IV. A STATISTICAL AND COMPARATIVE ANALYSIS OF INDIVIDUAL WRITTEN LANGUAGE SAMPLES¹

JOHN W. CHOTLOS
University of Iowa

I. INTRODUCTION

THE PRESENT INVESTIGATION is concerned with the relation of certain language variables to (1) the length of sample from which they are derived and (2) certain psychologically pertinent factors. In general, the language measures employed are based on a count of the number of different words (types) and the relationship of such measures to the total number of words, and to the factors of I.Q., C.A., locality (city, town, rural), and sex. Similar measures based on parts of speech categories and their relationship to I.Q., C.A., locality and sex will be reported. Finally the relationship of the reliability of these measures to the length of samples from which they are derived will be given attention.

Certain previous investigations have been concerned with closely related problems. To begin with, Carroll (3) has presented an equation describing the relation of the number of different words (D) to the total number of words (N) in a sample of language. A necessary condition to Carroll's formulation of this relationship is that a specified relationship hold between the frequency of a given word in a language sample and its rank in order of decreasing frequencies. Zipf (15) discovered that when he plotted frequency of a word against the number

of words having that frequency on logarithmic co-ordinates, the points approximated a straight line except for the few most frequently occurring words. From this fact he formulated the harmonic series law of word distribution, in which he states that the most frequent word in a large sample of language makes up $\frac{1}{10}$ of the sample, the second most frequent word $\frac{1}{20}$ of the sample, the third most frequent word $\frac{1}{30}$ of the sample, etc. This formulation can be put in the form of the equation

$$F = \frac{N}{10R}$$

in which F is the frequency of occurrence of any given word in a language sample, R is its rank in order of decreasing frequencies, and N is the total number of words in the sample.

Skinner (13) has also presented results pertaining to the relationship between F and R . In analyzing the results obtained from 1,000 responses to his verbal summator, he plotted ranks of words in order of decreasing frequencies (R) against frequency (F), expressed as a percentage of the total sample, on logarithmic coordinates, and found the points tended to fall on a straight line.

A deviation from linearity was again noted in the more frequently used words. Skinner (13) also reanalyzed the Kent-Rosanoff (8) data on free association response words in the same manner. He found that when the rank order of words in terms of mean frequency per thousand was plotted against mean frequency per thousand on logarithmic coordinates, the resulting curve was approximately linear

¹ This study was done in the Department of Psychology at the State University of Iowa as a dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy. It is part of a program of research on language behavior. The study was directed by Drs. Wendell Johnson and Don Lewis. Funds and assistance were provided by the Federal Work Projects Administration in connection with Iowa WPA Projects 4892 and 5960.

for the 100 responses most likely to occur. The equation

$$f = \frac{300}{R^{1.29}}$$

where f is the frequency with which a given association will occur in 1,000 responses and R is its rank in terms of mean-frequency per thousand, he finds to be descriptive of the 75 words having the strongest first associations in the Kent-Rosanoff list. He states that this formula is slightly less accurate for the total sample, and his calculated and observed points appear to agree satisfactorily. However, he states that this equation has little practical significance since the frequency and rank of a word must be ascertained before the equation can be used. Nevertheless, he feels that it has an important bearing on theories of language.

Carroll argued that if one accepts the harmonic series law of word frequency distribution, i.e.,

$$F = \frac{N}{KR}$$

where F is the frequency of any word in a language sample, R its rank in order of decreasing frequencies, N the total number of words in the verbal output sample, and K is a constant which is an indirect index of diversity, then it can be demonstrated that the following equation holds:

$$D = \frac{N}{K} (0.423 + K - \log_e N + \log_e K)$$

where D is the number of different words in a sample, N the total number of words in that sample and K is an empirically determined constant. This equation, if it can be shown to be applicable in general, has very important implications with regard to language since, in the

first place, if D is known for a specified N , predictions can be made to other N 's, and, secondly, the nature of the curve allows a determination of a maximum value of D , a value which can be correlated to a given type of vocabulary of the individual. Carroll tested this equation with a verbal output sample obtained by means of the verbal summator technique, and on several language samples from literature, and found the empirical points to fall very near to the computed curve.

The language samples which formed the protocols of these investigations into the relationship between D and N and between F and R , have been accumulated from different sources and massed into one language sample or are the product of verbally proficient writers. Thorndike (14) has emphasized that if language is to be viewed as behavior, the motivation, backgrounds, the individual characteristics of the writers or speakers must be taken into consideration. He further suggested that the relationship between F and R reported by Zipf may in some measure be a statistical artifact produced by combining language from varied sources, such combination resulting in a loss of individual variation. In the light of this criticism it seems desirable to apply the mathematical formulations of these relationships to samples of language which are the product of but a single individual, in order to test their adequacy more fully.

A second point of interest in the consideration of previous studies is that various attempts have been made to relate the number of different words in a sample to psychologically pertinent factors. Fairbanks (5) working with spoken language and Mann (9) with written language compared superior university freshmen and schizophrenic patients in

terms of the mean percentage of different words per 100-word segment, i.e., the 100-word type-token ratio. Both investigators found the mean type-token ratio for superior freshmen to be significantly greater than for schizophrenic patients, indicating a wider vocabulary range for freshmen than for schizophrenic patients. Fairbanks suggests that for spoken language, there might be a positive correlation between the 100-word type-token ratio and intellectual level. On the other hand, Mann (9) found that differences in intelligence test scores, level of educational attainment, and duration of hospital confinement had relatively little influence on the type-token ratio of patients in terms of accounting for differences between her two groups. In neither study were significant sex differences found. Fossum (6), studying spoken language obtained from junior college students in a regular speech class, found by means of a correlation technique that the 100-word type-token ratio based on 18 segments appeared to be related to parental occupation, correlation of .56, and to speaking rate, correlation of -.45, but not to vocabulary as measured by the Nelson-Denny Reading Test, correlation of .03, nor to intelligence as measured by the percentile score on the Ohio State Psychological Test, correlation of .09. Fossum found no sex differences in type-token ratio measures in his group.

Thirdly, counts of the number of words in parts of speech categories have been made by Fairbanks (5) and Mann (9). They related these counts to other variables under consideration. Fairbanks (5) found that for spoken language there were differences in the use of various parts of speech between her freshmen and schizophrenic groups. The schizophrenics used proportionately more pro-

nouns and verbs and proportionately fewer nouns and articles. On the other hand, for written language, Mann (9) found that the results of the grammatical parts of speech count were not significant, although there seemed to be a tendency for the patients to use more nouns than the freshmen, a result which does not agree with the comparable result obtained by Fairbanks.

Fourthly, Fossum (6) has attempted to relate the reliability of the type-token ratio to the length of spoken language sample. He found the correlation between 100-word type-token ratios for the two halves of 1,800-word samples to be .58 and he estimated by means of the Spearman-Brown prophecy formula that a sample of 14,000 words would be needed to give a reliability coefficient of .95.

Many important facts about language have been reported, and it is of undoubted importance to know whether the relationships already reported will hold for individual language samples. Many investigators have used verbal output samples in which the individual characteristics of the writers and speakers have been lost through massing of the data. In other instances, where individual variations have been of major interest in the investigation, the samples of individuals have been highly selected. It is proposed in this study to investigate language characteristics of individual verbal output samples in which the population sampled will allow for considerable generalization of the results. The question of whether or not individual verbal output samples will bear out the equations descriptive of massed language data is an important one. Furthermore, the relationship between language measures and I.Q., C.A., sex, etc., appears to need further investigation in popula-

tions less highly selected in these characteristics than the ones which have been so far reported.

II. THE PROBLEM

The object of this study may be oriented around the analysis of the number of different words (D) as a function of the total number of words (N). This relationship may be symbolized by the formula.

$$D = f(N)$$

In this equation it is apparent that we can hold N constant and study D in relation to other variables, we can study the variation in D with concomittant variation in N , and finally we can study the variation in D and N in relation to other variables, such as intelligence test score, age, etc. The basic unit of analysis is the language sample of a single individual. With this introduction the purpose of this study can be summarized in the following statements and questions:

1. To test empirically the equation derived by Carroll, namely,

$$D = \frac{N}{K} (425 + K - \log_e N + \log_e K)$$

where D is the number of different words in a sample of length N and K is an empirical constant to be determined from the data. It is a further purpose to test the assumptions under which this equation was developed.

2. Can the relationship between D and N be expressed by some empirically determined curve? If such a curve can be determined can the constants in this curve be given any rational meaning?
3. Does D for specified N 's differentiate I.Q. groups, age groups, location groups, and sex groups? And what is the extent and direction of these differences?
4. Do sections of the language samples, categorized by parts of speech, reveal any relationships or differences which are not apparent in the sample as a

whole? How are the parts of speech which go to make up the total sample interrelated?

5. What is the minimum size of sample that can be drawn to reveal the relationships and differences under investigation?

III. SUBJECTS AND PROCEDURES

As part of a remedial education survey, sponsored by the Iowa Child Welfare Research Station and financed by the Federal Work Projects Administration, approximately 1,000 public school children wrote manuscripts of 3,000 words each under conditions to be specified below. The collection and preliminary analysis of these manuscripts was carried out by Work Projects Administration personnel under the supervision of persons with background training in psychology, who had been given special training for this particular assignment.² The survey operated for a period of about two years in five counties of the state of Iowa. These counties are distributed throughout the state in such a way that no two counties were adjacent. Each county survey was operated as a unit, coordination being achieved by a state-wide supervisor stationed at the university. In each county all schools, including the one-room rural schools, were invited to participate in the program.

Unit supervisors were instructed to collect 3,000-word language samples for an allotted number of pupils in their respective units. As it took several hours for a child to write the number of words required, extensive cooperation from the school administrators and teachers was

² Professor George D. Stoddard served as general director of the survey, of which the language study was a part; Dr. Wendell Johnson was the technical director; Dr. C. Ecco Abermann and Mr. George Wischner served successively as statewide supervisors; and the present writer was the project statistician.

TABLE I
Factorial design of experimental sample

		Male			Female		
		City	Town	Rural	City	Town	Rural

I.Q.**	Age 1*	2	2	2	2	2	2
1	Age 2	2	2	2	2	2	2
	Age 3	2	2	2	2	2	2
I.Q.	Age 1	2	2	2	2	2	2
2	Age 2	2	2	2	2	2	2
	Age 3	2	2	2	2	2	2
I.Q.	Age 1	2	2	2	2	2	2
3	Age 2	2	2	2	2	2	2
	Age 3	2	2	2	2	2	2

* Age 1—149 months and under.

Age 2—150 to 179 months.

Age 3—180 months and over.

** I.Q. 1— 89 and under.

I.Q. 2— 90 to 109.

I.Q. 3—110 and over.

