QUANTITATIVE LINGUISTICS

Vol. 3

GLOTTOMETRIKA 2

edited

by

R. GROTJAHN

Studienverlag Dr. N. Brockmeyer

Bochum 1980

QUANTITATIVE LINGUISTICS

Editor R. Grotjahn, Bochum

Editorial Board: ND. Andreev, Leningrad
M.V.Arapov, Mocsow
B. Brainerd, Toronto
R. Grotjahn, Bochum
H. Guiter, Montpellier
D. Hérault, Paris
E. Hopkins, Bochum
W. Lehfeldt, Konstanz
W. Matthäus, Bochum
R.G. Piotrowski, Leningrad
B. Rieger, Aachen/Amsterdam
J. Sambor, Warsaw
D. Wickmann, Aachen

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Glottometrika. - Bochum: Studienverlag Brockmeyer.
2. / Ed. by R. Grotjahn. - 1980. - 218 S.
(Quantitative linguistics; Vol. 3)

NE: Grotjahn, Rüdiger (Hrsg.)

ISBN 3-88339-104-2

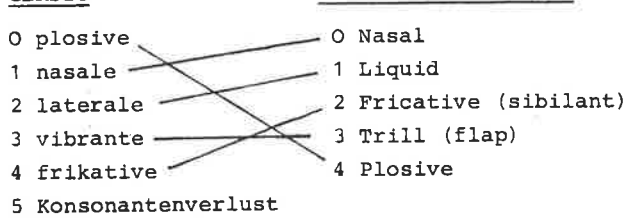
Querenburger Höhe 281
4630 Bochum 1

CONTENTS

E R R A T A

in GLOTTOMETRIKA 1

- In: Lehfeldt,W., Zur Messung der phonetischen Lautdifferenz. Eine begriffskritische Untersuchung.
Page 42, bottom: In the diagram the connecting lines are missing; the diagram should appear as follows:



- In: Rodrigue-Schwarzwald,O., The influence of demographic variables on linguistic performance.
Pp. 186 and 187 have been transposed.

Prolegomena to Menzerath's Law

G. Altmann, Bochum

1. Historical

The first step in the discovery of "MENZERATH's law" was made in 1928 when MENZERATH ascertained that "... a sound is the shorter the longer the whole in which it occurs (the law of quantity)" and "... the more sounds in a syllable the smaller its relative length" (MENZERATH 1928: 104). Formulated in another way, the duration of articulation of sounds shortens in long syllables.

Later on, analyzing German words, MENZERATH arrived at the following conclusion: "The relative number of sounds in the syllable decreases as the number of syllables in the word increases, or said in a different way: the more syllables in a word the shorter (relatively) it is" (MENZERATH 1954: 100). In other words, the longer the word the shorter its syllables.

From these two empirical findings in German there arise several tasks: (i) that of formulating general hypotheses in which no observational concepts (such as sound, syllable, word) occur; (ii) that of validating them theoretically, i.e. their derivation from plausible assumptions or their integration into a system of laws, respectively; (iii) that of testing them empirically on different languages and language entities; (iv) that of examining their possible consequences. Since all these tasks together can involve lengthy research, we can here merely programmatically discuss the problems mentioned.

2. Theoretical derivation

A more general statement including both statements of MENZERATH as special cases would be:

"The longer a language construct the
shorter its components (constituents)" (1)

or, in MENZERATH's concise wording: "The greater the whole the smaller its parts" (1954: 101). In the simplest mathematical formulation of this statement one could assume a constant decrease rate of the length of the component, i.e. to put

$$\frac{y'}{y} = -c$$

or in other words, to put the shortening proportional to the length, i.e. y' = -cy. Integration yields

$$y = ae^{-cx}. \qquad (2)$$

This result means that the length of the component (Y) is a monotonic decreasing function of the length of the construct (X) (cf. ALTMANN 1978).

One problem is, however, whether one should consider as a component only the immediately lower unit, e.g. the clause as the immediate component of the sentence, the syllable as the immediate (phonetic) component of the word etc., or other units too, e.g. the word as the component of the sentence, the phoneme/sound as the component of the word, etc. In the second case the simple relation (1) or (2) will be somewhat problematic. Similar problems arise in languages having nonsyllabic words, i.e. words consisting merely of one or more consonants but no vowels (and no syllabic consonants). Such words as e.g. the prepositions *k, v, s* in Russian belong phonetically as proclitics to the next word and contribute to its length; if one considers them as independent words of zero syllabic length then eo ipso the length of the syllable in such words is zero. In such cases the curve will not decrease monotonically but will reach a maximum at $x \neq 0$. The problem can be further complicated by the fact that the position of the construct in higher constructs or the position of the component in the construct may produce an effect, too. Therefore there is a possibility that some construct-component dependencies will not decrease monotonically.

As a refinement we can therefore tentatively assume that besides the constant decrease for "unambiguous cases" (-c) a

further member including an inverse proportionality of the decrease rate to the construct length can be added, i.e.

$$\frac{y'}{y} = -c + \frac{b}{x}. \qquad (3)$$

The hypothesis can be now formulated still more generally as

"The length of the components is a function of the length of language constructs" (4)

and this function can be given as a solution of the differential equation (3). According to how b and c are determined we obtain the particular functions given in Table 1. Presently we do not know how the individual curves are to be assigned to the individual construct-component relations and we do not know what the particular coefficients depend on. This will become apparent after a satisfactory empirical testing of the curves.

Table 1. Solutions of the differential equation (3)

$b = 0$		$y = Ae^{-cx}$	I
$b \neq 0$	$c = 0$	$y = Ax^{b}$	II
	$c \neq 0$	$y = Ax^{b}e^{-cx}$	III

3. Fitting

In order to make the fitting of the curves easier and thereby to stimulate research we bring a simple guide to the computation of the coefficients of the functions I to III in Table 1.

Function I. A straight line can be obtained by means of

$$\ln y = \ln A - cx,$$

therefore

$$\ln A = \frac{\sum_i x_i^2 \sum_i \ln y_i - \sum_i x_i \sum_i x_i \ln y_i}{n \sum_i x_i^2 - (\sum_i x_i)^2}$$

and

$$-c = \frac{n \sum_i x_i \ln y_i - \sum_i x_i \sum_i \ln y_i}{n \sum_i x_i^2 - (\sum_i x_i)^2} .$$

Here i=1.2,...,n where n is the number of observation; ln means always the natural logarithm.

Function II. A straight line can be obtained by means of

$$\ln y = \ln A + b \ln x,$$

therefore

$$\ln A = \frac{\sum_i (\ln x_i)^2 \sum_i \ln y_i - \sum_i \ln x_i \sum_i \ln x_i \ln y_i}{n \sum_i (\ln x_i)^2 - (\sum_i \ln x_i)^2}$$

and

$$b = \frac{n \sum_i \ln x_i \ln y_i - \sum_i \ln x_i \sum_i \ln y_i}{n \sum_i (\ln x_i)^2 - (\sum_i \ln x_i)^2} .$$

Function III. A straight line can be obtained by means of

$$\ln y = \ln A + b \ln x - cx,$$

therefore

$$b = \left\{ \sum_i \ln x_i \ln y_i \left[n \sum_i x_i^2 - (\sum_i x_i)^2 \right] - \sum_i x_i \ln y_i \left[n \sum_i x_i \ln x_i - \sum_i x_i \sum_i \ln x_i \right] + \sum_i \ln y_i \left[\sum_i x_i \sum_i x_i \ln x_i - \sum_i x_i^2 \sum_i \ln x_i \right] \right\} / D$$

$$c = \left\{ \sum_i x_i \ln x_i \left[n \sum_i \ln x_i \ln y_i - \sum_i \ln x_i \sum_i \ln y_i \right] - \sum_i x_i \ln y_i \cdot \left[n \sum_i (\ln x_i)^2 - (\sum_i \ln x_i)^2 \right] - \sum_i x_i \left\{ \sum_i \ln x_i \sum_i \ln x_i \ln y_i - \sum_i (\ln x_i)^2 \sum_i \ln y_i \right] \right\} / D$$

$$\ln A = \frac{1}{n} (\sum_i \ln y_i - b \sum_i \ln x_i + c \sum_i x_i)$$

$$D = n \left[\sum_i x_i^2 \sum_i (\ln x_i)^2 - (\sum_i x_i \ln x_i)^2 \right] + 2 \sum_i x_i \sum_i \ln x_i \sum_i x_i \ln x_i - \sum_i x_i^2 (\sum_i \ln x_i)^2 - (\sum_i x_i)^2 \sum_i (\ln x_i)^2 .$$

4. Illustrations

The derivation from a differential equation is not sufficient in order to award the statement (4) the status of a law. It remains a theoretically not fully validated hypothesis as long as it is set in relation to other laws of language i.e. until it is incorporated into a system of laws. Such a system does not exist at present. We merely suspect that it is somehow connected with the principle of least effort or with some not yet known principle of balance recompensating lengthening on one hand with shortening on the other.

Empirical validation consists of testing the hypothesis on data from many languages. Here the concepts "construct" and "component" can be interpreted in many ways if one replaces them with more empirical concepts such as sound, syllable, morpheme, word, word group, compound, phrase, clause, sentence, etc. One must ask whether the coefficients of the curves depend on the given language or on the given units or whether the text influences the course of the curves, too.

There is no doubt that this "law" is a stochastic one. We do not assume the existence of cases without any deviations, even if very long texts or the complete dictionary of a language is examined. A final confirmation of the hypothesis is impossible because of its universality but every favourable case can improve its degree of confirmation.

The hypothesis will be exemplified below by means of certain selected cases.

(a) Syllable length of Indonesian morphemes. This somewhat unusual relation is measurable only because in Indonesian every vowel represents a syllable. As a matter of fact the relation means the decrease of the number of consonants with increasing morpheme length. The mean syllable length (measured in phonemes) in morphemes of different length is given in Table 2. The means were computed from 13444 morphemes. The coefficients of the fitted curve were determined from the data.

