way more concrete than the index of contextuality in that the degree of consistency is related to the number of units being measured. More important, the degrees of freedom is taken into account. For the index of contextuality a higher value means less uniformity for the feature, while a higher value for the rejection size means more uniformity.

Table 6 gives values for these two indexes for a number of features as a basis for determining the relative similarity of Bach and Pike. As can be seen, the two writers agree relatively closely on the ratio of main to subordinate clauses of the passive type, but differ greatly on this same ratio for the there type.

We believe that the categorization of clause and sentence types used here is reasonable and simple, and that this sort of categorization would be readily applicable to other languages. In addition, statistical measures such as the index of contextuality and rejection size appear to be quite useful as indicators of the consistency of linguistic performance.

TABLE 6
INDEX VALUES FOR BACH AND PIKE COMPARED

| Feature | Degrees of Freedom | Index of Contextuality | Rejection Size |
|-----------------------------------|--------------------|------------------------|----------------|
| Word Level | 15 | 1.7 | 22200 |
| Clause Level | 4 | 11.1 | 1666 |
| <u>Be</u> Clause, S-M | 1 | 1.3 | 8300 |
| Active Clause, S-M | 1 | 1.5 | 7200 |
| Passive Clause, S-M | 1 | .65 | 16620 |
| <u>It</u> Clause, S-M | 1 | 5.3 | 2038 |
| <u>There</u> Clause, S-M | 1 | 29.15 | 370 |
| Clauses, Nested vs. Non-Nested | 1 | .1 | 10800 |
| Material in Parentheses | 1 | 2.9 | 3720 |
| Sentence Length in Words | 6 | 63.5 | 354 |
| Sentence Length in Clauses | 4 | 61.1 | 300 |
| Paragraph Length in Clauses | 2 | 76.35 | 154 |
| Paragraph Length in Sentences | 1 | 98.4 | 110 |

REFERENCES

- Bailey, Richard W. "Statistics and Style: A Historical Survey." Statistics and Style. Edited by Lubomír Doležel and Richard W. Bailey. New York: American Elsevier, 1969.
- Bailey, Richard W., and Doležel, Lubomír, eds. An Annotated Bibliography of Statistical Stylistics. Ann Arbor: Michigan Slavic Contributions, Bibliographical Series No. 2, 1968.
- Doležel, Lubomír. "A Framework for the Statistical Analysis of Style." Statistics and Style. Edited by Lubomír Doležel and Richard W. Bailey. New York: American Elsevier, 1969.
- Edmundson, H.P. "Mathematical Models in Linguistics and Language Processing." Automated Language Processing. Edited by Harold Borko. New York: John Wiley and Sons, 1967.
- Huddleston, R. D.; Hudson, A.; Winter, E.O.; and Henrici, A. Sentence and Clause in Scientific English. London: Communication Research Centre, University College London, 1968.
- Kaufman, S. I. "Ob Imennom Kharaktere Tekhnicheskovo Stilja." Voprosy Jazykoznanija, X, No. 5 (1961), 104-06.
- Simpson, Harold. "A Descriptive Analysis of Scientific Writing." Unpublished Ph.D. dissertation, The University of Michigan, 1965.
- Streeter, Victor J. "Homogeneity in a Sample of Technical English." Unpublished Ph.D. dissertation, The University of Michigan, 1969.