*** The numbers refer to the number of randomly selected subjects in the cell.

necessary. The plan called for collecting an equal number of samples from city, town and rural school children, from equal numbers of boys and girls, and an equal number from each grade from four through twelve. Localities with a population of 25,000 or over were called cities, other localities and consolidated schools were considered as town schools, and rural schools of the one-room variety were considered as rural for purposes of this study. Since, as a rule, the one-room rural schools have only eight grades, it was necessary to classify town school pupils who had a rural school background and whose parents were farmers, as rural in order to fill out the rural categories at the older ages. In collecting this sample the pupils were matched by sex for grade, age (within six months), I.Q. (within five I.Q. points) and socio-economic level (within the limits of 1920 U. S. census occupational classification system). No pupil under eight years nor over eighteen years of age was included in the sample.

The writing was done under the supervision of a worker who remained in

the classroom throughout the writing session. Writing sessions averaged about forty minutes in length and, on the whole, four or five writing sessions were required for a child to complete his assigned task.

The worker in charge read the following instructions before the children began to write:

"You are to write about anything you want to write about. Just make it up as you go along. That is, don't write anything you have memorized such as stories or poems. Just start with the first thing you think of and try to keep on writing steadily."

If a child stopped writing for longer than five minutes or complained that he couldn't go on, the worker was instructed not to tell him what to write about, but to tell him to write on whatever he was thinking about. No positive suggestion as to topics was allowed. Legibility of the manuscript was emphasized and speed was not encouraged. Each day's writing was handed in to the monitor at the close of the session. The worker counted the number of words written, entered the count in his record

TABLE 2

Means and standard deviations of distributions of ages in months for the total group and for the main sub-groups

Group	Mean	Standard Deviation
Age Groups		
149 months and under	129.806	12.684
150 to 179 months	164.389	8.473
180 months and over	189.639	7.779
I.Q. Groups		
89 and under	163.917	26.186
90-109	162.083	26.319
110 and over	157.833	26.455
Location Groups		
City	165.194	24.120
Town	159.861	28.136
Rural	158.778	26.469
Sex Groups		
Male	160.796	25.887
Female	161.759	27.020
Total Group	161.278	26.443

TABLE 3

Means and standard deviations of distributions of Otis intelligence test scores (I.Q. units) for the total group and the main sub-groups

Group	Mean	Standard Deviation
Age Groups		
149 months and under	101.278	14.251
150 to 179 months	101.528	14.297
180 months and over	99.028	14.937
I.Q. Groups		
89 and under	83.833	3.670
90-109	101.861	4.461
110 and over	116.139	4.596
Location Groups		
City	101.056	14.012
Town	100.111	14.240
Rural	100.667	13.383
Sex Groups		
Male	100.556	14.453
Female	100.667	13.300
Total Group	100.611	13.888

and dismissed the subject when he had reached the prescribed quota.

From this basic sample of approximately 1,000 language samples, an experimental sample of 108 was selected to conform to the factorial design in Table 1. Since a complete record was available on each child it was possible to sort the larger sample of manuscripts into the fifty-four cells of the design and select at random two subjects for each cell. There was no matching by sex in the experimental sample as was the case in the survey sample. Intelligence was tested by means of the Otis Quick-Scoring Mental Ability Tests. The Alpha test was administered to pupils in grades one through four, the Beta test to pupils in grades five through nine, and the Gamma test to pupils in grades ten through twelve. In classifying the subjects according to the design in Table 1, no distinction was made between the various forms of these three tests. Ages were computed as of the day the children started writing. Criteria for the

locality levels of the design are the same as those mentioned above for the collection of the basic sample. The design permits a distribution of 36 subjects at each of three I.Q. levels: (1) 89 and under, (2) 90 to 109, and (3) 110 and over; 36 subjects at each of three age levels: (1) 12 years, 5 months and under, (2) 12 years, 6 months to 14 years, 11 months, and (3) 15 years and over; 36 subjects at each of three locality levels: (1) city, (2) town, and (3) rural; and 54 subjects in each of the sex groups. Furthermore, many combinations of I.Q., age, locality and sex levels are possible.

Means and standard deviations for distributions of I.Q. and age for the total experimental sample and for the main sub-groups, i.e., in terms of I.Q., age, locality and sex, are presented in Tables 2 and 3.

Following the collection of the language samples, the manuscripts were typed and edited. The definition of a word, it should be realized, is crucial in a study of this type. Quite a bit of free-

dom is permitted in defining a unit of language and the way in which the unit is defined is necessarily a condition that is important in connection with any statements made about language phenomena. For this reason the rules for editing the manuscripts are presented in full. These rules define the fundamental language unit better than any formal definition could. The following rules were followed in editing the samples:

1. Type all words exactly as they are written by the subject. Record each correction by writing it in parenthesis after the word for which it is a correction.
2. Correct each misspelling, recording the word as spelled by the subject and writing the correction after it, in accordance with (1) above.
Classify as a misspelling any word which as spelled by the subject does not constitute a standard English word (current edition of the Century Dictionary to be used as authority) or a recognizable 'slang', nonstandard word (recognizable to the present investigators).
3. Classify as a substitution and correct in accordance with (1) above, any of the following.
 - a. Any correctly spelled homonym or an apparently 'intended' word; e.g., "their" substituted for an apparently intended "there", "bare" for "bear", "four" for "for", etc. Judgment in such cases will involve reasonable interpretation of context.
 - b. Any correctly spelled non-homonymous substitution which apparently distorts the 'intended' sense; e.g., "of your own" for "of your own", "is would be" for "it would be". Judgment in such cases will, again, involve reasonable interpretation of context.
4. Do not insert any word apparently or obviously omitted by the subject. For example, if the subject writes, "It would fun to play ball," do not insert the word "be" at the point where the subject obviously omitted it.
5. Record slang or non-standard words as

written by the subject. When a slang or non-standard word has a standard equivalent, record this equivalent in parenthesis after the slang word; e.g., write "sneaked" in parenthesis as a correction for "snuk". Any slang or non-standard word having no standard equivalent is to stand as written by the subject and misspellings of such words when recognizable are to be recorded in accordance with (1) above.

6. Any proper name which consists of more than one word is to be counted as one word; e.g., "John Jones" is one word; "East St. Louis" is one word; but "East St. Louis, Illinois" is two words, since they constitute two proper names, the name of a city and the name of a state; "The Chicago and Northwestern Railroad" are three words since (1) "the" is never to be regarded as an integral part of a proper name, always being counted as a separate word and (2) any class-name to which a proper name is attached is to be counted as a separate word; e.g., "railroad", "hotel", "theatre", "street", etc., even in such an example as "the Hotel Roosevelt", "the" and "Hotel" are to be counted as separate words. A proper name is one that designates the sole bearer of the name, as: there is only one "Chicago and Northwestern" railroad, only one "Great Atlantic and Pacific" tea company, only one "General Motors" corporation, etc. The names given above in quotes, therefore, are proper names and each is counted as one word. In "A 1940 Multi-Motored Amphibian P45 Boeing Transport", on the other hand, the various words are qualifying adjectives; there are many Transports—Boeing names one type, and it, rather than "Boeing Transport" is a proper name in this case; there are many Boeing Transports, and P45 merely serves as an adjective—there might be P44, P46, etc. Again "Amphibian" serves as an adjective, and so for the terms, "a", "1940" and "multi-motored". In the example of "Dubuque Senior High School", "Dubuque" is an adjective, of course; "Senior High School", however, is not one word in the same sense

that "Chicago and Northwestern" is one word; "Chicago and Northwestern" designates the only railroad that goes by that name, but there are thousands of senior high schools; therefore, "senior" and "high" are to be regarded as adjectives; "Dubuque Senior High School" is to be regarded as four words.

"Mrs.," "Mr.," "Miss" and other modes of address are to be counted as separate words and not as integral parts of proper names; e.g., "Mr. John Jones" are two words.

Titles are not integral parts of proper names, but are to be counted as separate words; e.g., "Doctor Jones" are two words; "Senator Hill" or "Professor Smith" are two words.

Abbreviated titles which consist of more than one unit, e.g., M.D., or Ph.D., or unabbreviated titles which consist of more than one word, e.g., "Speaker of the House" or "Dean Emeritus", are to be counted as single words.

7. Any number is to be counted as one word and all figures are to be written or changed to longhand words. "One", "twenty-seven", "one thousand sixteen" are each a single word. Where time is denoted in numbers, it should be counted as a number; e.g., "7:35," write as "seven thirty-five" and count as one word.

Where street numbers are denoted, write as customarily spoken: e.g., "1220 Harrison Street" write as "twelve twenty Harrison Street".

When numbers are placed at the beginning of sentences for no obvious reason they are not to be included in the typewritten copy and are not to be counted. For example, in one or two cases it was noted that the sentences had been numbered by the child. Such numbers are not to be counted.

8. Contractions are recorded as written: e.g., "didn't" is not to be changed to "did not"; "didn't" is one word.
9. Record abbreviations as written by the subject, with full term in parenthesis.
10. Hyphenated words properly hyphenated (Century Dictionary to be used as

authority) are to be counted as single words; e.g., "hitch-hiker" is one word. Two words improperly hyphenated are to be counted as two words. Corrections in such cases are to be made according to instructions given above.

11. Any two words, as corrected and tabulated, are different unless spelled exactly alike, except:
 - a. Plurals and possessives, and contractions involving apostrophes, are to be differentiated even though they are spelled alike.
 - b. Any word which begins with a capital letter solely by virtue of its place in the sentence is not to be classified as different from a word spelled as it is in all other respects.
12. All recognizable words are to be counted and tabulated except in the case where some symbol is used to indicate a previously written word. These symbols are not to be counted as that word. For example,

"John is going to town tonight.
 " " " " school.
 " " " " be late."

There are nine written words and only nine are to be counted.

13. Sentences are to be left as the child has written them. Do not change the pronoun to agree with the noun in the following example, "Ruth has one side on his paper", nor change the tense in this example, "It is raining last night". Count these as they stand.

After the manuscripts were typed and edited, the types and tokens were recorded and tabulated separately for each manuscript. This tabulation made it possible to abstract the following language measures:

1. *The number of types in any 100-word segment, 500-word segment, 1000-word segment, and in the total 3,000 words.* Thus, 30 measures from 100-word segments, six measures from 500-word segments, three measures from 1,000-word segments and one measure for the total manuscript were computed. For each subject these were averaged to give the mean number of types in 100, 500, and

1,000-word segments, respectively. The mean number of types in a specified segment can be symbolized by D , with a subscript to denote the size of the base from which it is computed. The four measures described above can be symbolized by D_{100} , D_{500} , $D_{1,000}$, and $D_{3,000}$.

2. *The type-token ratio for one-hundred-word, five-hundred-word, one-thousand-word and three-thousand-word segments.* The type-token ratio can be defined as the mean percentage of types in any specified segment. If the number of tokens is symbolized by N , then the type-token ratio can be symbolized by D_i —, where the subscript i specifies the N_i size of the sample on which the type-token ratio was based. In our case, the following type-token ratios were computed:

$$R_{100} = \frac{D_{100}}{100} \quad R_{500} = \frac{D_{500}}{500} \quad R_{1,000} = \frac{D_{1,000}}{1,000}$$

$$R_{3,000} = \frac{D_{3,000}}{3,000}$$

in which R is a symbol for type-token ratio.