Table 2. Syllable length of Indonesian morphemes

Morpheme length (in syllables) x	1	2	3	4	5
Mean syllable length (in phonemes) y	3.10	2.53	2.29	2.12	2.09
y_c	3.10	2.52	2.27	2.15	2.08

Using curve (III) we obtain

$$y_c = 2.9603x^{-0.36853}\ e^{0.04764x}.$$

The computed values are in the third row of Table 2. The value of $F_{1,3} = 212.1$ is highly significant as compared with $F_{1,3(.05)} = 10.1$. When evaluating data of this kind one should take care to provide a sufficient number of instances for each construct length otherwise the higher values especially can considerably deviate from the "smooth" course of the curve. The identity of the individual syllables is irrelevant when ascertaining their mean length therefore no special method of syllable identification is necessary.

(b) Syllable length in English words. The computations were based on the data by ROBERTS (1965: 50) consisting of

15465010 words. The results are given in Table 3. The expected values computed according to

$$y_c = 3.06x^{0.14997}\ e^{-0.09968x}$$

are given in the third line. The shortening of the syllable with increasing word length is evident. The F-test yields $F_{2,4} = 129.3$ which is highly significant as compared with $F_{2,4(.05)} = 6.9$.

Table 3. Syllable length in English words (types)

Word length in syllables x	1	2	3	4	5	6	7
Mean syllable length in phonemes y	2.79	2.73	2.68	2.51	2.41	2.23	2.00
Computed values y_c	2.77	2.78	2.67	2.53	2.36	2.20	2.04

(c) Syllable duration in Bachka-German (presented with kind permission of S. Geršić). The measurement of syllable duration (in milliseconds) dependent on word length (measured in number of syllables) was performed on 488 words (with 3961 repetitions whose means were used for computation) of spoken texts of the German dialect of Bachka (Yugoslavia) and yielded the results given in Table 4. The values computed from

$$y = 269.96e^{-0.087735x}$$

are in the third row of the table. The F-test yields $F_{1,3} = 36.57$ as compared with $F_{1,3(.05)} = 10.1$.

For the time being it is not possible to bring the coefficients of the curves into relation to other entities of the examined languages.

Table 4. Syllable length in Bachka-German words

Word length in syllables x	1	2	3	4	5
Mean syllable duration in milliseconds y	239.91	240.92	204.28	183.95	177.06
Computed values y_c	247.28	226.51	207.49	190.06	174.09

5. Consequences

A lawlike statement can also be incorporated in a system in such a way that one shows the consequences which follow from it. Since the consequences themselves can acquire the status of laws the original statement can take the status of a high level law. Since at the moment lingistics is in need of laws the only possibility to corroborate the status of the examined hypothesis is to present (in a non formal way) some of its possible consequences. To this aim we present the following open list of hypotheses.

In the *phonetic domain* we assume that

(i) with increasing word length the sounds will be shortened. Complications can arise by means of the influence of sentence length, of the position of the word in the sentence etc. This hypothesis can in turn have some further consequences:

(ii) the circumstance that in relatively long words the sounds tend to be shortened can result in phonetic changes. Therefore the hypothesis seems plausible that in longer words more changes occur than in short ones (the frequency of use should be considered, too). The changes should more often concern consonants and long vowels, more seldom short vowels.

(iii) From (ii) it follows that in languages with a greater mean word length more phonetic/phonological changes occur during the same time interval (ceteris paribus) than in languages with a smaller mean word length.

(iv) Syllable length depends on word length. The curve need not necessarily decrease monotonically (cf. ex. b and c).

This hypothesis also has consequences for the *morphological domain*. If in longer words the syllables tend to be shortened then it is plausible to assume that

(v) if (monosyllabic) affixes with several consonants are added to a word then the inventory of consonants in the new word will be reduced. This phenomenon is well known from several languages.

(vi) When adding affixes or compounding two words epenthetic vowels without etymology are often inserted. Thereby the new word becomes longer but the mean syllable length decreases. This is a contrary process to that in (v).

(vii) The longer the word the shorter its morphemes, i.e. morpheme length depends "somehow" on word length. From this automatically follows that

(viii) partial reduplications must be more frequent in the languages of the world (not necessarily in a single language) than total ones and with partial reduplications usually a consonant is eliminated.

(ix) From (vii) also follows the hypothesis that short roots (morphemes) build more compounds than long ones.

(x) From (vii) follows mutatis mutandis that the more elements (roots) in a compound the shorter they are.

In the *domain of the sentence* we state two hypotheses, namely

(xi) in long sentences (measured in number of clauses) the clauses are shorter and vice versa otherwise the sentence looses its clearness.

(xii) Somewhat more problematic is the relation of sentence length to the word length. A monotonic decrease of word length can hardly be assumed.

These hypotheses are by no means the only possible ones, yet they are the most evident ones. Some of them will, perhaps, be rejected and other ones will be set up. The law evidently interferes with the development of language, too.

It is the task of the research (1) to test the hypotehses on as much data as possible, i.e. on texts as well as dictiona-

ries of many languages; (2) to examine the range of "MENZERATH's law" by setting up new hypotheses; (3) to specify the curves to particular data and to bring the coefficients into relation to other language phenomena; (4) to integrate "MENZERATH's law" into a system of laws or to develop the principles from which it follows.

References

ALTMANN, G., Towards a theory of language. In: ALTMANN, G. (Ed.), Glottometrika 1. Bochum, Brockmeyer 1978, 1-25

HALLE, M., The sound pattern of Russian. The Hague, Mouton 1959.

MENZERATH, P., Über einige phonetische Probleme. Actes du premier congrès international de linguistes. Leiden, Sijthhoff 1928, 104-105.

MENZERATH, P., Die Architektonik des deutschen Wortschatzes. Bonn, Dümmler 1954.

ROBERTS, A.H., A statistical linguistic analysis of American English. The Hague, Mouton 1965.

The Theory of Runs as an Instrument for Research in Quantitative Linguistics

Rüdiger Grotjahn, Bochum

0. *Introduction*

In the use of statistical methods in linguistics and also in many other disciplines usually only the *frequency* of the research units is taken into consideration. In many cases, however, the *order* of the units is also of interest for a quantitative analysis.

For example, if we toss a coin ten times, the following sequences are possible results of the experiment:

H H H H H T T T T T (a)

H T H T H T H T H T (b)

If we view only the frequency of occurrence of the elements H (heads) and T (tails), then for both cases $p_H = p_T = 0.5$ is valid, and the hypothesis of a random sequence with $p_1 = p_2 = 0.5$ must be accepted. If we look at the order of the elements, however, then both sequences seem to vary from a random sequence. In sequence (a) there is apparently a tendency toward repetition, in sequence (b), in contrast, a tendency toward alternation.

Analogous tendencies occur in many areas of language. Research in quantitative linguistics should therefore not stop with the analysis of the frequency of linguistic phenomena, but must also take order into consideration. Thus it is necessary in the framework of the analysis of verse, for example,

to inquire not only about the frequency of occurrence of stressed and unstressed syllables in a text, but also to test, with the help of statistical methods, whether a tendency toward ordering of the stressed and unstressed syllables can be recognized. It can be further studied whether verses of the same type occur in clusters, whether the sequence of short and long sentences in a text tends to be ordered, whether an alternation of consonants and vowels can be observed in a certain language, or, in the framework of psycholinguistic research, whether certain types of associations tend to cluster. There are many other interesting questions of this type.

In order to study these questions, we can use, for example, correlational analysis (e.g. FUCKS 1955; 1968), the theory of random clumping (cf. ROACH 1968; GREGER 1972), the model of Markov chains (cf. BRAINERD 1976; GROTJAHN 1979: 213-224) or the theory of runs. In the following we will specifically examine the theory of runs which until now has hardly been used in linguistic research.[1] Our goal is not primarily to present concrete results of analysis. We are mainly concerned with presenting the bases of this method relatively explicitly and understandably, and to give some examples of its possible use. It is hoped that the theory of runs can thus be made available to as many researchers in quantitative linguistics as possible.

2. *The Concept of Runs*

Central to the theory of runs is the concept of a run. The term 'run' is understood as a sequence of identical elements which either precede or follow sample elements of a different type. A run can be formally defined as follows (cf. FISZ 1970: 483 f):

Suppose we are given a sample with the values $x_1, x_2, \ldots, x_n$. The sequence $x_j, \ldots, x_{j+m}$ with $j=1,2,\ldots,n$ and $m=0,1,\ldots,n-j$ is called a run, if

$$x_{j-1} \neq x_j = \ldots = x_{j+m} \neq x_{j+m+1}.$$

Since m may equal 0, a run can also consist of a single element.

Frequently one differentiates between binary and multiple runs. In the case of binary runs, the sample only consists of two different types of elements, while in the case of multiple runs, there are more than two different types of elements. For example, consider the following sample with three different kinds of elements:

$$\underline{a_1\ a_1\ a_1}\ \underline{a_2}\ \underline{a_1}\ \underline{a_3\ a_3}\ \underline{a_2}$$

This sequence of $n=8$ elements consists of $n_1=4$ a_1-elements, $n_2=2$ a_2-elements and $n_3=2$ a_3-elements. It contains an a_1-run of three elements, an a_1 run of one element, two a_2-runs of one element and one a_3-run of two elements. The total number of a_1-runs is $r_1=2$, of a_2-runs $r_2=2$ and of a_3-runs $r_3=1$. The total number of runs in the sample is $r_1+r_2+r_3=r=5$.

Example 1.1: As an example of binary runs let us look at the sequence of stressed and unstressed syllables in the first two lines of Goethe's "Erlkönig":

Wer reitet so spät durch Nacht und Wind?
Es ist der Vater mit seinem Kind;

If the stressed syllable is s and the unstressed syllable u, the two lines correspond to the binary sequence

u s u u s u s u s u s u u s u s

with $n_1 = 8$ $n_2 = 10$ $n = n_1 + n_2 = 18$

$r_1 = 8$ $r_2 = 8$ $r = r_1 + r_2 = 16$

As could be expected in a text of verse of this kind, there seems to be a tendency toward alternation of stressed and unstressed syllables.

Example 1.2: As an example of multiple runs let us consider the number of syllables per line in "Erlkönig". The data are presented in Table 1.

Table 1: Number of Syllables (S) per Line (L) in "Erlkönig"

L-No.	S	L-No.	S	L-No.	S	L-No.	S
1	9	9	8	17	9	25	12
2	9	10	9	18	10	26	11
3	9	11	9	19	11	27	11
4	9	12	10	20	11	28	9
5	10	13	11	21	11	29	10
6	9	14	10	22	9	30	10
7	9	15	9	23	9	31	9
8	8	16	9	24	10	32	9

It is immediately apparent that among the 32 lines there are only lines with 8, 9, 10, 11, or 12 syllables and that these types of lines occur also with different frequencies. In addition, there seems to be a tendency toward ordering of the line types. Table 2 shows the number of runs of the different line types.