R and D , as defined above are equivalent measures so long as the number of tokens, N , on which they are based is equal for the different individuals in a distribution. However, in instances where a distribution is made up of measures based on a varying N , the two measures are not equivalent. In this study, where R and N gave equivalent measures, R was preferred because it made comparison with results of other studies possible.

3. *The cumulative type frequency curve* is obtained by cumulating the types added in each successive 100-word segment. It is the curve that results when D , the number of types, is computed as a function of N , the number of tokens. Since each language sample was sectioned into thirty segments, there are thirty points available in the computation of this curve.
4. *The frequency and rank of each type* was computed from the data. The frequency of a type is the number of

times it occurred in the language sample. A ranking of the types, in whole number steps, and in order of decreasing frequency, was made. Types of equal frequency were given an average rank number. The numerical position of a word in this sequence is its rank. The frequency of any type is symbolized by F_i and its rank by R_i .

Following the computation of the above language measures, the language sample was split into four subsamples. The division was made on the basis of these parts of speech: nouns, verbs, adjectives, and adverbs. It did not seem profitable to analyze the pronouns, prepositions, conjunctions and articles at this time since these parts of speech are much more limited in the number of available types. An attempt was made to classify the words on a functional rather than on a formal basis. In each instance the function of the word in the context in which it was found determined its classification. For example, the word *run* in "He will run" and "She had a run in her stocking" are considered as two different types in this classification, the first a verb and the second a noun, whereas in the previously described definition they would be counted as one type according to editing rule number 11. Curme's (2) text, *A Grammar of the English Language*, was used as a reference and final authority in case of doubt. On the basis of this grammatical analysis these additional measures were abstracted.

5. *The number of nominal, verbal, adjectival and adverbial tokens.*
6. *The number of nominal, verbal, adjectival and adverbial types.*
7. *The type-token ratio for nouns, verbs, adjectives and adverbs.* This measure is not equivalent to (2) above since the number of tokens on which the number of types is based is not equal from individual to individual. Usually type-token ratios are not directly comparable unless they are based on the same number of tokens for each individual, but in this instance, it is felt that distributions of type-token ratios

derived from a varying number of tokens can be justifiably used because the total number of tokens is the same in all the manuscripts. If the parts of speech type-token ratios are weighted by their respective number of tokens, their sum will be found to equal d , the number of types in the language sample.

8. *Percentage of nounal, verbal, adjectival and adverbial types.* This measure was computed by summing all the nounal, verbal, adjectival and adverbial tokens for each individual and then finding the percentage which the separate nounal, verbal, adjectival and adverbial types are of this total.

IV. RESULTS

RELIABILITY OF DATA

It should be realized that the recording, tabulating, and counting operations in this study were unusually extensive. Attainment of absolute accuracy in a study of this type is both expensive and extremely difficult in a reasonable-length of time. Errors were reduced to a minimum by having all operations done twice and by constant supervision of the workers. Further, the procedures for recording, tabulating, and counting the data were so set up as to make possible continuous checking throughout the operations. One set of verbal samples was tabulated independently by two units of the project in order to get some indication of the accuracy of the work. Forty verbal samples comprised this set and the correlation between the two counts of the number of types in each manuscript was .983.

It will be recalled that our data consist of 3,000-word language samples, the individual words of which have been tabulated in such a fashion as to allow for the determination of the number of different words, types, in 100-word segments, or in any segment which is a mul-

tipile of 100 words. In order to test the reliability of the type-token ratios the technique of correlating 'split-halves' was employed. The segments of each of the 108 language samples were split into two halves, the first 1,500 words constituting one-half and the last 1,500 words constituting the other. Mean type-token ratios for 100- and 500-word segments were computed for these two halves and the correlation between the halves computed. The product-moment correlation coefficient for the mean type-token ratios for 100-word segments was .829, and for the 500-word segment type-token ratios the correlation coefficient was .826. Since these type-token ratios were computed for only half of the 3,000 words, it is desirable to have some estimate of what the correlation would be if the whole sample were used as a basis for computing the ratios. Such an estimate can be made by means of the Spearman-Brown prophecy formula. Estimated reliability coefficients for the full length of 3,000 words are .906 for the 100-word type-token ratio and .904 for the 500-word ratio. An assumption basic to the use of the Spearman-Brown prophecy formula is that, in this case, the language sample be homogeneous throughout its length in the above two measures. If the two halves are homogeneous, i.e., measure the same aspect of language, then one would expect the mean type-token ratios for the group to be approximately the same for the two halves. These means were found to be 62.58 and 62.64 for the first and last halves, respectively, with regard to the 100-word type-token ratio, and 40.63 and 40.65 for the first and last halves, respectively, with regard to the 500-word type-token ratio. On this basis, the assumption of homogeneity throughout the verbal sample for these two measures appears to be tenable.

It may be, however, that 3,000 words is an insufficient number to reveal any trends in the behavior of these two measures. One long sample of 18,000 words was available and was used to test the hypothesis of homogeneity on a longer language sample. This sample was ob-

population variance is obtained from the deviations of the individual type-token ratios and the other from the variation in the means. None of these three analyses proved significant, the F-ratios being less than one in the case of the 100-word TTR, 1.756 for the 500-word TTR and

TABLE 4
Type-token ratio measures of an 18,000-word language sample

	Mean 100-word TTR*	Mean 500-word TTR	Mean 1,000-word TTR	Mean 3,000-word TTR
3,000-word Sub-sample				
1	71.57**	50.33	42.40	31.20
2	69.13	48.37	40.80	29.57
3	68.87	47.13	40.00	29.27
4	70.13	48.60	41.07	30.37
5	69.83	47.27	38.50	27.40
6	70.40	45.97	36.40	24.10

* TTR is an abbreviation for type-token ratio.

** All type-token ratios are expressed as percentages.

tained under the same conditions and rules as the shorter samples, except that the subject volunteered to do the task. Several subjects volunteered to write long samples, but only in this one instance was the task carried beyond the 3,000 word quota. It is recognized that one subject, as such, has no statistical status and that no generalization whatever can be made to the language behavior of other children. It is offered as a particular case of language behavior and because it may be provocative of future leads. The child who wrote this long sample was fifteen years old, attended senior high school, had an I.Q. of 120, and came from a town school. The long sample was sectioned into six 3,000 word sub-samples and the 100, 500, 1,000, and 3,000-word type-token ratios computed for each sub-sample. These results are presented in Table 4. Since the type-token ratios in the first three columns, i.e., for 100, 500 and 1,000 words, are means, it was possible to do an analysis of variance in which one estimate of the

1.646 for the 1,000-word TTR. For significance at the five per cent level of confidence, F-ratios of 2.30, 2.54 and 3.20 respectively, would be required. The 3,000-word TTR could not be tested by this technique since there is but one measure of it in each sub-sample.

On the basis of this analysis one would infer that the differences between the various sub-sample TTR's for 100, 500, and 1,000 words can be attributed to chance and that the hypothesis of homogeneity has not been discounted. However, this statement holds only if the six sub-samples are randomly selected from a population of such sub-samples. Since there appears to be a downward trend in the magnitude of these type-token ratios with an increase in the number of words written, except in the case of the 100-word type-token ratio, this assumption of randomness may not be fulfilled. If the sub-samples cannot be assumed to be randomly selected, the results of the F-test can be ignored and the data interpreted in

terms of a trend. It is the opinion of this investigator that in this one instance the results presented in Table 4 are indicative of a trend toward a reduction in the use of types with an increase in the number of words written. With the exception of the 100-word TTR, the TTR's may not be considered as homogeneous throughout the length of this sample of language. Factors operating to produce such an effect may be (1) reduction in the number of topics available to the child tending to a greater repetition of types related to topics already discussed, (2) change in motivating conditions, such as loss of interest, boredom, competition with other activities, and (3) an adaptation to the writing situation tending to produce stereotyped behavior. Further, the fact that the 100-word TTR does not show this trend may be indicative of chance factors operating in the other three ratios, or it may indicate that the various type-token ratios are not measuring the same aspect of language.

In any event, it may be considered as demonstrated that for 3,000 words the 100-word and 500-word TTR's are homogeneously distributed throughout the length of the sample, but that, on the basis of one case, this homogeneity may not be assumed to be necessarily present much beyond 3,000 words, except perhaps for the 100-word TTR's.

A problem closely related to that of reliability is posed by this question: "What is the minimum number of words that need to be sampled from one individual to obtain an adequate measure of his language behavior in terms of type-token ratios?" The answer to this question is of more practical interest than theoretical, since the number of words sampled has been, necessarily, arbitrarily determined. If a positive answer

can be made to this question much laborious and time-consuming work entailed in language studies may be partially eliminated.

Since no child wrote his full quota of 3,000 words in one day and since there appears to be practically no carry-over in any given child's topics from day to day, it is felt that the first part of a child's output could be considered as relatively independent of his last part. Correlations of TTR's for the first and last part of the sample, based on successively larger numbers of words, should give an indication of the reliability of the TTR as the base number of words is increased. For this purpose the following correlation coefficients were computed for 108 pairs of subjects:

- (1) TTR of first 100 words against TTR of last 100 words $r = .378$
- (2) Mean 100-word TTR of first 500 words against mean 100-word TTR of last 500 words. $r = .669$
- (3) Mean 100-word TTR of first 1,500 words against mean 100-word TTR of last 1,500 words. $r = .826$
- (4) First 500-word TTR against last 500-word TTR. $r = .657$
- (5) Mean 500-word TTR for first 1,500 words against mean 500-word TTR of last 1,500 words. $r = .829$
- (6) First 1,000-word TTR against last 1,000-word TTR. $r = .813$

If it is assumed that the various type-token ratios under consideration are equivalent measures, we see that the 1,000-word TTR is practically as good a measure of the individual's language as the average 100 and 500-word TTR's for 1,500 words. Further, the 1,000-word TTR gives a reliability which compares favorably with that estimated for average 100 and 500-word TTR's based on the full 3,000 words, .813 as compared to .904 and .906. For most purposes, a type-token ratio based on 1,000 words will

prove adequate and in some instances TTR's based on only 500 words might prove useful.