Table 2: Number of Runs of Line Types from Table 1

	Syllables					
	8	9	10	11	12	
n_i	2	16	7	6	1	n=32
r_i	1	7	6	3	1	r=18

2.1 The theory of runs is most frequently used to test the randomness of binary sequences. Since binary runs are much more simple to work with mathematically than multiple runs, let us first consider in the following derivation of the formulas for the distribution of runs[2] the case that the sample consists of a sequence of n_1 elements of type a_1 and n_2 elements of type a_2.

To find the joint probability function of r_1 and r_2, i.e.

$f(r_1,r_2)$, we must first establish the number of possibilities for dividing n_1 a_1-elements in r_1 runs. For example, if we look at the sequence $a_1a_1a_1a_1a_1a_1a_1$, the $n_1=7$ a_1-elements could be divided into, say, $r_1=4$ runs by choosing among the $n_1-1=6$ possible spaces $r_1-1=3$. In this way we arrive at the sequence $a_1a_1/a_1a_1a_1/a_1/a_1$. Hence there are

$$\binom{n_1 - 1}{r_1 - 1}$$

possibilities for dividing n_1 a_1-elements in r_1 runs. Similarly, we arrive at

$$\binom{n_2 - 1}{r_2 - 1}$$

possibilities for dividing n_2 a_2-elements in r_2 runs.

We must now differentiate between two cases.

(1) There are *different* types of runs at the beginning and at the end of the sample. In this case $r_1=r_2$, and the sample can either begin with an a_1-run and end with an a_2-run or begin with an a_2-run and end with an a_1-run. Thus for the case $r_1=r_2$, there are two further possibilities which must be taken into consideration in deriving the total number of possible arrangements of r_1 a_1-runs and r_2 a_2-runs.

(2) There is the *same* type of run at the beginning and at the end of the sample. In this case $|r_1-r_2|=1$ and either $r_1=r_2+1$ or $r_1=r_2-1$. If $r_1=r_2+1$, the sequence begins and ends with the element a_1. If $r_1=r_2-1$, the sequence begins and ends with the element a_2. Thus, the total number of possibilities remains the same.

If we consider further that the number of the possible arrangements of a sequence of the length $n_1+n_2=n$ is

$$\binom{n}{n_1} = \binom{n}{n_2},$$

then the probability function which we have looked for is

$$f(r_1,r_2) = \frac{c\binom{n_1-1}{r_1-1}\binom{n_2-1}{r_2-1}}{\binom{n}{n_1}} \qquad (1)$$

with $c = 2$ for $r_1 = r_2$
$c = 1$ for $r_1 \neq r_2$,

where the binomial coefficient $\binom{a}{b}$ for $a<b$ is defined to be zero.

For the application of the theory of runs, the distribution of the total number of runs plays a major role. We arrive at this in the following way:
If $r_1=r_2$, then $r_1=r/2=r_2$. If we substitute $r/2$ for both r_1 and r_2 in (1), we get

$$f(r) = \frac{2\binom{n_1-1}{\frac{r}{2}-1}\binom{n_2-1}{\frac{r}{2}-1}}{\binom{n}{n_1}} \qquad (2)$$

for r even.

For the case, that $r_1=r_2+1$ or $r_1=r_2-1$, then

$$r_1 = \frac{r+1}{2} \quad \text{or} \quad r_1 = \frac{r-1}{2}$$

and

$$r_2 = \frac{r-1}{2} \quad \text{or} \quad r_2 = \frac{r+1}{2} .$$

If we substitute accordingly in (1), we obtain

$$f(r) = \frac{\binom{n_1-1}{\frac{r-1}{2}}\binom{n_2-1}{\frac{r-3}{2}} + \binom{n_1-1}{\frac{r-3}{2}}\binom{n_2-1}{\frac{r-1}{2}}}{\binom{n}{n_1}} \qquad (3)$$

for r odd.
The computation of (3) is in many cases relatively cumbersome; it can be simplified, however, with the help of a recursion formula (see GROTJAHN 1979: 148).

SWED/EISENHART (1943) have computed tables for the cumulative distribution function of r, i.e. F(r), for $n_1,n_2\leq 20$ (printed for example in OWEN 1962: 377 ff). For $n_1,n_2>20$ the approximation by the normal distribution is usually used (see Part 2.3).

In order to find the distribution of the number of a_1-runs or a_2-runs regardless of whether $r_1=r_2$ or $r_1\neq r_2$, we use the marginal distributions of $f(r_1,r_2)$. For the derivation of $P(R=r_1)=f(r_1)$, we substitute r_1 for r_2 in (1). Again we must take into account that $r_2=r_1$ or $r_2=r_1+1$ or $r_2=r_1-1$:

$$\binom{n}{n_1}\cdot f(r_1) = 2\binom{n_1-1}{r_1-1}\binom{n_2-1}{r_1-1}+\binom{n_1-1}{r_1-1}\binom{n_2-1}{r_1-2}+\binom{n_1-1}{r_1-1}\binom{n_2-1}{r_1}$$

$$=\binom{n_1-1}{r_1-1}\cdot\left[\binom{n_2-1}{r_1-1}+\binom{n_2-1}{r_1-1}+\binom{n_2-1}{r_1-2}+\binom{n_2-1}{r_1}\right]$$

$$=\binom{n_1-1}{r_1-1}\left[\binom{n_2}{r_1-1}+\binom{n_2}{r_1}\right] = \binom{n_1-1}{r_1-1}\binom{n_2+1}{r_1}$$

Thus we get

$$f(r_1) = \frac{\binom{n_1-1}{r_1-1}\binom{n_2+1}{r_1}}{\binom{n}{n_1}} . \qquad (4)$$

Analogously we obtain

$$f(r_2) = \frac{\binom{n_1+1}{r_2}\binom{n_2-1}{r_2-1}}{\binom{n}{n_1}} . \qquad (5)$$

Until now we have only looked at the distribution of the number of runs for fixed n_i. Therefore we could also have written

$f(r_1,r_2|n_1,n_2)$, $f(r|n_1,n_2)$ and $f(r_2|n_1,n_2)$ in formulas (1) to (5). It is however also possible to view n_i as a random variable. For example, if we know the probability of occurrence of the elements of the sequence in the population, or if we can assume under the null hypothesis that $n_1/n = p = n_2/n = q$, then we can test with the theory of runs not only the number of runs, given n_i, but we can also inquire about the probability that, say, n_1 a_1-elements and n_2 a_2-elements occur in the sequence and that they are arranged in a certain way. Since the probability of the occurrence of two events A and B is given by

$$P(AB) = P(A|B)\cdot P(B),$$

we obtain the formulas for such joint probabilities simply by multiplying the corresponding conditional probability functions - in our case the formulas (1) to (5) - with the binomial probability

$$P(n_1,n_2) = \binom{n}{n_1} p^{n_1} q^{n_2} .$$

The resulting probability functions are:

$$f(r_1,r_2,n_1,n_2) = c\binom{n_1 - 1}{r_1 - 1}\binom{n_2 - 1}{r_2 - 1} p^{n_1} q^{n_2}$$

with $c = 2$ for $r_1 = r_2$
$c = 1$ for $r_1 \neq r_2$ (6)

$$f(r,n_1,n_2) = 2\binom{n_1 - 1}{\frac{r}{2} - 1}\binom{n_2 - 1}{\frac{r}{2} - 1} p^{n_1} q^{n_2} \qquad (7)$$

for $r_1 = r_2$

$$f(r,n_1,n_2) = \left[\binom{n_1-1}{\frac{r-1}{2}}\binom{n_2-1}{\frac{r-3}{2}} + \binom{n_1-1}{\frac{r-3}{2}}\binom{n_2-1}{\frac{r-1}{2}}\right] p^{n_1} q^{n_2}$$

for $r_1 \neq r_2$ (8)

$$f(r_1,n_1,n_2) = \binom{n_1 - 1}{r_1 - 1}\binom{n_2 + 1}{r_1} p^{n_1} q^{n_2} \qquad (9)$$

$$f(r_2,n_1,n_2) = \binom{n_1 + 1}{r_2}\binom{n_2 - 1}{r_2 - 1} p^{n_1} q^{n_2} \qquad (10)$$

The distributions which have been derived so far can be generalized for the case of k different types of elements (see MOOD 1940; DAVID/BARTON 1962: 119 ff). The exact distributions are, however, as MOOD (1940) himself states, not very useful in practice. Therefore in the case of multiple runs, the normal approximation is usually used.

2.2 *Moments*

In order to be able to use the normal approximation, we need the mean and the standard error of the number of runs. First we consider the case where the n_i are given, i.e. we compute the conditional moments. To simplify the notation, we do not indicate the conditions in the following formulas.

The expectations of r_1 and r_2 are relatively simple to derive from the factorial moments (see WILKS 1962: 148).[3] For r_1 the gth factorial moment is

$$E(r_1^{[g]}) = \sum_{r_1=g}^{n_1} r_1^{[g]} \frac{\binom{n_1 - 1}{r_1 - 1}\binom{n_2 + 1}{r_1}}{\binom{n}{n_1}}$$

where $r_1^{[g]} = r_1(r_1 - 1) \ldots (r_1 - g + 1)$.

Using the identity

$$\sum_{r_1=g}^{n_1} \binom{n_1 - 1}{r_1 - 1}\binom{n_2 + 1 - g}{r_1 - g} = \binom{n - g}{n_1 - g}$$

we get

$$E(r_1^{[g]}) = \frac{(n_2+1)^{[g]}\binom{n-g}{n_1-g}}{\binom{n}{n_1}} . \qquad (6)$$

In the same way we arrive at

$$E(r_2^{[g]}) = \frac{(n_1+1)^{[g]}\binom{n-g}{n_2-g}}{\binom{n}{n_2}} . \qquad (7)$$

It follows from (6) and (7)

$$E(r_1) = \frac{n_1(n_2+1)}{n} \qquad (8)$$

$$E(r_2) = \frac{n_2(n_1+1)}{n} \qquad (9)$$

$$V(r_1) = E(r_1^{[2]}) + E(r_1) - (E(r_1))^2$$

$$= \frac{n_1(n_1-1)\, n_2(n_2+1)}{n^2(n-1)} \qquad (10)$$

$$V(r_2) = E(r_2^{[2]}) + E(r_2) - (E(r_2))^2$$

$$= \frac{n_2(n_2-1)\, n_1(n_1+1)}{n^2(n-1)} . \qquad (11)$$

To derive the covariance of r_1 and r_2, we use the relation

$$COV(r_1,r_2) = E(r_1r_2) - E(r_1)E(r_2) . \qquad (12)$$

For computing $E(r_1r_2)$ we again use the method of the factorial moments. Analogous to the computation of formula (6) and with the help of (1) we get

$$E\left[(r_1-1)^{[g_1]}(r_2-1)^{[g_2]}\right] = \frac{(n_1-1)^{[g_1]}(n_2-1)^{[g_2]}}{\binom{n}{n_1}} \binom{n-g_1-g_2}{n_1-g_2} \qquad (13)$$

where $(x-1)^{[g]} = (x-1)(x-2)\ldots(x-g)$.

For $g_1 = g_2 = 1$ it follows from (13), (8) and (9) after simple but tedious computations

$$E(r_1r_2) = \frac{n_1n_2(n_1n_2+n-1)}{n(n-1)} . \qquad (14)$$

From (8), (9) (14), and (12) we obtain, again after elementary but unwieldy computations

$$COV(r_1,r_2) = \frac{n_1(n_1-1)\, n_2(n_2-1)}{n^2(n-1)} . \qquad (15)$$

The mean of the total number of runs can now be directly computed from $E(r_1)$ and $E(r_2)$. It follows from (8) and (9)

$$E(r) = E(r_1) + E(r_2) = \frac{2n_1n_2+n}{n} \qquad (16)$$

and from (10), (11) and (15) we get

$$V(r) = V(r_1) + V(r_2) + 2\,COV(r_1,r_2)$$

$$= \frac{2n_1n_2(2n_1n_2-n)}{n^2(n-1)} . \qquad (17)$$

If we substitute n_i for n_1 and $n-n_i$ for n_2 in (8) and (10), we arrive at the mean and the variance of the number of runs of a certain type a_i in the case of k different types of elements:

$$E(r_i) = \frac{n_i(n-n_i+1)}{n} \qquad (18)$$

$$V(r_i) = \frac{n_i(n_i-1)(n-n_i)(n-n_i+1)}{n^2(n-1)} \qquad (19)$$

If we substitute n_i for n_1 and n_j for n_2 in (15), we get the formula for the covariance of k different types of runs:

$$COV(r_i,r_j) = \frac{n_i(n_i-1)\, n_j(n_j-1)}{n^2(n-1)} \qquad (20)$$

From (18), (19), and (20) we can derive the mean and the variance of the total number of runs of k different types of elements:

$$E(r) = \sum_{i=1}^{k} E(r_i) = \sum_{i=1}^{k} \frac{n_i(n - n_i + 1)}{n}$$

$$= n + 1 - \frac{\sum_{i=1}^{k} n_i^2}{n} \qquad (21)$$

$$V(r) = \sum_{i=1}^{k} V(r_i) + \sum_{\substack{i=1 \\ i\neq j}}^{k} \sum_{j=1}^{k} COV(r_i, r_j)$$

In order to facilitate the evaluation of the double sum, we expand the summation to the case i=j. The additional sum which thus results is subtracted:

$$V(r) = \sum_{i=1}^{k} \frac{n_i(n_i - 1)(n - n_i)(n - n_i + 1)}{n^2(n - 1)}$$

$$+ \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{n_i(n_i - 1)\, n_j(n_j - 1)}{n^2(n - 1)}$$

$$- \sum_{i=1}^{k} \frac{n_i(n_i - 1)\, n_i(n_i - 1)}{n^2(n - 1)}$$

After several rearrangements we finally have

$$V(r) = \frac{\sum_{i=1}^{k} n_i^2 \left[\sum_{i=1}^{k} n_i^2 + n(n + 1)\right] - 2n \sum_{i=1}^{k} n_i^3 - n^3}{n^2(n - 1)} \qquad (22)$$

For the derivation of the moments in the case that n_1 and n_2 are not fixed, but are to be seen as random variables, we use the relation

$$E(Y) = E[E(Y|X)]. \qquad (23)$$

This relation allows us to compute the moments we are looking for by using the moments already derived.

For example, for $E(r_1, n_1, n_2)$ using formula (8) we get

$$E(r_1, n_1, n_2) = E[E(r_1 | n_1, n_2)] = E\left[\frac{n_1(n - n_1 + 1)}{n}\right]$$

$$= E(n_1) - \frac{1}{n}E(n_1^2) + \frac{1}{n}E(n_1).$$

If we take into account that for binomially distributed random variables $E(n_1) = np$ and $E(n_1^2) = n^2p^2 + npq$, we get

$$E(r_1, n_1, n_2) = npq + p^2. \qquad (24)$$

With the help of the relation (23) and the relations used for the computation of the conditional moments, we also obtain the remaining moments:

$$E(r_2, n_1, n_2) = npq + q^2 \qquad (25)$$

$$E(r, n_1, n_2) = 2npq + p^2 + q^2 \qquad (26)$$

$$V(r_1, n_1, n_2) = np(1 - 4p + 6p^2 - 3p^3) + p^2(3 - 8p + 5p^2) \qquad (27)$$

$$V(r_2, n_1, n_2) = nq(1 - 4q + 6q^2 - 3q^3) + q^2(3 - 8q + 5q^2) \qquad (28)$$

$$COV(r_1, r_2, n_1, n_2) = -npq(3pq - 1) - pq(2 - 5pq) \qquad (29)$$

$$V(r, n_1, n_2) = 4npq(1 - 3pq) - 2pq(3 - 10pq) \qquad (30)$$

For the case $p = q = 1/2$, which often occurs in practical use, we have

$$E(r,n_1,n_2) = \frac{n + 1}{2} \qquad (31)$$

$$V(r,n_1,n_2) = \frac{n - 1}{4}. \qquad (32)$$

2.3 *Normal Approximation*

Using the derived means and variances, we can now approximate the distribution of runs by the normal distribution. As can be shown, for $n \to \infty$ the variable

$$z = \frac{x - E(x)}{\sqrt{V(x)}} \qquad (33)$$

tends to be normally distributed with mean 0 and variance 1. By means of the transformation (33) we now can very easily test the significance of the number of runs.

If we want to test the significance of the total number of runs for given n_i, *without* a preceding computation of the mean and the variance, we can use the two following formulas which we obtain after several rearrangements by inserting the appropriate characteristics in (33):

$$z = \frac{n(r - 1) - 2n_1n_2}{\sqrt{\frac{2n_1n_2(2n_1n_2 - n)}{n - 1}}} \qquad (34)$$

$$z = \frac{\Sigma\, n_i^2 - n(n - r)}{\sqrt{\frac{\Sigma\, n_i^2\,[\Sigma n_i^2 + n(n + 1)] - 2n\,\Sigma\, n_i^3 - n^3}{n - 1}}} \qquad (35)$$

with $i = 1,\ldots,k$

Formula (34) is an approximation for binary runs, (35) for multiple runs. (34) is sufficiently exact for $n_1,n_2>10$ and (35) also gives us satisfying results even for relatively small samples. More information concerning asymptotical distributions of runs can be found in MOOD (1940) or DAVID/BARTON (1962).

With the help of the normal distribution we can also test whether there is a difference between the number of runs in two *different* samples. The corresponding transformation is:

$$z = \frac{x - y - [E(x) - E(y)]}{\sqrt{V(x) + V(y)}} \qquad (36)$$

The normal approximation can be improved by a correction for continuity. With this correction we take into account that, in transforming a discrete random variable into a normal variable, each value x of the discrete variable corresponds to an interval (x-0.5; x+0.5). Therefore, for the probabilities the relations $P(X \geq x) = P(X \geq x-0.5)$ and $P(X \leq x) = P(X \leq x+0.5)$ hold. Hence, if we want to carry out a *left-tail test*, the transformation into a normal variable is

$$z = \frac{x + 0.5 - E(x)}{\sqrt{V(x)}}. \qquad (37)$$

In the case of a *right-tail test*, we have

$$z = \frac{x - 0.5 - E(x)}{\sqrt{V(x)}}. \qquad (38)$$

For the *two-tail test* we use

$$z = \frac{|x - E(x)| - 0.5}{\sqrt{V(x)}}. \qquad (39)$$

In a comparison of two different samples, the formula corrected for continuity is

$$z = \frac{(x - 0.5) - (y + 0.5) - [E(x) - E(y)]}{\sqrt{V(X) + V(y)}}$$

$$= \frac{x - y - 1 - [E(x) - E(y)]}{\sqrt{V(x) + V(y)}} \qquad (40)$$

for $x > y$.

For smaller sample sizes the corrected formulas should definitely be used.

3. *Some Instances of Application*

We will now give several examples of the use of the above formulas.

Example 3.1: We considered above the sequence of stressed and unstressed syllables in the first two lines of Goethe's "Erlkönig" (see Example 1.1). These two lines correspond to a binary sequence with n_u=10 unstressed syllables, n_s=8 stressed syllables and r_u=r_s=8 runs of the type u or s.

We now test whether there is a significant difference between the frequencies of stressed und unstressed syllables; let the level of significance be $\alpha = 0.05$. If we assume that the probability of both a stressed and an unstressed syllable is $p_s = p_u = 0.5$, then the null hypothesis (H_0) and the alternative hypothesis (H_A) can be formulated as follows:

$$H_0 : p_u = p_s = 0.5$$

$$H_A : p_u \neq p_s$$

With the help of the binomial distribution we get

$$P(X \leq 8) = P(X \geq 10) = \sum_{x=0}^{8} \binom{18}{x} 0.5^{18} = 0.4073.$$

Since the test is two-tailed, the obtained value must be doubled. Because 2(0.4073) >> 0.05, there is no reason to doubt that the sequence we are examining has the character of a random sequence with regard to the number of stressed and unstressed syllables. It must be kept in mind, however, that this conclusion is based on the assumption that $p_s = p_u = 0.5$.