Finally we seek to answer the question: "How strongly are the type-token ratios interrelated?" An answer to this question will also give a partial answer to the question: "Do the different type-token ratios employed in this study measure the same aspect of language?" An answer to these questions was sought by intercorrelating the four type-token ratios (all four based on 3,000 words). These correlation coefficients are as follows:

	R_{100}	R_{500}	$R_{1,000}$
R_{500}	.934		
$R_{1,000}$	.870	.948	
$R_{3,000}$	.745	.925	.952

The multiple-correlation coefficient of R_{100} , R_{500} , and $R_{1,000}$ with $R_{3,000}$ is .99. Besides the fact that the four TTR's are highly interrelated it is noted that the greater the difference between the base number of words the less the correlation. Since, for each individual, these TTR's are based on the same language sample, this result is to be anticipated, to some extent, on *a priori* grounds. Further, inspection of the scatter-diagrams of these intercorrelations revealed that the relationships are linear. On these grounds, we would judge the four TTR's to measure essentially the same aspect of language.

GROUP DIFFERENCES IN LANGUAGE MEASURES

Differences between the levels of I.Q., C.A., locality and sex were investigated by means of the analysis of variance technique. The analysis of the factorial design was carried out for the following language measures:

1. Language measures derived from the total sample:

- a. Type-token ratio³ for 100-word segments.
 - b. Type-token ratio for 500-word segments.
 - c. Type-token ratio for 1,000-word segments.
 - d. Type-token ratio for the total 3,000 words.
2. Language measures derived from sections of the samples categorized by the following parts of speech: nouns, verbs, adjectives, and adverbs. Each of these four sections of the samples were analyzed separately in terms of the following measures:
 - a. Number of tokens.
 - b. Number of types.
 - c. Type-token ratio computed as the total number of types divided by the total number of tokens of each category.
 - d. The percentage which the types of each category is of the total number of types in the four categories.

From these measures, a total of twenty analyses of variance was made. In the tables presenting these results the following symbols are used to designate the various levels of the factors under consideration:

1. I.Q. levels
 - I_1 —I.Q. 89 and under.
 - I_2 —I.Q. 90 to 109, inclusive.
 - I_3 —I.Q. 110 and over.
2. Chronological age levels
 - A_1 —C.A. of 12 years, five months and under
 - A_2 —C.A. of 12 years, six months to 14 years, 11 months, inclusive.
 - A_3 —C.A. of 15 years and over.
3. Locality levels
 - L_1 —City
 - L_2 —Town
 - L_3 —Rural
4. Sex
 - S_1 —Boys
 - S_2 —Girls

In the twenty applications of the analysis of variance technique, 208 interac-

³ All type-token ratios are presented as percentages rather than as decimal fractions.

TABLE 5
Summary of results of analysis of variance for 20 language measures

	I.Q.		C.A.		Locality		Sex	
	Level of signif. F-test	Rank order* of means of groups	Level of signif. F-test	Rank order of means of groups	Level of signif. F-test	Rank order of means of groups	Level of signif. F-test	Rank order of means of groups
R ₁₀₀	1%	I ₁ I ₂ I ₃	1%	A ₁ A ₂ A ₃				
R ₉₀	1%	I ₁ I ₂ I ₃	1%	A ₁ A ₂ A ₃				
R ₈₀	1%	I ₁ I ₂ I ₃	1%	A ₁ A ₂ A ₃				
R ₇₀	1%	I ₁ I ₂ I ₃	1%	A ₁ A ₂ A ₃				
Number of Tokens								
Nouns	5%	I ₁ I ₂ I ₃			1%	L ₁ L ₂ L ₃		
Verbs					1%	L ₁ L ₂ L ₃	5%	Boys Girls
Adjectives								
Adverbs					5%	L ₁ L ₂ L ₃	5%	Boys Girls
Number of Types								
Nouns	1%	I ₁ I ₂ I ₃						
Verbs	5%	I ₁ I ₂ I ₃						
Adjectives	5%	I ₁ I ₂ I ₃	5%	A ₁ A ₂ A ₃				
Adverbs	5%	I ₁ I ₂ I ₃	1%	A ₁ A ₂ A ₃				
Type-Token Ratio for								
Nouns	1%	I ₁ I ₂ I ₃	5%	A ₁ A ₂ A ₃				
Verbs	5%	I ₁ I ₂ I ₃			5%	L ₁ L ₂ L ₃		
Adjectives	5%	I ₁ I ₂ I ₃	5%	A ₁ A ₂ A ₃				
Adverbs	5%	I ₁ I ₂ I ₃			5%	L ₁ L ₂ L ₃		
Per cent of Total Types								
Nouns								
Verbs	5%	I ₁ I ₂ I ₃						
Adjectives			1%	A ₁ A ₂ A ₃				
Adverbs	1%	I ₁ I ₂ I ₃			5%	L ₁ L ₂ L ₃		

* Rank order of means in terms of increasing magnitude.

Legend: I₁—I.Q. 89 and under; I₂—I.Q. 90 to 100 inclusive; I₃—I.Q. 110 and over.

A₁—C.A. 12-5 years and under; A₂—C.A. 12-5 to 14-11 years inclusive; A₃—C.A. 15 years and over.

L₁—City; L₂—Town; L₃—Rural.

tion variances were computed, and since only three of these were significant, and then only at the five per cent level of confidence, it was felt that, on the grounds that these measures are highly interrelated, it could be safely assumed that there is no interaction among the four factors under consideration as far as these language measures are concerned. By chance one should expect about ten or eleven of the interaction variances to be significant at the five per cent level of confidence.

Because the main purposes of this analysis were exploratory, it was decided that the more conservative error variance estimate should be used to test the main effects. In 19 of the 20 individual analyses, the I x A x L x S variance afforded the more conservative estimate of the error variance and consequently was used as the error term even though the degrees of freedom available for the F-test were much reduced. In the one instance where the I x A x L x S variance

was less than the majority of the interaction variances, an error variance was computed by summing all of the sums of squares of the interaction terms and dividing by the sum of the degrees of freedom of all of the interaction terms.

In cases where the F-test was significant, differences between levels were tested by means of Fisher's t-test. The error variance used as a basis for these t-tests was computed as the residual variance after the variation due to the main effects had been deducted from the total variance. This procedure is permissible on the hypothesis of no interaction and since this hypothesis of no interaction seems tenable, the standard error of difference was computed from these residuals because it permitted a greater number of degrees of freedom in evaluating *t*.

Summary of Results of Analysis of Variance for Each of 20 Language Variables

The evidence garnered from the ap-

plication of the analysis of variance technique to 20 language variables, reveals their capacity to differentiate groups classified according to the factors investigated. A summary in tabular form is presented in Table 5.⁴ Considering the fact that the results of the analyses of variance of these 20 variables are rather ponderous, a skeletonized version of these results would appear to be more appropriate than a detailed account. The summary will collect the significant results for each of the pertinent factors.

A. The I.Q. factor. Of the 20 analyses of variance involving the I.Q. factor 14 resulted in significant F-values. Of these, seven are significant at the one per cent level and seven at the five per cent level of confidence. In general, the direction of differences of the means of I.Q. levels for these variables is in a numerical increase in the value of the measure for increases in I.Q. level. The I.Q. factor seems to be the most strongly related to these language measures.

1. Segmental type-token ratios

The 100, 500, 1,000 and 3,000-word type-token ratios all give F-values significant at the one per cent level of confidence. Means of the three I.Q. levels for the segmental type-token ratios are positively related to I.Q. level.

2. Variables dependent on counts of nouns

Three of the four measures derived from counts of nounal types and tokens resulted in significant F-ratios. The type-token ratio and number of types, respectively, are significant at the one per cent level of confidence, while the percentage of nounal types failed to reach either criterion of significance. In each case, differences in means of these measures

among I.Q. levels is in favor of an increase in the mean value of the measure with an increase in I.Q. level.

3. Variables dependent upon a count of verbs

Three of the four measures involving counts of verbal types and tokens result in significant F-values, all significant at the five per cent level of confidence. These variables gave significant F-values: number of verbal types, type-token ratio for verbs, and percentage of verbal types. The direction of differences among means of I.Q. levels for the number of types and type-token ratio, respectively, is positively related to I.Q. levels, while the direction of mean differences for the percentage of verbal types is reversed, the low I.Q. group using a greater percentage of verbal types than either of the other two I.Q. groups.

4. Variables dependent upon a count of adjectives

Measures based on a count of the number of adjectival types and tokens result in only two of the four measures giving significant F-ratios for the I.Q. factor. Both of these are significant at the five per cent level of confidence. The direction of mean differences for these two significant measures is positively related to I.Q. level.

5. Variables dependent upon a count of adverbs

Of the four measures involving adverbs, the type-token ratio for adverbs and the percentage of adverbial types gave significant F-ratios, the latter significant at the one per cent level of confidence and the former at the five per cent level. I.Q. levels prove to be positively related to the mean type-token ratio for adverbs and negatively related to the mean percentage of adverbial types.

B. The C.A. factor. Nine of the 20

⁴A complete presentation and discussion of the results of the analysis of variance are on file at the S.U.I. library.

language measures, when each is submitted to an analysis of variance in terms of the factorial design, result in a significant F-value for the C.A. factor. Of these nine measures, six give F-values significant at the one per cent level and the remaining three give F-values significant at the five per cent level of confidence. For each of these nine measures, the means of C.A. levels increase in value with an increment in C.A. level; in other words, the older the child the higher the score in terms of these nine measures.

1. Segmental type-token ratios

The four segmental type-token ratios, when submitted to an analysis of variance, result in F-values which, for the C.A. factor, are significant at the one per cent level of confidence in each case. The direction of mean differences is positive, i.e., the older the child the higher the numerical value of the type-token ratios.

2. Variables dependent upon a count of nouns

Only the type-token ratio for nouns resulted in a significant F-value, and this at the five per cent level of confidence. The older children tended to have a greater type-token ratio for nouns.

3. Variables dependent upon a count of verbs

None of the four variables derived from counts of verbal types and tokens results in significant F-values for the C.A. factor.

4. Variables dependent upon a count of adjectives

Three of the four variables involving counts of adjectives prove significant for the C.A. factor. For this factor, the number of adjectival types is significant at the one per cent level. This evidence points to an increase in the use of adjectival tokens and types as the children grow older.

5. Variables dependent upon a count of adverbs

Of the variables in this category only the number of adverbial types gives a significant F-value (at the one per cent level) for the C.A. factor in the analysis of variance. The tendency is for an increase in the use of adverbial types with age, although the two older groups are reversed, but the difference in means between these two groups is not statistically significant.

C. The locality factor. Six of the 20 measures, when submitted to an analysis of variance, gave significant F-ratios for the locality factor. Two of these six are significant at the one per cent level and the remaining four at the five per cent level of confidence. No general trend in differences among the means of the city, town and rural groups was noted.

1. Segmental type-token ratios

Segmental type-token ratios do not differentiate locality groups. None of these four measures gave significant F-ratios for the locality factor.

2. Variables dependent upon a count of nouns

For the locality factor, only one measure, number of nounal tokens, gave a significant F-value (at the one per cent level). The city group uses, on the average, a greater number of nouns than do town or rural groups, while the rural group uses more than do the town group.