If we now look at the *order* of the elements, then there seems to be a tendency toward alternation of stressed and unstressed syllables - as is to be expected in a text of this kind. We test this hypothesis by computing the probability that among $n_1 + n_2 = 10 + 8 = 18 = n$ sample elements at least $r = r_1 + r_2 = 8 + 8 = 16$ runs occur, i.e. we compute

$P(r \geq 16 | n_1=10, n_2=8)$. For even r we use formula (2), for odd r formula (3):

$$P(r = 16) = \frac{2\binom{9}{7}\binom{7}{7}}{\binom{18}{10}} = 0.00165$$

Because $\binom{a}{b} = 0$ for a<b, we get for P(r = 17)

$$P(r = 17) = \frac{\binom{9}{8}\binom{7}{7} + \binom{9}{7}\binom{7}{8}}{\binom{18}{10}} = \frac{\binom{9}{8}\binom{7}{7}}{\binom{18}{10}} = 0.00021$$

Since for $n_1 = 10$ and $n_2 = 8$ a maximum of r = 17 runs can occur, the upper limit of summation is 17 and P(r = 18) = 0. This is also shown when we compute P(r = 18) with the help of (2):

$$P(r = 18) = \frac{2\binom{9}{8}\binom{7}{8}}{\binom{18}{10}} = 0$$

The probability we are looking for is thus

$$P(r \geq 16) = P(r = 16) + P(r = 17) = 0.00186.$$

Since 0.00186 < 0.05, we can reject the null hypothesis "The arrangement of the sample elements is due to chance" in favour of the alternative hypothesis "There is a tendency toward regular alternation". We could have, however, also obtained this result without any computation by using the tables of SWED/EISENHART (1943), where the upper limit for $\alpha = 0.05$ is given as r = 14).

As the example shows, a conclusion about the randomness of the arrangements of the sample elements can be drawn, with the theory of runs, even for very small samples. In the above example, however, not much is gained by merely stating the non-randomness of the sequence, since only a comparison, for example with syllable sequences in other texts of verse or prose, makes further interpretation of the results possible.

Example 3.2: We will therefore compare "Erlkönig" - this time in its entirety - with another of Goethe's ballads, "Der Totentanz". Since n_1 and $n_2 > 20$, we use the normal approximation to test the significance of the total number of runs of stressed and unstressed syllables. The necessary data are presented in Table 3.

Table 3: Observed and Expected Number of Runs of Stressed (s) and Unstressed (u) Syllables in "Erlkönig" and in "Der Totentanz"

Syllables	Erlkönig			Der Totentanz		
	n_1	$\hat{r}_1$	$E(\hat{r}_1)$	n_1	$\hat{r}_1$	$E(\hat{r}_1)$
s	128	126	75.22	175	175	113.48
u	180	126	75.39	320	176	113.78
	308	252	150.61	495	351	227.26
	$V(\hat{r}) = 72.42$			$V(\hat{r}) = 103.18$		

Let us illustrate the computation of the means and the variances using "Der Totentanz". From formulas (8), (9), (16), and (17) we get

$$E(\hat{r}_1) = \frac{175(320 + 1)}{495} = 113.4848$$

$$E(\hat{r}_2) = \frac{320(175 + 1)}{495} = 113.7778$$

$$E(\hat{r}) = 113.4848 + 113.7778 = \frac{2(175)320 + 495}{495} = 227.2626$$

$$V(\hat{r}) = \frac{2(175)320\,[2(175)320 - 495]}{495^2(495 - 1)} = 103.1751$$

In the same manner we obtain the values for "Erlkönig". Because of the meter, runs of more than one stressed syllable do not occur in "Erlkönig", with the exception of the beginnings of lines 22 and 28, and not at all in "Der Totentanz". Therefore in "Der Totentanz" the number of runs of stressed syllables is equal to the number of stressed syllables, and in both poems the expected number of runs is clearly smaller than the observed number.

Since the sample size is relatively large, we can use the normal approximation without a correction for continuity. According to (33) the transformation is

$$z = \frac{r - E(r)}{\sqrt{V(r)}} . \qquad (41)$$

We obtain the following values:

"Erlkönig": $$\hat{z} = \frac{252 - 150.61}{\sqrt{72.42}} = 11.91$$

"Der Totentanz": $$\hat{z} = \frac{351 - 227.26}{\sqrt{103.18}} = 12.18$$

These two values could have been obtained without computing $E(\hat{r})$ and $V(\hat{r})$ directly from formula (34). For "Erlkönig", for instance, the computation is

$$\hat{z} = \frac{308(252 - 1) - [2(128)180]}{\sqrt{\frac{2(128)180[2(128)180 - 308]}{308 - 1}}} = 11.91.$$

The $\hat{z}$-value is highly significant in both cases. The hypothesis of a tendency toward regular alternation of stressed and unstressed syllables can thus be regarded as confirmed.

With the help of the normal distribution we can now also test whether there is a difference between the two texts. According to (36) the test statistic is

$$z = \frac{r_A - r_B - [E(r_A) - E(r_B)]}{\sqrt{V(r_A) + V(r_B)}} . \qquad (42)$$

With this criterion we do not test the difference between two numbers of runs directly, but indirectly by means of the difference of both numbers from their expected values. A direct comparison is only possible if for the number of elements a_1 and a_2 in the samples to be compared $n_{1A} \cdot n_{2A} = n_{1B} \cdot n_{2B}$ holds. In this case $E(r_A) = E(r_B)$, and (42) becomes

$$\hat{z} = \frac{r_A - r_B}{\sqrt{V(r_A) + V(r_B)}} . \qquad (43)$$

By comparing "Erlkönig" and "Der Totentanz", we get

$$\hat{z} = \frac{|252 - 351 - (150.61 - 227.26)|}{\sqrt{72.42 + 103.18}} = 1.69.$$

For the two-tail test, when $\alpha = 0.05$, the critical value is 1.96. The difference between the observed and the expected number of runs is thus seen as the same for both texts.

Example 3.3: Now in the following comparison between a text of verse and one of prose in respect to long (i.e. naturally long and long by position) and short syllables, there is a more definite difference, as opposed to Example 3.2. For the two samples we have used the beginnings of Virgil's *Aeneid* and Caesar's *Bellum Gallicum*. The data are shown in Table 4.

To test the difference in the total number of runs, we again use the normal approximation. Because of the relatively small sample size, we make a correction for continuity this time. For the *Aeneid*, where a tendency toward alternation and thus a higher number of runs is to be expected, we carry out a one-tail test, for *Bellum Gallicum*, however, a two-tail test.

Table 4: Observed and Expected Number of Runs in the *Aeneid* and in *Bellum Gallicum*:

Syllable Type	Aeneid			Bellum Gallicum		
	n_i	$\hat{r}_i$	$E(\hat{r}_i)$	n_i	$\hat{r}_i$	$E(\hat{r}_i)$
long	45	30	17.50	48	15	14.82
short	27	14	17.25	20	13	14.41
	72	44	34.75	68	28	29.24
	$V(\hat{r}) = 15.57$			$V(\hat{r}) = 11.48$		

Using formulas (38) and (39), we get for the *Aeneid*

$$\hat{z} = \frac{\hat{r} - 0.5 - E(\hat{r})}{\sqrt{V(\hat{r})}} = \frac{44 - 0.5 - 34.75}{\sqrt{15.57}} = 2.22$$

and for *Bellum Gallicum*

$$\hat{z} = \frac{|\hat{r} - E(\hat{r})| - 0.5}{\sqrt{V(\hat{r})}} = \frac{|28 - 29.24| - 0.5}{\sqrt{11.48}} = 0.22 .$$

The $\hat{z}$ values clearly show the difference between verse and prose. In the prose text the observed and the expected number of runs largely agree. In the verse text, on the other hand, we see that the meter limits the possible choices and increases the sequential probability structure of language. In terms of information theory this can be interpreted as a tendency toward increasing the redundancy of language. If we compare the two texts with each other by means of (40), the difference also becomes clear:

$$\hat{z} = \frac{\hat{r}_A - \hat{r}_B - 1 - [E(\hat{r}_A) - E(\hat{r}_B)]}{\sqrt{V(\hat{r}_A) + V(\hat{r}_B)}}$$

$$= \frac{44 - 28 - 1 - (34.75 - 29.24)}{\sqrt{15.57 + 11.48}} = 1.82$$

Since we are using a one-tail test, the critical value is 1.645 for $\alpha = 0.05$. The deviation from the expected value is thus more salient in the *Aeneid*.

Example 3.4: In order to test the hypothesis that the sequence of consonants (C) and vowels (V) in texts from natural language is not only produced by chance, MULLER (1974) determined the theoretical frequencies of the sequences CV, VC, VV, and CC by means of the probability of occurrence of consonants and vowels, estimated from a larger sample, and compared them with the observed frequencies using a χ^2 test. Since this procedure is relatively awkward, MULLER further suggested that, for large samples, the randomness of the sequence of consonants and vowels should be tested by comparing the observed variance with the theoretical variance computed on the basis of a binomial model.

The hypothesis studied by MULLER can easily be tested with the help of the theory of runs. Table 5 shows the number of runs of consonants and vowels in a French written text, in a German text, and in a sequence generated by a random procedure. The data for the random sequence and for the French text are based on Table 1A and 1B in MULLER (1974: 32). The German text consists of the first 100 phonemes of Heinrich Böll's "Ansichten eines Clowns".

Again we test the significance of the total number of runs by using the normal approximation. Since we expect a tendency toward alternation in French and in German, we carry out a one-tail test for these texts, for the random sequence, however, a two-tail test. In all cases we test at the 0.05 level. For the French text we obtain

$$\hat{z} = (83 - 50.02) / \sqrt{23.78} = 6.76,$$

for the German text

$$\hat{z} = (73 - 48.12) / \sqrt{21.95} = 5.31$$

and for the random sequence

$$\hat{z} = |43 - 50.02| / \sqrt{23.78} = 1.44.$$

As could be expected, both the French and the German texts show an obvious tendency toward alternation of consonants and vowels. For the random sequence, however, the deviation from the expected value can be considered as produced by chance. If one compares the French and the German texts with each other by means of formula (42), we obtain $\hat{z} = 1.20$, and hence a non-significant result.