3. Variables dependent upon a count of verbs

For the locality factor, the number of verbal tokens and type-token ratio for verbs, respectively, gave significant F-values. Only the number of verbal tokens is significant at the one per cent level of confidence. The rank order of means for number of verbal types is town, city and rural in order of decreasing magnitude, while for the type-token ratio for verbs

the corresponding rank order is city, rural and town.

4. Variables dependent upon a count of adjectives

For the locality factor, measures based on counts of adjectival types and tokens do not produce any significant F-values.

5. Variables dependent upon a count of adverbs

Three of the four variables derived from counts of adverbial types and tokens successfully differentiate locality groups as judged by the F-test. These three measures are the number of adverbial tokens, adverbial type-token ratio, and percentage of adverbial types. All three are significant at the five per cent level of confidence. City groups tend to use the least number of adverbial tokens and types, but have the greatest adverbial type-token ratio. The town group uses the most adverbial tokens while the rural group uses a greater percentage of adverbial types.

D. The sex factor. For the sex factor, only two of the 20 language variables gave significant results in terms of the analysis of variance. These two measures are number of verbal tokens and number of adverbial tokens, respectively, both significant at the five per cent level of confidence. In both instances girls use a greater number of these classes of tokens than do boys.

E. General summary. In general, it may be said that in terms of the language measures employed, the higher the I.Q. and the higher the age level the more highly differentiated is the language structure of the writers. The use of a proportionately greater number of nouns and adjectives characterizes high I.Q. and older age groups, while the use of a proportionately greater number of verbs characterizes the low I.Q. and younger age groups. Adverb usage is not

clearly differentiating among these groups.

On the basis of the analysis of variance one would predict that a correlation exists between the type-token ratios and I.Q. score and between type-token ratios and C.A. The correlation of type-token ratios with I.Q. scores might be attenuated by allowing C.A. to be unrestricted and on the other hand the correlation between C.A. and type-token ratios might be attenuated if I.Q. is allowed an unrestricted range due to the possible counteracting influence of these two factors, although there might be reinforcement rather than attenuation. In any event, it is desirable to determine the relationship between I.Q. and type-token ratio with the effect of C.A. reduced to a minimum, and to determine the relationship between C.A. and type-token ratio with the influence of I.Q. minimized. One method of accomplishing this end is to determine the correlation of I.Q. with type-token ratio within C.A. levels and of C.A. with type-token ratio within I.Q. levels. These correlations may be viewed as empirically determined partial correlation coefficients, in the first case with C.A. held constant and in the second with I.Q. held constant. In each case three measures of the partial correlation coefficient are obtained.

On the basis of the assumed equivalence of the four type-token ratios being studied, only the correlations of the 3,000-word type-token ratio with C.A. and I.Q. were computed. The correlation between 3,000-word type-token ratio and I.Q. and between 3,000-word type-token ratio and C.A. within I.Q. and C.A. groups was also computed.⁵ These

⁵ See Lindquist, E. F. *Statistical Analysis in Educational Research*, Houghton Mifflin Co., 1940, pp. 219-228.

TABLE 6

Table of correlations of 3,000-word type-token ratio with C.A. and I.Q.

Correlation of	<i>r</i>	<i>N</i>
<i>R</i> _{3,000} With C.A. (I.Q. unrestricted)	.332	108
<i>R</i> _{3,000} With C.A. Within I ₁	.463	36
<i>R</i> _{3,000} With C.A. Within I ₂	.395	36
<i>R</i> _{3,000} With C.A. Within I ₃	.463	36
<i>R</i> _{3,000} With I.Q. (C.A. unrestricted)	.517	108
<i>R</i> _{3,000} With I.Q. Within A ₁	.646	36
<i>R</i> _{3,000} With I.Q. Within A ₂	.552	36
<i>R</i> _{3,000} With I.Q. Within A ₃	.548	36
Within Groups Correlation of		
<i>R</i> _{3,000} With C.A.	.325	108
<i>R</i> _{3,000} With I.Q.	.420	108

correlations are presented in Table 6. In each case the unrestricted correlation along with the correlation within groups as well as the empirically determined partial correlations are presented. A comparison of these two correlations reveals that the unrestricted correlation and the within groups correlations tend to be very much alike in this instance. This result is to be expected since the range in C.A. and I.Q. is reduced relatively much more than is the range of the type-token ratio scores. It would seem that a better index of the strength of the relationship of the 3,000-word type-token ratio to I.Q. and to C.A. is the correlation of 3,000-word type-token ratio with I.Q. within C.A. levels on one hand and 3,000-word type-token ratio with C.A. within I.Q. levels on the other, since in each case counteracting influences are somewhat reduced. However, inasmuch as the locality and sex factors were not significant, the within groups correlations permits the estimation of the significance of the correlation of I.Q. and of C.A. with 3,000-word type-token ratio. With 100 degrees of freedom, both correlations are significant at the one per cent level of confidence when tested

by means of Fisher's *t*-test.⁶ The *t*-value for the within groups correlation of C.A. with 3,000-word type-token ratio is 3.51; the equivalent measure for the within groups correlation of I.Q. with 3,000-word type-token ratio is 4.64.

The positive results of these correlations are indicative of a relationship between type-token ratios and I.Q. and between type-token ratios and C.A., but the relationships are not strong enough to predict a type-token ratio for an individual from knowledge of his age and his I.Q. score. On the assumption that the correlation of C.A. and I.Q. is zero, the multiple correlation coefficient of type-token ratio with C.A. and I.Q. is only .608, a result which suggests that C.A. and I.Q. are not sufficient factors for completely determining the type-token ratio.

CUMULATIVE TYPE FREQUENCY CURVE

In a recent article, Carroll (3) states that the equation

$$D = \frac{N}{K} - (.423 + K - \log_e N + \log_e) \quad (1)$$

where *D* is the number of different words in a language sample of length *N*, *N* is the total number of words in that sample, and *K* is an empirically determined constant, held for the language samples he had under investigation. If this formula could be demonstrated to hold generally, it would be a powerful tool in language research.

Carroll deduced equation (1) by means of certain logical and statistical considerations from the following equation,

$$F = \frac{N}{KR^*} \quad (2)$$

where *F* is the frequency with which any given word occurs, *R* is its rank in order

⁶ See Lindquist, E. F. *Op. cit.*, pp. 210-211.

FIG. 1.—Subject no. 12. Graphic representation of the relationship of the rank of a word (R) in decreasing frequency order to the frequency of occurrence (F). The upper plot shows the reduction line in terms of $\log F$ and $\log R$. Empirical points are shown in their relation to the curve described by the indicated equation.

of decreasing frequency, N the number of words in the sample from which F and R are computed, and K is an empirically determined constant which has the same meaning as the K in equation (1). A necessary condition to the applica-

bility of equation (1) is that equation (2) hold for the data and particularly, that the exponent of R have a value of 1.0.

To determine whether equation (2) holds for the language samples under

investigation, 18 of the 108 language samples were selected in such a fashion that two randomly selected cases came from each of nine C.A., I.Q. groups. The

TABLE 7

Estimates of parameters, a and K in the fitted equation $F = 3,000/KR^a$ for eighteen language samples

Subject Number	Total Ranks		First Twenty Ranks Eliminated	
	a	K	a	K
29	0.938	6.716	1.255	2.028
49	0.825	10.549	1.021	4.636
79	0.865	9.080	1.000	5.206
3	0.859	9.709	0.995	5.625
96	0.855	10.211	0.991	5.931
78	0.796	11.700	1.041	4.636
8	0.916	8.513	0.929	8.378
4	0.850	10.989	0.937	7.382
56	0.817	11.646	0.949	6.701
12	0.828	11.619	0.948	6.969
101	0.852	10.142	1.049	4.370
86	0.815	12.948	0.880	9.294
44	0.825	11.891	0.941	7.246
97	0.827	14.741	1.065	3.909
43	0.809	14.495	0.808	14.528
6	0.861	10.003	1.010	5.372
100	0.829	11.299	0.907	8.319
22	0.841	12.448	0.850	12.165
Means	0.845		0.974	

variables F and R were measured for each sample. Equation (2) can be reduced to

$$\log F = -a \log R + \log \frac{N}{K}$$

which is seen to be linear in $\log F$ and $\log R$. If a plot of $(\log F, \log R)$ can be considered linear, then the line of best fit can be determined by means of the method of least squares. A graphical representation of a representative plot of $(\log F, \log R)$ and of (F, R) along with the best fitting curve is presented in Figure 1. Similar curves were computed for each of the 18 selected samples of language. It can be noted that the fit is not good for the lower ranks, and since Car-

roll states that equation (2) holds only for ranks greater than about 20, the best fitting straight line was fitted to each of the 18 plots of $(\log F, \log R)$ for a series of points in which the first 20 (approximately) points corresponding to the lower 20 ranks were eliminated. Estimates of the parameters, a and K , for 18 language samples are presented in Table 7.

If equation (1) is to have any generality, then equation (2), which as was noted previously is the harmonic series law of word frequency distribution, must hold for language samples in general. Specifically it must be demonstrated that a plot of $(\log F, \log R)$ is linear and that the value of the exponent of R is 1.0. On the basis of the 18 plots of values of $(\log F, \log R)$, it was judged that the assumption of linearity is reasonable, although the possibility of some other function giving a better fitting reduction line should not be excluded. As for the exponent of R , it can be noted that the curves in which the first 20 ranks were eliminated result in values of a which are much closer to 1.0 than are the values of a for the entire series of ranks. Further, we note that when the first 20 ranks are eliminated, several of the 18 equations give estimates of a which are practically 1.0. On the other hand we note quite a range in the a values, from 0.808 to 1.235, and we are confronted with the problem of determining if these estimates of a are sufficiently close to 1.0 to support the assumption that the value of this parameter is 1.0.

On the assumption that the value of the parameter a is 1.0, we can, by means of the t -test test the hypothesis that the mean value of a for these 18 language samples differs from 1.0 within the limits of chance. The mean of the 18 a -values differs from 1.000 by 0.026 and results

TABLE 8
Table of K -values for eighteen language samples computed at successive
five-hundred-word points

Subject Number	500 Words	1,000 Words	1,500 Words	2,000 Words	2,500 Words	3,000 Words	Mean
79	5.69	6.03	6.24	6.45	6.56	6.56	6.24
3	6.49	6.67	6.81	6.86	6.99	7.10	6.82
96	6.94	6.97	7.14	7.22	7.28	7.28	7.14
6	7.14	7.08	7.22	7.29	7.30	7.30	7.22
78	7.13	6.88	6.82	6.88	6.91	6.99	6.94
49	6.50	6.75	6.87	6.87	6.90	6.94	6.81
101	6.66	6.65	6.82	6.96	7.09	7.13	6.89
56	6.90	7.15	7.15	7.28	7.31	7.32	7.19
12	6.59	6.78	7.00	7.17	7.42	7.41	7.06
44	7.47	7.38	7.46	7.39	7.31	7.41	7.40
4	6.92	7.21	7.20	7.37	7.36	7.36	7.24
97	6.88	6.89	6.84	6.87	6.90	6.98	6.89
8	7.08	6.75	6.93	6.99	7.06	7.13	6.99
100	7.43	7.24	7.25	7.36	7.40	7.43	7.35
86	7.47	7.45	7.47	7.57	7.53	7.62	7.52
22	7.33	7.37	7.60	7.81	7.85	7.92	7.65
43	7.91	8.02	7.80	7.93	7.97	8.05	7.95
29	6.17	6.22	6.44	6.48	6.58	6.68	6.43
Means	6.928	6.972	7.059	7.153	7.207	7.256	

in a t -ratio of 1.178, which, with 17 degrees of freedom, gives a probability value greater than 0.2 but less than 0.3, a probability interval which, judged by the ordinary criterion of significance, is not significant. This statistical test is not entirely satisfactory, however, since the assumption we wish to test with reference to the value of the parameter a is that it is 1.0 in each individual case and not that the mean of a distribution of randomly selected language samples is 1.0, although the latter is necessarily true if the former holds. On the basis of this test, it is reasonable to accept the hypothesis that the mean value of the parameter a might be 1.0, and, in some degree, the assumption of a harmonic series law of word frequency distribution is seen to be validated when approximately the first 20 ranks have been eliminated.

Equation (1) is peculiar in that it contains only one parameter, K . An estimate of this parameter can be determined from any one point on the cumulative type-frequency curve. If K

is constant, as it must be if equation (1) is to hold, then estimates of K computed at various points along the cumulative type-frequency curve should differ from one another in a chance fashion. Since we can get a distribution of K -values for each individual sample, as well as a distribution of K -values at each successive point along the curve for a group of individuals, we can derive two estimates of the population variance, one based on the variance due to the difference in means at successive points and the other a remainder variance computed from the total sums of squares after the variation due to individuals and to successive points along the curve have been deducted. On the assumption that the various estimates of K differ only by chance, the F -ratio of the two estimated variances should be non-significant.

For the same 18 language samples used to test equation (2), K -values were computed at each successive 500-word point along the cumulative type-frequency curve. Mean K -values for the 18 samples

TABLE 9
Results of analysis of variance of K -values for eighteen language samples

Factor	Sums of Squares	d.f.	Variance	F
Individuals	17.0892	17		
Sample size	1.5524	5	0.3105	18.356
Error	1.4377	85	0.0169	
Total	20.0793	107		

Inter-mean K -value differences					
	500 Words	1,000 Words	1,500 Words	2,000 Words	2,500 Words
1,000 Words	0.044*				
1,500 Words	0.131	0.087			
2,000 Words	0.225	0.181	0.094		
2,500 Words	0.279	0.235	0.148	0.054	
3,000 Words	0.328	0.284	0.193	0.103	0.049

* Differences greater than 0.0848 are significant at the 5% level of confidence. Differences greater than 0.1116 are significant at the 1% level of confidence.

were also computed at each successive 500-word point as well as the mean K -value for each sample. These data are presented in Table 8. It is noted that there appears to be a systematic tendency for the value of K to increase as the base number of words from which it is computed increases, although this tendency is not of equal strength in all cases.⁷

The data in Table 8 were subjected to an analysis of variance in order to determine if the variation in mean K -values at successive 500-word points could, statistically, be allocated to chance factors. The error variance used to test the sample size variance, i.e., variation derived from the means at successive 500-word points, was the interaction variance of individuals and sample size. The results of this analysis of variance are presented in Table 9. The F -ratio of 18.356 is significant at the one per cent level of confidence. This result may be interpreted as meaning that the mean K -values computed at successive 500-word points along the curve cannot be

considered as representing populations which are equally variable or which have equal means. On the basis of a significant F -test, inter-mean K -values were tested by means of the t -test. Of the 15 differences among the six means, 12 give a t -value significant at least at the five per cent level and nine give t -values significant at the one per cent level. If the means of the K -values differed by chance only, one would expect less than one of the 15 differences to be significant at the five per cent level when tested by means of the t -test.

A further test of the adequacy of the hypothesis that K is constant is afforded by an analysis of the behavior of K in the 18,000-word sample. K -values were computed at successive 1,000-word points throughout the 18,000 words. Since K cannot be validly estimated at very large values of N it was necessary to determine whether or not K could be validly computed at each of the successive 1,000-word points along this curve. In equation (1), D is a double-valued function of N , i.e., there are two values of N which will satisfy the equation for any given D . Disregarding the portion of the curve described by equation (1) for nega-

⁷ In one case, subject number 79, the value of D at $N = 3,000$ was not large enough for a valid computation of K . In order to complete the design the value of K computed at 2,500 words was used as the best estimate of K at $N = 3,000$.

tive values of N , the curve may be said to have its origin at point (0.0), to rise to a maximum and to fall indefinitely for all values of N beyond this maximum. The usable portion of the curve is from its origin to its maximum. The maximum point on the curve is determined from K and for each value of K a maximum N can be computed beyond which the value of D computed from equation (1) is not valid. If for each K there is a maximum N , then for each value of N there is a minimum value of K which is valid for that value of N . These minimum values of K for specified N 's can be determined by setting the first derivative of equation (1) equal to zero and solving for K . The first derivative of equation (1) is

$$\frac{dD}{dN} = \frac{1}{K} (K + \log_e K - \log_e N - 0.577) \quad (4)$$

and setting

$$\frac{dD}{dN} = 0$$

$$\log_e N = K + \log_e K - 0.577 \quad (5)$$

as the equation from which we can determine the maximum point on the curve for specified values of N . In terms of the independent variable, N , the limits of the usable portion of the curve derived from equation (1) are from $N = 0$ to

$$N = E_{(K + \log_e K - 0.577)} \quad (6)$$

From equation (5) minimum values of K , which for a specified N give a maximum, can be solved. From this value of K , then, the minimum D value for the specified N and K can be computed from equation (1). These minimum D and K values, along with the empirically determined K -values at each successive 1,000-word point of the cumulative type-frequency curve, are presented in Table 10.

Apparently K -values can be validly

TABLE 10
K-values computed at successive 1,000-word points of the 18,000 word sample

N	D	K	Minimum K for given N	Minimum D for which K is valid
1,000	434	7.82	5.73	175
2,000	732	8.03	6.35	316
3,000	936	8.00	6.68	449
4,000	1,175	8.18	6.94	576
5,000	1,376	8.27	7.13	701
6,000	1,532	8.27	7.29	823
7,000	1,700	8.34	7.42	943
8,000	1,889	8.42	7.54	1,061
9,000	2,021	8.45	7.64	1,178
10,000	2,193	8.51	7.74	1,292
11,000	2,344	8.56	7.83	1,405
12,000	2,465	8.58	7.90	1,519
13,000	2,585	8.62	7.97	1,631
14,000	2,703	8.64	8.04	1,741
15,000	2,786	8.64	8.10	1,852
16,000	2,884	8.66	8.16	1,961
17,000	2,966	8.66	8.21	2,071
18,000	3,025	8.67	8.26	2,179

computed throughout the length of this sample. However, it is to be noted that if we were making a prediction of D from 3,000 words, the number of words collected from the other subjects, the estimate of the maximum number of different words in this child's vocabulary would fall somewhere between 2,585 and 2,703 words, i.e., at the point along the N -axis beyond which computation of D from $K = 8.00$ is no longer valid. The data do not justify the prediction that the child's vocabulary is exhausted after writing between 13,000 and 14,000 words, as new words are being added throughout the length of the sample.

On the other hand, it is noted that beyond N equal to about 11,000 or 12,000 the value of K remains fairly stable. This sample, like the previously discussed samples, gives estimates of K which show a systematic tendency to increase for successive values of N .

The evidence points to a rejection of the hypothesis that, for these language samples, K is a constant. Indirectly, it

FIG. 2.—A graphic representation of the relationship of the number of different words (D) to the sample size (N). The upper curve is the reduction plot in terms of $\log D$ and $\log N$. Plotted from the data computed from language sample written by subject no. 96. Empirical points are shown in their relation to the curve described by the indicated equation.

means that equation (1) will not adequately describe the data derived from these samples and that D will vary systematically from the obtained values of D . We may, on this evidence, reject the generality of equation (1) for language

samples like those used in this investigation, with the qualification that 3,000 words may be an insufficient number of words on which to base an estimate of K for prediction of D 's.

An attempt was made to find an em-

TABLE 11

Estimates of parameters, a and b , in the fitted function $D = bN^a$ for 18 language samples

Subject Number	a	b
29	0.618	3.373
49	0.657	3.192
79	0.702	1.656
3	0.725	1.963
96	0.713	2.466
78	0.600	5.023
8	0.646	3.733
4	0.740	2.109
56	0.710	2.570
12	0.743	1.945
101	0.683	2.754
86	0.722	2.606
44	0.690	3.170
97	0.599	4.989
43	0.729	2.786
6	0.679	3.221
100	0.665	3.750
22	0.794	1.614

pirically fitted equation to represent the cumulative type-frequency curves. Since it is reasonable to assume that, for any individual, there is a limit to the number of types at his command, it was felt that an equation of the hyperbolic form would best agree with the character of the phenomena at hand, inasmuch as the hyperbolic curves are characterized by asymptotes, which can be correlated to the limit of the writing vocabulary of the individuals in similar situations. However, it was found that the data could not be satisfactorily reduced to a linear form of the hyperbolic curve, at least not a simple hyperbolic curve with two parameters. Possible reasons for this failure will be discussed later.

Of the attempts to fit curves with simple equations to these data, only a plot of $\log D$ and $\log N$ resulted in what could be considered a linear relationship. The resulting linear function is of the form

$$\log D = a \log N + \log b$$

in which a and b are empirically determined constants. If the above equation

TABLE 12

Results of analysis of variance of parameter a for groups of I.Q., C.A. and type-token ratio

Factor	Sums of Squares	d.f.	Variance	F
I.Q.	0.004449	2	0.002225	1.642
C.A.	0.010095	2	0.005048	3.723
TTR	0.015547	1	0.015547	11.474*
Error	0.016261	12	0.001355	
Total	0.046352	17		

Factor	Mean	Difference	t
I.Q.			
89 and under	0.669		
90 to 109 (inc.)	0.707		
110 and over	0.693		
C.A.			
149 mo. and under	0.656		
150 to 179 mo. (inc.)	0.709		
180 mo. and over	0.704		

TTR for 3,000 words

Less than .231	0.660		
Greater than .231	0.719	0.059	3.411*

* Significant at the one per cent level of confidence.

holds, then $\log D$ is a linear function of $\log N$, and D is a power function of N , of the general form

$$D = bN^a.$$

Equations of this form were fitted to the cumulative type-frequency data for the 18 language samples used in fitting equation (2). Estimates of parameters a and b arrived at by means of a least squares solution of the $(\log N, \log D)$ plot, are presented in Table 11. A graphical representation of this relationship for a typical plot of $(\log N, \log D)$ and (N, D) is shown in Figure 2. The curve presented in Figure 2, is typical of the other fitted curves in that the fit for larger values of N is not satisfactory and would make prediction beyond the limits of these data rather hazardous.