Table 5: Number of Runs of Consonants (C) and Vowels (V) in a French Text, in a German Text, and in a Random Sequence

	French			German			Random Sequence		
	n_i	$\hat{r}_i$	$E(\hat{r}_i)$	n_i	$\hat{r}_i$	$E(\hat{r}_i)$	n_i	$\hat{r}_i$	$E(\hat{r}_i)$
C	57	42	25.08	62	36	24.18	57	22	25.08
V	43	41	24.94	38	37	23.94	43	21	24.94
	100	83	50.02	100	73	48.12	100	43	50.02
	$V(\hat{r}) = 23.78$			$V(\hat{r}) = 21.95$			$V(\hat{r}) = 23.78$		

Testing the number of runs has one important advantage over MULLER's procedure: it can be used even if the probabilities of occurrence of consonants and vowels in the particular language or language variety are not known. Furthermore, the test for runs also seems to have greater power. Since $z^2 = \chi^2_1$, we only need to square the obtained $\hat{z}$ values in

order to be able to compare them with MULLER's χ^2 values. We get $6.67^2 = 45.74$ and $1.44^2 = 2.07$. The corresponding χ^2 values in MULLER are, however, 43.94 and 1.87. Considering further that the χ^2 values from the test for runs have only one degree of freedom, the corresponding values in MULLER, however, two, we get relatively important differences in the related probabilities.

As we have shown in Part 2, it is also possible to test the total number of runs when n_1 and n_2 are not fixed, but are considered to be random variables. In this case, though, we must know the probability of occurrence of the elements of the sequence in the population. MEIER (1967: 252 f.) computed a relative frequency of $\hat{p} = 0.6269$ for consonants and of $\hat{q} = 0.3731$ for vowels using a sample of 48,222 sounds from German prose texts. We disregard the sampling error, which is small because of the very large sample size, and consider both values to be probabilities of occurence of consonants and vowels in German prose. We now test the total number of runs in Böll's *Ansichten eines Clowns* again, this time, however, under the assumption that p and q are known. We first calculate the mean and the variance using formulas (26) and (30). We get

$$E(r,n_1,n_2) = 2(100)0.6269(0.3731) + 0.6269^2 + 0.3731^2 = 47.3115$$

and in the same manner

$$V(r,n_1,n_2) = 27.6003.$$

It is shown that the deviation from the mean is somewhat larger than for fixed n_i. Since the variance is also much larger this time, the z test results in a value of only 4.89, which is, of course, also highly significant. Both tests thus point to a strong tendency toward alternation of vowels and consonants.

Example 3.5: FUCKS (1968; 1971) established, using correlational analysis, that the lengths of sentences in sequence are not independent of each other, but that long sentences tend to occur with long ones, and short sentences with short ones. In order to check this hypothesis, we determine the

number of syllables per sentence in 30 sequential sentences in Goethe's *Letter 605*, compute the median, note whether a sentence is larger (l) or smaller (s) than the median, and test by means of the total number of runs of the elements l and s whether there is a "clumping effect". Because of this procedure, this test is often called "Test of the total number of runs above and below the median". The data are presented in Table 6.

Table 6: Analysis of the Number of Syllables per Sentence in Goethe's *Letter 605*

Sent. No.	Syll.	Type	Sent. No.	Syll.	Type	Sent. No.	Syll.	Type
1	20	s	11	14	s	21	9	s
2	26	l	12	23	s	22	9	s
3	34	l	13	19	s	23	9	s
4	40	l	14	19	s	24	35	l
5	42	l	15	24	s	25	65	l
6	32	l	16	67	l	26	21	s
7	37	l	17	28	l	27	13	s
8	19	s	18	15	s	28	32	l
9	3	s	19	11	s	29	37	l
10	29	l	20	30	l	30	28	l

Since the number of values is even, the median does not fall on an observed value and a category "equal to the median" does thus not occur. If, however, the number of values in an investigation is odd, then we can ignore the value which falls on the median in order to avoid a category "equal to the median". We obtain a value of 25 for the median by computing the arithmetic mean of the values 24 and 26. Since the number of values below the median equals the number of values above the median, we have $n_1 = n_2$ and equations (16) and (17) become

$$E(r) = n_1 + 1 \qquad (44)$$

$$V(r) = \frac{n_1(n_1 - 1)}{2n_1 - 1} . \qquad (45)$$

We get $E(\hat{r}) = 16$ and $V(\hat{r}) = 7.24$. From Table 6 we compute for the total number of runs $\hat{r} = 12$. As was to be expected, $\hat{r}$ is smaller than $E(\hat{r})$ and thus indicates a clumping effect. We now look for the probability of obtaining a value of $\hat{r} \leq 12$. The transformation into a normal variable, taking the correction for continuity into account, gives, according to formula (37)

$$\hat{z} = \frac{\hat{r} + 0.5 - E(\hat{r})}{\sqrt{V(\hat{r})}} = \frac{12 + 0.5 - 16}{\sqrt{7.24}} = -1.30.$$

Since the critical value of -1.645 is not reached, FUCKS' results are not confirmed by our sample. It is, however, entirely possible that a significant result could have been obtained from a somewhat larger sample.

Example 3.6: Until now we have only investigated binary runs. In the following two examples, we now concern ourselves with multiple runs.

If we consider the arrangement of dactyls, spondees, and trochees in Latin hexameter, then we observe that - in addition to the well-known fact that trochees are limited to the sixth foot and that the so-called *spondiaci*, i.e. verses with a spondee in the fifth foot, are very rare - certain limitations appear also for the first four feet. For example, there seems to be a tendency to prefer spondees after dactyls. Because of the strong disassociative effect of the fifth (dactylic) foot, this tendency seems to increase with proximity to the fifth foot (cf. GROTJAHN 1979, Chapter 13). Table 7 shows the number of runs of dactyls (d), spondees (s), and trochees (t) in the first 30 lines of Virgil's *Aeneid*.

Table 7: Number of Runs of Dactyls (d), Spondees (s), and Trochees (t) in the first 30 lines of Virgil's *Aeneid*

Type	n_i	$\hat{r}_i$	$E(\hat{r}_i)$
d	89	61	45.49
s	81	56	45.00
t	10	10	9.50
	180	127	99.99
	$V(\hat{r}) = 40.76$		

The values shown in Table 7 for $E(\hat{r}_i)$ and $V(\hat{r})$ are calculated as follows: Using formula (18), we have, e.g. for the number of runs of dactyls, the mean

$$E(\hat{r}_d) = \frac{n_d(n - n_d + 1)}{n} = \frac{89(180 - 89 + 1)}{180} = 45.49.$$

Analogously $E(\hat{r}_s)$ and $E(\hat{r}_t)$ are computed, and finally $E(\hat{r})$ and $V(\hat{r})$ by means of formulas (21) and (22). Now we test, using the total number of ternary runs, whether there is, as expected, a tendency toward alternation. The transformation to a normal variable, according to (38), results in

$$\hat{z} = \frac{\hat{r} - 0.5 - E(\hat{r})}{\sqrt{V(\hat{r})}} = \frac{127 - 0.5 - 99.99}{\sqrt{40.76}} = 4.15.$$

The $\hat{z}$ value is highly significant. The hypothesis of a tendency toward alternation is thus confirmed.

Example 3.7: If we consider only the first four feet in Latin hexameter, a total of 16 different hexameter types,

according to the combinations of dactyls (d) and spondees (s), can be seen. We now test whether the arrangement of these 16 types in a running hexameter text can be viewed as due to chance. Table 8, which is partly based on the computational analyses of Latin hexameter verse by OTT (1970; 1974), shows the number of runs of the 16 types of lines in the first 300 lines of both Horace's *De Arte Poetica* and Lucrece's *De Rerum Natura*. Again formulas (18), (21), and (22) were used to calculate the means and the variances.

Table 8: Number of Runs of Different Hexameter Types in Horace's *De Arte Poetica* and Lucrece's *De Rerum Natura*

Type		Horace			Lucrece		
		n_i	$\hat{r}_i$	$E(\hat{r}_i)$	n_i	$\hat{r}_i$	$E(\hat{r}_i)$
1	dddd	4	4	3.96	6	6	5.90
2	sddd	12	12	11.56	8	8	7.81
3	dsdd	14	13	13.39	6	5	5.90
4	ddsd	17	17	16.09	9	9	8.76
5	ddds	24	24	22.16	29	27	26.29
6	ssdd	12	11	11.56	3	3	2.98
7	sdsd	21	19	19.60	7	7	6.86
8	sdds	16	15	15.20	20	20	18.73
9	dssd	25	20	23.00	21	20	19.60
10	dsds	37	34	32.56	27	25	24.66
11	ddss	22	21	20.46	40	34	34.80
12	sssd	10	9	9.70	4	4	3.96
13	ssds	24	23	22.16	18	18	16.98
14	sdss	26	23	23.83	32	28	28.69
15	dsss	25	22	23.00	51	40	42.50
16	ssss	11	11	10.63	19	18	17.86
		300	278	278.87	300	272	272.29
		$V(\hat{r}) = 19.24$			$V(\hat{r}) = 24.84$		

As can be seen from Table 8, a widespread agreement exists between the observed and expected numbers of runs. This is also shown by testing the total number of runs using the z transformation, where we only get $|\hat{z}| = 0.20$ for *De Arte Poetica* and $|\hat{z}| = 0.06$ for *De Rerum Natura*. The arrangement of the line types can thus be viewed as due to chance. However, the test which was carried out probably does not have a very high power for 16 different types of elements. Possible tendencies toward structuring, such as avoiding the immediate contact of certain line types, can certainly not be discovered in this manner.

If we test, however, the 16 lines for equidistribution, then we cannot sustain the hypothesis of a random sequence with $p_1 = p_2 = \ldots = p_{16}$. For this we use the *minimum discrimination information statistic* developed by KULLBACK et al., which is preferable to the usual χ^2 test in many cases (see KULLBACK et al. 1962; BLÖSCHL 1966; ALTMANN 1980). The formula for testing the hypothesis of equidistribution is

$$2I = \sum_{i=1}^{k} 2n_i \ln n_i - 2n \ln n + 2n \ln k. \qquad (46)$$

Under the null hypothesis the test statistic 2I is asymptotically distributed as χ^2 with df = k-1. Since several extensive tables for $2n_i \ln n_i$ are available (e.g. KULLBACK et al. 1962; BLÖSCHL 1966), we need only to add the tabulated values in order to compute 2I. For *De Arte Poetica* we obtain

$$2\hat{I} = [2(4) \ln 4 + \ldots + 2(11) \ln 11] - 2(300) \ln 300 + 2(300) \ln 16 = 56.89.$$

For df = 15 and $\alpha = 0.05$ the critical χ^2-value is 25.00. Since $2\hat{I} = 56.89$ lies far above this value, there is a definite tendency toward a preference for certain line types. As Table 8 shows, this is primarily the case in lines with dactyls in the first feet. This tendency is even more pronounced in *De Rerum Natura*, where $2\hat{I} = 156.91$. If we compare the texts

with each other, we get $2\hat{I} = 43.62$ with df = 15 and hence a significant difference.

Example 3.8: In Example 1.1 we investigated the number of runs of lines of 8, 9, 10, 11, and 12 syllables in Goethe's "Erlkönig". If we calculate the mean and the variance of the total number of runs for the data in Table 2, we obtain, using formulas (21) and (22), $E(\hat{r}) = 22.19$ and $V(\hat{r}) = 4.85$. Since the observed total number of runs is only 16, the impression that lines of the same length tend to cluster seems to be confirmed. The transformation to a normal variable, according to formula (39), results in $|\hat{z}| = 1.68$. For a two-tail test at the 0.05 level the critical value is 1.96. A tendency toward structuring in the use of the different line lengths can thus not be supported.

4. *Conclusion*

Only part of the theory of runs and only a few of the possibilities for its use are presented in this article. For example, two independent, continuously distributed samples can also be tested by means of the total number of runs as to whether they are from the same population. For this we arrange the $n_1 + n_2 = n$ elements of the pooled sample in ascending order of size and mark whether a value comes from sample 1 or sample 2. If the populations are different, it can be expected that values from the same sample tend to cluster. This procedure, which was developed by WALD/WOLFOWITZ (1940) and which is occasionally used, for example when the requirements for the application of the t test are not fulfilled, is, however, one of the weakest nonparametric tests for a comparison of two independent samples (see BRADLEY 1968: 264).

In addition to the total number of runs we can also test the *length* of runs for significance (see BRADLEY 1968: 255 ff). Furthermore, we can investigate the distribution of so-called *runs up and down*. In this manner we can also discover trends and cycles in a sample. For this we compare two adjacent sample

elements with regard to size or equality, mark the result of the comparison with + or - and compute the distribution of the total number of runs of the n-1 plus or minus signs (see BRADLEY 1968: 271 ff; GIBBONS 1971: 62 ff). In this way we could test, for example, whether a text shows a tendency toward alliteration.

SUMMARY

While most statistical methods only take the *frequency* of the sample elements into account, but not their arrangement, the *theory of runs* also makes it possible to test the *arrangement* of sample elements for randomness. In this article the theory of runs is presented rather explicitly, and its use is clarified by several examples. The intention is to make the theory of runs available to as many researchers in quantitative linguistics as possible.

Notes

1 The following presentation of the theory of runs is partly based on GROTJAHN (1979: 143-161). An early example of the application of this theory in the study of verse is WORONCZAK's article (1961).

2 With regard to the following derivations see primarily MOOD (1940), DAVID/BARTON (1962: 85 ff, 119 ff), WILKS (1962:144 ff), BRUNK (1965: 354 ff), BROWNLEE (1965: 224 ff), BRADLEY (1968: 253 ff), FISZ (1970: 483 ff), GIBBONS (1971: 50 ff), and GROTJAHN (1979: 145-153).

3 There are several possibilities for computing E(r) and V(r). In addition to the method of factorial moments, the method of indicator function also allows a fairly simple derivation (see DAVID/BARTON 1962: 85 ff, MORAN 1968: 38 f, GIBBONS 1971: 56 f).

References

ALTMANN, G., The Homogeneity of Metric Patterns in Hexameter. In: GROTJAHN, R. (Ed.), Hexameter Studies. Bochum, Brockmeyer 1980 (forthcoming)

BLÖSCHL, L., Kullbacks 2Î-Test als ökonomische Alternative zur Chi^2-Probe. Psychologische Beiträge 9, 1966, 379-406

BRADLEY, J.V., Distribution-Free Statistical Tests. Englewood Cliffs, N.J., Prentice Hall 1968

BRAINERD, B., On the Markov Nature of Text. Linguistics 176, 1976, 5-30

BROWNLEE, K.A., Statistical Theory and Methodology in Science and Engineering. New York, Wiley 1965

BRUNK, H.D., An Introduction to Mathematical Statistics. Waltham, Mass., Blaisdell 1965

DAVID, F.N./BARTON, D.E., Combinatorial Chance. London, Charles Griffin 1962

FISZ, M., Wahrscheinlichkeitsrechnung und mathematische Statistik. Berlin, VEB Deutscher Verlag der Wissenschaften 1970

FUCKS, W., Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen. Köln/Opladen, Westdeutscher Verlag 1955

FUCKS, W., Nach allen Regeln der Kunst. Stuttgart, Deutsche Verlags-Anstalt 1968

FUCKS, W., Possibilities of Exact Style Analysis. In: STRELKA, J. (Ed.), Patterns of Literary Style. University Park, Pa., Pennsylvania State University Press 1971, 51-75

GIBBONS, J.D., Nonparametric Statistical Inference. New York, McGraw-Hill 1971

GREGER, K., Zufälliges Zusammenklumpen. Aus der Schulpraxis - Für die Schulpraxis 25, 1972, 211-214

GROTJAHN, R., Linguistische und statistische Methoden in Metrik und Textwissenschaft. Bochum, Brockmeyer 1979

KULLBACK, S./KUPPERMAN, M./KU, H.H., An Application of Information Theory to the Analysis of Contingency Tables, With a Table of 2n ln n, n = 1(1)10,000. Journal of Research of the National Bureau of Standards - B. Mathematics and Mathematical Physics 66, 1962, 217-243

MEIER, H., Deutsche Sprachstatistik. Hildesheim, Olms 1967

MOOD, A.M., The Distribution Theory of Runs. Annals of Mathematical Statistics 11, 1940, 367-392

MORAN, P.A.P., An Introduction to Probability Theory. Oxford, Clarendon Press 1968

MULLER, Ch., Pour une mesure statistique de certaines contraintes idiomatiques. In: DAVID, J./MARTIN, R. (Eds.), Statistique et Analyse Linguistique. Paris, Kliencksieck 1974, 29-37

OTT, W., Metrische Analysen zur Ars Poetica des Horaz. Göppingen, Kümmerle 1970

OTT, W., Metrische Analysen zu Lukrez De Rerum Natura Buch I. Tübingen, Niemeyer 1974

OWEN, D.B., Handbook of Statistical Tables. Reading, Mass., Addison-Wesley 1962

ROACH, S.A., The Theory of Random Clumping. London, Methuen 1968

SWED, F.S./EISENHART, C., Tables for testing randomness of grouping in a sequence of alternatives. Annals of Mathematical Statistics 14, 1943, 66-87

WALD, A./WOLFOWITZ, J., On a Test Whether Two Samples are from the Same Population. Annals of Mathematical Statistics 11, 1940, 147-162

WILKS, S.S., Mathematical Statistics. New York, John Wiley 1962

WORONCZAK, J., Statistische Methoden in der Verslehre. In: DAVIE, D. et al. (Eds.), Poetics, Poetyka, Poètika, Vol. I, Warszawa, Państwowe Wydawnictwo Naukowe 1961, 607-624

Zur numerischen Erfassung der Schwierigkeit des Sprechbewegungsablaufs

Werner Lehfeldt, Konstanz

In letzter Zeit lenkt die Frage nach der Beschaffenheit der Sprache, in der die Linguistik über Sprache redet, eine immer stärkere Aufmerksamkeit auf sich. Immer mehr Linguisten werden sich dessen bewußt, daß die Chancen der Sprachwissenschaft, bessere, tiefere Einsichten in ihren Gegenstand zu gewinnen, nicht zuletzt davon abhängen, daß es gelingt, eine linguistische Metasprache zu schaffen, deren Aufbau einer Reihe formaler Kriterien genügen muß. Eines der wichtigsten Kriterien, die hier zu beachten sind, lautet, daß die Begriffe der angestrebten Metasprache ein System bilden, d.h., zueinander in eindeutig definierten Beziehungen stehen müssen.

Selbstverständlich wäre es eine Illusion, wollte man glauben, es könne sozusagen auf einen Schlag gelingen, eine solche Metasprache zu konstruieren, die alle Bereiche der Linguistik gleichermaßen abdeckte. Es ist klar, daß viele Einzelbemühungen erforderlich sind, damit wir uns dem angestrebten Ziel annähern können, und aus diesem Grunde sind alle Bemühungen zu unterstützen, die darauf abzielen, einzelne begriffliche und terminologische Teilsysteme in der angedeuteten Weise zu konstruieren (vgl. etwa MEL'ČUK 1963a; 1963b; 1975; 1978; SOVA 1970; ŠELOV 1974).

Die Notwendigkeit, die Sprache der Linguistik zu untersuchen und zu verbessern, gilt ganz allgemein, ihr sehen sich alle sprachwissenschaftlichen Disziplinen gegenüber. In besonderem

Maße gilt sie für den mit quantitativen Methoden arbeitenden Zweig der Sprachwissenschaft, die sogenannte quantitative Linguistik. Sie stellt sich indes für diese in einer besonderen Form: Zu den Aufgaben der quantitativen Linguistik gehört es, eine quantitative Begriffssprache zu entwickeln, d.h. die Ausprägungen von Eigenschaften sprachlicher Größen in einer exakten, kurzgefaßten, "handlichen" Form zu erfassen, d.h. sie - nach Möglichkeit - zu metrisieren. Es handelt sich hierbei nicht um eine Aufgabe, der man sich nur nach Lust und Laune zuzuwenden brauchte. Vielmehr ist an die Erfüllung der genannten Voraussetzung die Möglichkeit gebunden, die übrigen, eigentlichen Ziele der quantitativen Linguistik zu erreichen, vor allem das Ziel, sprachliche Erscheinungen und Zusammenhänge zwischen ihnen aus einer Theorie heraus zu erklären und/oder vorauszusagen.