In order to determine what relationship exists between the estimates of these parameters and the factors of I.Q., C.A. and 3,000-word segmental type-token ra-

TABLE 13

Results of analysis of variance of estimates of parameter b for groups of I.Q., C.A., and type-token ratio

Factor	Sums of Squares	d.f.	Variance	F
I.Q.	1.211853	2	0.605927	—
C.A.	1.220563	2	0.610284	—
TTR	1.606827	1	1.606827	—
Error	12.739341	12	1.061612	—
	16.778584	17		

Factor	Mean
I.Q.	
89 and under	2.946
90 to 109, inclusive	2.620
110 and over	3.255
C.A.	
149 months and under	3.279
150 to 179 months, inclusive	2.646
180 months and over	2.895
TTR for 3,000 words	
0.231 and less	3.239
0.232 and over	2.641

tio, three levels each of I.Q. and C.A. as previously defined and two groups, a low and a high, categorized according to the magnitude of the 3,000-word segmental type-token ratio, were subjected to an analysis of variance. The results of the analysis of variance of these two parameters for these groups is presented in Tables 12 and 13. The results of the analysis for parameter a indicates that only the type-token ratio factor results in a significant F-value. This result is to be expected since the type-token ratio is a function of D , one of the variables in the equation. Inasmuch as the constant a determines the rate at which new words are added, it is not surprising that the difference in means between the low and high 3,000-word TTR groups should result in a significant difference, when tested by means of the t-test, in favor of a greater magnitude in the value of a for the group with a larger type-token ratio. The I.Q. and C.A. factors show no systematic differences be-

TABLE 14

Estimate of parameters, a and b in the equation $D = bN^a$ for 3,000-word sections and for the total sample of the 18,000-word language sample

Sample Section	a	b
Total 18,000 Words	0.697	3.499
First 3,000 Words	0.729	2.825
Second 3,000 Words	0.752	2.239
Third 3,000 Words	0.785	1.694
Fourth 3,000 Words	0.752	2.193
Fifth 3,000 Words	0.712	2.884
Sixth 3,000 Words	0.680	3.236

tween means of the various groups that are of great enough magnitude to give significant F-values. Since these factors have been shown to be associated in systematic manner to the 3,000-word segmental type-token ratio the marked differences between the two lower levels of I.Q. and C.A. might presumably be attributed to this association.

The results of the analysis of variance of estimates of parameter b for these groups indicate no significant factors. Apparently the exponent of N , i.e., the parameter a , is more influential in determining the differentiating characteristic of the curves than is the co-efficient of N , i.e., the constant b .

In order to determine if the power function will hold beyond 3,000 words, a curve was fitted to the data of the 18,000-word sample. The sample was divided into six 3,000-word samples and a curve fitted to each 3,000-word section as well as to the total sample. Estimates of the parameters of the power function for the total sample and for each section is presented in Table 14 and a graphical representation of the relationship for the 18,000 words is presented in Figure 3. Again, it is noted that the fit for the larger values of N , beyond N equal about 14,000, is poor. The empirical points diverge considerably from the curve.

FIG. 3.—A graphic representation of the relationship of the number of different words (D) to the sample size (N) and of $\log D$ to $\log N$ for the 18,000-word language sample. Empirical points are shown in their relation to the curve described by the indicated equation.

A second empirically fitted curve was derived by generalizing equation (1). A transformation of equation (1) can be made by writing the function as follows,

$$\log_e N = -K \left(\frac{D}{N} \right) + (\log_e K + K + 0.423), \quad (7)$$

which, since the terms in the right-hand bracket are all constants can be written as,

$$\log_e N = -K \left(\frac{D}{N} \right) + C. \quad (8)$$

Equation (8) is a more general function,

of which equation (7) is one special instance. In this form equation (8) is seen to be linear in $(D/N, \log_e N)$. Thus, if a plot of the variables D/N and $\log_e N$ can be considered as linear, we can proceed to make a least squares solution and get estimates of K and C .

We note, further that if Carroll's formulation of the relationship, i.e., equation (1), is to hold the two parameters of the fitted equation must have a definite relationship. Equation (1) makes the parameter C the following function of K ,

$$C = \log_e K + K + 0.423. \quad (9)$$

In this way we have a further test as to the adequacy of Carroll's formulation, one that is more satisfactory since it uses all the data in the language sample rather than a few selected points along the cumulative type frequency curve. If equation (9) is found to be tenable, we have direct evidence in substantiation of equation (1); if not, we can substitute an empirically determined curve of the same type as equation (1) but of a more general nature, as follows:

$$D = \frac{N}{K} (C - \log_e N). \quad (10)$$

In order to test the adequacy of equation (8), an average value at each successive 100-word point along the cumulative type-frequency curve was computed for each of the I.Q. levels of the experimental design. There were 36 language samples at each I.Q. level. An average series of points was felt desirable in order to give the empirical curve a greater stability at each point and also to smooth out chance fluctuations along the curve. The I.Q. levels were chosen because the variable D more clearly differentiated I.Q. levels than C.A., locality or sex levels.

A graphical representation of the rela-

tionship between $(D/N, \log_e N)$ and (D, N) , as well as the best fitting curve for each of the three I.Q. levels, is presented in Figure 4. It can be seen from Figure 4 that the reduction curves are reasonably linear, and that the fit of the curves to the empirical points appears to be good. The empirically derived equations for these three curves are:

(1) for the group composed of I.Q. 89 and under,

$$\log_e N = -9.910 \frac{D}{N} + 9.853 \quad \text{or}$$

$$D = \frac{N}{9.910} (9.853 - \log_e N)$$

(2) for the group composed of I.Q. 90 to 109, inclusive,

$$\log_e N = -10.081 \frac{D}{N} + 10.272 \quad \text{or}$$

$$D = \frac{N}{10.081} (10.272 - \log_e N)$$

(3) for the group composed of I.Q. 110 and over,

$$\log_e N = -9.551 \frac{D}{N} + 10.321 \quad \text{or}$$

$$D = \frac{N}{9.551} (10.321 - \log_e N)$$

K -values computed from equation (1) for these three curves at $N = 3,000$ are 6.97, 7.24, and 7.40, respectively. Further, the restriction basic to equation (1), that

$$C = K + \log_e K + 0.423$$

does not appear to be fulfilled. On the basis of these results, we would be compelled to reject the generality of equation (1) to these data.

A plot of $(D/N, \log_e N)$ was also made for the 18,000-word sample. The plot of $(D/N, \log_e N)$ and (D, N) as well as the best fitting curves in each instance, is presented graphically in Figure 5. Apparently the function described above represents the data reasonably well

FIG. 4.—Graphic representation of the relationship of the number of different words (D) to the language sample size (N) and of the ratio D/N to $\log_e N$. Plotted for each I.Q. level by averaging D at each successive 100-word point along the N -axis for each of the 36 subjects in that I.Q. level. Empirical points are shown in their relation to the curve described by the indicated equation.

throughout the 18,000 words. The empirically determined equation for this long sample is,

$$\log_e N = -11.268 \frac{D}{N} + 11.670 \quad \text{or}$$

$$D = \frac{N}{11.268} (11.670 - \log_e N)$$

Equations (1) and (10) have similar properties. Inasmuch as in equation (1) the equivalent of parameter C is a function of K , the maximum point on the curve is also a function of K . However, in the empirically determined curve; the parameter C is independent of the con-

FIG. 5.—A graphic representation of the relationship of the number of different words (D) to the language sample size (N) and of the ratio D/N to $\log_e N$ from the data computed from the 18,000-word language sample. Empirical points are shown in their relation to the curve described by the indicated equation.

stant K , and it is noted that the maximum point on the curve is a function of C and not K . For any given curve of this form, the maximum point on the curve is reached when

$$\frac{N}{\log_e N} = \frac{E_{(C-1)}}{C-1} \quad \text{or}$$

If the value of N is greater than $E_{(C-1)}$, computations of D for these values of N are spurious.

V. DISCUSSION

Crucial to the interpretation of the results of this study is the manner of selection and definition of a unit of language. Any one of several language units might have been chosen—syllable, word, phrase, or clause, among others. The simplest and least ambiguous unit to define is the word. For this among other reasons, the unit of language used in this study is the word, and it is generally referred to as a *token* in order to differentiate it from a unit of vocabulary, the *type*. This definition of a language unit is in conformance with that used by Zipf (15), Carroll (3), Fairbanks (5), Mann (9), and Fóssum (6) in studies which are comparable to this one. The word, as a unit of language, results in a statistical measure which involves fundamentally simple enumeration. All words are given equal weight in the determination of relationships. It may be questioned whether such a definition of language is entirely satisfactory, since it is known that a large proportion of words in connected discourse must necessarily be made up of the structural or interstitial words which represent relationships among other words and, in and of themselves, carry no meaning independent of the immediate verbal context in which they appear. On this basis, any sample of language may be divided into at least two classes of words, (1) the structural words and (2) the content words. It is possible, though less practicable, to give each word a weight in accordance with its classification in the above terms, or in accordance with its frequency of use in some standard language sample, and thus derive a more suitable measure of language. In any event, the procedure used will, to some degree, affect the character of the results obtained.

The classification of word units ac-

cording to certain rules and the counting of the resulting classes results in a measure of the number of types, or what may be appropriately called the vocabulary of the sample. Here again we note that our measure is a statistical one, in that it involves classification and enumeration. Once more, each type is given an equal weight in arriving at the numerical value of the measure, regardless of the function each type plays in the language structure. In this instance, however, greater liberty is given us in setting up our classes. What is psychologically the more fruitful method of selecting these classes can only be speculated upon. It might be argued that '*fall*', '*fall in*', '*fall out*', '*fall short*', '*fall apart*' should each be classified as a unit rather than as two units on the grounds that each represents a unitary symbol. Such an argument is undoubtedly valid, and if such a procedure were followed it might conceivably alter the results.

The language measures used in this study represent but a fraction of those already in use or that have been suggested for use. Sanford (11) in a recently published investigation demonstrated the utility of some 234 language measures which he used to describe personality differences between two individuals whose language he investigated. Busemann (2) has suggested and used the adjective-verb quotient for the study of personality. Boder (1), following Busemann's suggestion, has also employed the adjective-verb quotient, this time in the study of various types of literature. Mann (9) also applied this measure in comparing schizophrenic patients with university freshmen; she also employed an adjective-noun quotient and an adverb-verb quotient. Johnson (7) has suggested several language measures, among which are included the ones used in this

study. Measures of language, not in terms of word counts, but in terms of themes and topics have been investigated by Skinner and his co-workers (12). This enumeration will give the reader at least a suggestion of the abundance of language measures already in use or proposed. Many others, of course, are possible. The ones used in the present study by no means exhaust the possibilities of language analysis.

Further qualifications of the present results arise with regard to the fact that pertinent determining factors may have been omitted from consideration. Two such factors may be (1) number of topics discussed by each individual and (2) rate of verbal output per unit of time. With regard to the first of the above factors, if we are permitted to assume that, other conditions being equal, the number of types in a sample of language is positively correlated to the number of topics discussed in that sample, then it may appear plausible that, insofar as the more intelligent individuals can sustain a discussion on one topic for a greater number of words and, thus, for a specified number of words discuss fewer topics than less intelligent individuals, who of a necessity must shift topics more frequently to write a given quota, the relationship of intelligence to the number of types will be somewhat attenuated unless the number of topics is given some weight in the determination of this relationship. Again, it seems likely that age differences may be accentuated due to a wider range of interests, ambitions, opportunities, etc. of the older children. The behavior of these language measures within selected topics or fields of writing, as for instance in fiction and scientific writing, may be profitably investigated. In any event, consideration of this factor of number and type of topic,

perhaps by an analysis of co-variance technique, will broaden our understanding of language in terms of these measures.

The second factor mentioned above, namely, rate of verbal output per unit of time, might be investigated by somehow weighting the number of types produced by reference to the rate of verbal output. The analysis of co-variance technique is an appropriate method of carrying out such a weighting. Fossum (6), for spoken language, reported a negative correlation of $-.45$ of 100-word type-token ratio with rate of verbal output per unit of time. Whether the relationship, if any, is in the same direction for written language is yet to be determined.

These two above-mentioned factors do not exhaust all of the possibly pertinent factors which may need to be controlled, although they do offer perhaps the greatest promise of successful manipulation. Psychological factors such as motivation, interests, attitudes, ambitions, emotional states, etc. which are admittedly more difficult to control, nevertheless may prove to be significant determining factors in differentiating individuals in terms of these language measures. It was noted that in some instances children who expressed a dislike for the task of writing 3,000 words seemed to show a tendency to write in short jerky sentences with many of the sentences beginning with the same pattern of words, producing a stereotyped effect. On the other hand, children who expressed interest and a liking for this task tended to keep their discussion varied throughout the manuscript and thus probably produced a greater number of types than they would have if the motivation had been less adequate. For example, the boy who produced the greatest number of types in the present group of chil-

dren expressed a desire to be a writer of western fiction. Additional factors that may need to be given consideration in studying language in terms of these measures are: (1) differential effects of fatigue; (2) physiological conditions; (3) socio-economic background in the immediate family sense; (4) season of the year; (5) personality variables; etc. This enumeration certainly does not include all the possibly influential factors and serves only to demonstrate the complexity of the problem.

One implication of the results of this study is that the segmental type-token ratios based on successively larger segments are for practical purposes equivalent measures, insofar as they differentiate among individuals. This implication is inferred from the fact that the intercorrelation of these segmental type-token ratios are uniformly high and linear. However, there are instances in which the value of the type-token ratio for 100-word segments places individuals near the top of the group for this measure while for these same individuals the value of the 3,000-word type-token ratio places them near the bottom of the group. Such individuals would seem to be making efficient use of the vocabulary available to them, and if such an interpretation is justified, development of the notion of efficiency of vocabulary usage should result in fruitful research.

Certain implications arise from the curve-fitting aspects of the present study. In general, tests of the applicability of the equation presented by Carroll (3) point toward a rejection of the generality of this equation to these data. One possible reason for this may lie in the fact that the language samples used in this investigation were much different in certain characteristics from the language samples used by Carroll. The main

points of difference are: (1) one type of language sample used by Carroll consisted of the verbal output of a group of subjects which was combined into one unit, while in this study each language sample represents the performance of but one child. It may well be that the aforementioned equation will hold for the first type of language sample but not for the second; (2) the other type of language sample used by Carroll was obtained from the field of literature, and the individuals who produced the writing probably represented a highly selected group of verbally skilled individuals. In this study the verbal output of 'normal' Iowa school children, who are comparatively unskilled in language arts, was the object of investigation; (3) great differences in age, intelligence and environmental background undoubtedly existed between the two groups of subjects who produced the language studied in the two investigations. There is the possibility that, in general, the equation presented by Carroll will hold for some types of language samples but not for others.

On the other hand, there may be some question as to whether the notion involved in this equation is too narrow for psychological utility. Although an equation with only one parameter may be found to be adequate to describe the relationship between D and N , it would appear to be highly unlikely that the rate at which new words are added in connected discourse is so simply explained. The factor of number of topics, for example, has already been discussed as a possible factor influencing the rate at which new words are added, and probably other factors are also to be considered.

Attempts to fit a curve to represent the relationship between D and N reveals

that the function is of no simple nature. The study of the relationship between frequency of occurrence of a word and its rank demonstrates the important fact that the more frequently occurring words do not fit the same general function as do the rarer words. Considerations of the language structure suggests that perhaps it may be feasible to divide the words of any language sample into two categories. First, there are the interstitial or structural words which form the core of framework structure of language. Since these words carry little meaning beyond the verbal context in which they appear they may be termed intensional words in contrast to the second type of words, the content and action words that have, directly or indirectly, an extensional reference. Since the more frequently occurring words are of the intensional type, we have a basis for making an analysis of the language into two parts. In view of the fact that the extensional words serve to represent or are symbolic of the interests, attitudes, ambitions, etc. of the writer they would appear to be of greater psychological interest. There is some suggestion that if such a division of words could be made, a hyperbolic equation could be fitted to each division of words in the sample. One of the vitiating factors in attempts to find a lawful relationship between D and N may well lie in the difference between the ways in which these two types of words reach their maximum. The intensional words appear to reach a maximum in terms of D very rapidly, while the extensional words rise comparatively much more slowly and reach a maximum at a much later point on the curve.

VI. SUMMARY AND CONCLUSIONS

Three-thousand-word written language samples were obtained from 108

Iowa school children who had been selected to fill the cells of a factorial design which consisted of three levels each of I.Q., C.A., and locality (city, town, rural) as well as two equal groups of boys and girls. The subjects were asked to write about whatever they wanted to write about and in a free-writing situation for a short time each day until they had reached their quota of 3,000 words.

Each language sample was edited and the words tabulated according to a set of predetermined rules. From these tabulations, 20 language measures were obtained for each sample, and individual cumulative type-frequency curves were computed for a selected group of subjects. Language measures obtained were: (1) 100-word type-token ratio, (2) 500-word type-token ratio, (3) 1,000-word type-token ratio, (4) 3,000-word type-token ratio; number of tokens for (5) nounal, (6) verbal, (7) adjectival, (8) adverbial categories; number of types for (9) nounal, (10) verbal, (11) adjectival, (12) adverbial categories; type-token ratio for (13) nouns, (14) verbs, (15) adjectives, (16) adverbs; percentage of (17) nounal, (18) verbal, (19) adjectival, (20) adverbial types of the total types of these four parts of speech categories.

These data were analyzed in three ways in order to determine (a) the reliability of type-token ratios; (b) the ability of these measures to differentiate groups of individuals classified according to levels of I.Q., C.A., locality and sex; and (c) the mathematical relationship, if any, between the number of different words (D) and the size of the sample (N).

On the basis of these analyses the following conclusions can be drawn:

1. Segmental type-token ratios derived from samples of 3,000 words are highly reliable in (a) the agreement between

two independent sets of operations used to arrive at the numerical value of the type-token ratio and in (b) the relative constancy of the type-token ratio over a short span of time (about a week). A further implication, indicated by the fact that the reliability coefficient is a positive function of the size of sample, is that, in general, type-token ratios computed from a sample of 1,000 words in length are, for practical purposes, as reliable as are those computed from samples of 3,000 words in length and should prove satisfactory in all instances except when a high degree of precision is needed.

2. The results of the analyses of variance of the 20 language measures may be summarized briefly in the statement that the implication of these results is that the language measures employed can be used to characterize groups classified according to I.Q., C.A., locality and possibly sex, although the results for sex are practically negative. On the whole,

the more highly developed the individual in terms of intelligence and age, the more highly differentiated his language structure appears to be. This is shown particularly by the type-token ratios. It may also be said that high I.Q. groups are characterized by the use of a proportionately greater number of nouns while the low I.Q. groups are characterized by the use of a greater percentage of verbs and adverbs.

3. An equation presented by Carroll was discussed and tests of its adequacy to describe these data were carried out. Its generality to these data was not substantiated, although a more general form of the same equation was found to give a fairly good fit. The relationship of the number of different words to the size of sample was found to be a complex one. Empirically fitted equations to represent these data are of such a character as to make prediction beyond the limits of the data hazardous.

BIBLIOGRAPHY

1. BODER, D. P. The adjective-verb-quotient: a contribution to the psychology of language. *Psychol. Rec.*, 1940, 3, 310-343.
2. BUSEMANN, A. *Die Sprache der Jugend als Ausdruck der Entwicklungsrhythmik*. Jena: Fischer, 1925, pp. x + 98.
3. CARROLL, J. B. Diversity of vocabulary and the harmonic series law of word-frequency distribution. *Psychol. Rec.*, 1938, 2, 379-386.
4. CURME, G. O. *Parts of Speech and Accidence*. Boston: D. C. Heath and Co., 1935.
5. FAIRBANKS, H. Studies in language behavior: II. The quantitative differentiation of samples of spoken language. *Psychol. Monogr.*, 1944, 56, 19-38.
6. FOSSUM, E. C. An analysis of certain characteristics of oral language of junior college students. Univ of Iowa Ph.D. Dissertation. 1941.
7. JOHNSON, W. Studies in Language Behavior: I. Objectives and methods. *Psychol. Monogr.*, 1944, 56, 1-15.
8. KENT, G. H., & ROSANOFF, A. J. A study of association in insanity. *Amer. J. Insanity*, 1910, 67, 57-96.
9. MANN, M. B. Studies in Language Behavior: III. The quantitative differentiation of samples of written language. *Psychol. Monogr.*, 1944, 56, 41-74.
10. SANFORD, F. H. Speech and personality. *Psychol. Bull.*, 1942, 39, 811-845.
11. SANFORD, F. H. Speech and personality: a comparative case study. *Character & Pers.*, 1942, 19, 169-198.
12. SKINNER, B. F. The verbal summator and a method for the study of latent speech. *J. Psychol.*, 1936, 2, 71-107.
13. SKINNER, B. F. The distribution of associated words. *Psychol. Rec.*, 1937, 1, 71-76.
14. THORNDIKE, R. L. On the number of words of any frequency of use. *Psychol. Rec.*, 1937, 1, 379-406.
15. ZIFF, G. K. *The Psychobiology of Language*. Boston: Houghton Mifflin Co., 1935, pp. ix + 336.