Aus diesem Zusammenhang wird ohne weiteres ersichtlich, weshalb sich die quantitative Linguistik im heutigen Stadium ihrer Entwicklung intensiv mit den methodologischen und den empirischen Problemen der Metrisierung sprachlicher Erscheinungen und Relationen auseinanderzusetzen hat. Insbesondere der vertiefenden Beschäftigung mit den methodologischen Aspekten dieser Tätigkeit kommt gegenwärtig große Bedeutung zu (vgl. hierzu etwa ALTMANN 1978), und zwar deshalb, weil selbst solche Linguisten, die prinzipiell die Ziele der quantitativen Linguistik und damit die Notwendigkeit der Metrisierung anerkennen, oftmals keine hinreichend präzisen Vorstellungen darüber haben, auf welche Bedingungen man bei Metrisierung und Messung zu achten hat. Um das Bewußtsein für dieses methodologische Problem zu schärfen, kann man verschiedene Wege gehen. Eine Möglichkeit besteht darin, Begriffskritik zu üben, d.h. Metrisierungsvorschläge, wie sie in der wissenschaftlichen Literatur anzutreffen sind, insbesondere auf ihre formale

Stichhaltigkeit hin zu überprüfen (als Beispiel hierfür vgl. LEHFELDT 1978).

Eine andere Möglichkeit ist die, einen Metrisierungsvorschlag in statu nascendi zu beobachten, d.h. die Schritte nachzuvollziehen und zu analysieren, die nach der Meinung eines Autors u. a. zur Metrisierung bestimmter sprachlicher Größen bzw. Eigenschaften führen sollen. Diesen Weg wollen wir in der vorliegenden Arbeit gehen, wobei neben methodologischen Fragen auch solche empirischer Natur zu Sprache kommen sollen. Und zwar wollen wir uns hier anhand einiger Arbeiten von GERHART LINDNER zum Sprechbewegungsablauf mit der Aufgabe beschäftigen, eine Vorschrift zu entwickeln, wie man den Grad der Schwierigkeit oder der Kompliziertheit des Übergangs von einem Laut zum anderen beim Sprechen messen kann. Neben den praktischen, insbesondere den logopädischen Aspekten, denen LINDNERs Hauptaugenmerk gilt, ist diese Aufgabe auch unter theoretischen Fragestellungen interessant. So muß man sie beispielsweise gelöst haben, wenn man den artikulatorischen Kontrastaufbau solcher Einheiten wie Silben, Allomorphen oder Wortformen und seine Gesetzmäßigkeiten studieren will.

Wir gehen mit LINDNER von der Vorstellung aus, daß im normalen Sprechvorgang die Organe, die die an der Erzeugung lautsprachlicher Zeichen beteiligten Funktionsgruppen Atmung, Stimmgebung und Lautbildung konstituieren, sich in harmonischer Dauerbewegung befinden, daß die Sprechorgane aus den Stellungen, die für die Bildung eines bestimmten Lautes charakteristisch sind, übergehen in die Stellungen, die für die Bildung des diesem folgenden Lautes kennzeichnend sind, wobei "sich diese Bewegungen in klanglichen Übergängen von einem Laut zum anderen bemerkbar machen" (LINDNER 1961, 4; vgl. LINDNER 1969a, 450).

Der Gedanke liegt nahe, daß die Bewegungen derjenigen Organe, die den Funktionsgruppen Stimmgebung und Lautbildung zugerechnet werden, beim Übergang von einem Laut zum anderen verschieden schwierig oder kompliziert sind. Um dieser Vorstellung eine präzisere Gestalt zu geben, ist es erforderlich, die Schwierigkeit des Übergangs von Laut zu Laut meßbar zu machen. Es soll hier darum gehen, einige Schritte anzugeben, die zu diesem Ziel hinführen können, sowie einige Probleme zu diskutieren, deren Lösung noch aussteht.

Den Ausgangspunkt unserer Betrachtungen bildet die theoretische Analyse des Sprechbewegungsablaufes, die LINDNER (1969a, 1975) für das Deutsche erarbeitet hat. Das Ziel dieser Analyse besteht darin, ein Modell des Sprechbewegungsablaufes zu entwikkeln, das von den individuellen Besonderheiten verschiedener Sprecher abstrahiert, d.h., in dem die wirklichen Bewegungsabläufe in einer auf das funktional Wesentliche vereinfachten Form abgebildet werden (vgl. LINDNER 1975, 26 f.).

Das Objekt von LINDNERs modellhafter Untersuchung des Sprechbewegungsablaufes sind zweigliedrige Lautverbindungen, die "kleinsten Bewegungseinheiten des Sprechaktes" (1969a,451). Die einzelnen Laute, die eine zweigliedrige Lautverbindung konstituieren, sind hier natürlich wie die gesamte Analyse als Abstraktionen anzusehen, da "1. von den *Individualqualitäten* des Sprechers und 2. von *Varianten* abstrahiert wird, die sich um einen Mittelwert zentrieren. Beim Begriff des Lautes werden alle seine wesentlichen Bildungsmerkmale einbezogen, sowohl in genetischer als auch in gennematischer wie auch in perzeptiver Hinsicht" (LINDNER 1975, 68).

Die zweigliedrigen Lautverbindungen sind nach LINDNER (1961, 43) "das Abbild der wesentlichen Sprechbewegungen"; "in ihnen

sind diejenigen Bewegungen enthalten, die die Sprechorgane vollführen müssen, um aus der Stellung, die sie zur Bildung des ersten Lautes einnehmen mußten, in diejenige zu gelangen, die zur Bildung des zweiten Lautes notwendig ist" (LINDNER 1969a, 452; vgl. auch LINDNER 1969b, 22; 1975, 70). Es darf hier nicht übersehen werden, daß die Beschränkung auf die Betrachtung zweigliedriger Lautverbindungen auch innerhalb des Modells zu einer stark vereinfachten Darstellung des Sprechbewegungsablaufes führt. Es lassen sich auf dieser Grundlage nämlich bestenfalls Aussagen über die "unbedingt notwendigen Sprechbewegungen" (LINDNER 1975, 71) gewinnen, während all diejenigen Organbewegungen unberücksichtigt bleiben, die "als *koartikulatorisch gesteuerte Bewegungen* ... für die Aussprache des gesamten Zeichens notwendig sind" (LINDNER 1975, 71). Innerhalb der "gleichen" Lautverbindungen können die koartikulatorisch der Vorbereitung folgender Laute dienenden Organbewegungen verschieden sein, je nachdem, um was für Laute es sich handelt. Von all diesen Unterschieden wird innerhalb des Modells abstrahiert, so daß deutlich ist, daß es auch modellintern nur um eine erste Approximation an die Sprechwirklichkeit geht.

Um den Sprechbewegungsablauf zwischen je zwei Lauten modellhaft zu erfassen und zu beschreiben, geht LINDNER von der grundlegenden Annahme aus, daß es möglich sei, von der Kenntnis der Einstellungen, in denen sich die Sprechorgane bei der Erzeugung des ersten bzw. des zweiten Lautes befinden, auf die Bewegungen zu schließen, die die Sprechorgane beim Übergang von den ersten zu den zweiten Einstellungen vollziehen: "Jeder Laut verlangt zu seiner Produktion, daß die Sprechorgane ganz *bestimmte Einstellungen* einnehmen. Also müssen, wenn die Folgen von Lauten bekannt sind, die Bewegungen, die die Sprechorgane bei der Realisierung der Laut*folge* vollziehen, erschlossen werden können" (LINDNER 1975, 70). Es ist also notwendig, folgende Schritte nacheinander zu vollziehen: Zunächst muß für eine gegebene Sprache die Menge derjenigen Laute bestimmt werden, die überhaupt in die Analyse eingehen sollen. Anschließend sind die Sprechorgane festzulegen, deren Einstellungen für jeden einzelnen Laut beschrieben werden sollen. Für jedes Sprechorgan ist ein Einstellungsinventar anzugeben. Nachdem die Einstellungen aller Sprechorgane für alle Laute ermittelt sind, kann man sich der Aufgabe zuwenden, die Bewegungen zu erschließen, die zwischen jeweils zwei Lauten einer nach bestimmten Kriterien festgelegten Menge von zweigliedrigen Lautverbindungen vollzogen werden.

Was den ersten Schritt anlangt, so bezieht sich LINDNER auf "das für die deutsche Hochlautung typische Mindestinventar" (LINDNER 1975, 80), das 20 Vokale (einschließlich der Diphthonge) sowie 22 Konsonanten umfaßt (vgl. Tabelle 1). Im Hinblick auf den zweiten Schritt berücksichtigt LINDNER "die Organe, die zur Beschreibung der aktiven Bewegungen unmittelbar gebraucht werden" (LINDNER 1975, 127), v.a. diejenigen Organe des Ansatzrohres, "die sich aktiv bewegen können und durch ihre Bewegungen zur Bildung oder Auflösung von Sekundärschallquellen oder zur Veränderung des Hohlraumes des Ansatzrohres und damit *zur Klangveränderung aktiv beitragen*" (LINDNER 1975, 52 f.), d.h. neben der Glottis den Unterkiefer, die Lippen, die Zunge und das Gaumensegel. Die Zunge wird in die Organteile Zungenspitze, Zungenrücken, Medianfläche und Zungenrand unterteilt, wobei die Grenzen zwischen diesen Teilen "durch *funktionelle Besonderheiten* und ihre *Bedeutung* beim Sprechen *gesetzt*" (LINDNER 1975, 56) werden. Die Einstellungsinventare, die LINDNER den einzelnen Sprechorganen zuordnet, sind aus der linken Spalte von Tabelle 1

zu ersehen (zu Einzelheiten vgl. LINDNER 1975, 131-149).

Für die angestrebte Analyse des Sprechbewegungsablaufes ist weiterhin wichtig, daß LINDNER die Sprechorgane bezüglich eines jeden Lautes in zwei Gruppen einteilt, und zwar 1. in die Gruppe der an der Bildung eines Lautes *wesentlich* beteiligten Organe und 2. in die Gruppe der an der Bildung eines Lautes *nichtwesentlich* beteiligten Organe. "Die wesentlich beteiligten Organe *müssen* sich bei der Bildung eines bestimmten Lautes in einer bestimmten Einstellung befinden oder müssen diese Einstellung im Verlauf des Bewegungsvollzuges durchlaufen" (LINDNER 1975, 128; vgl. auch LINDNER 1969a, 453, 1969b, 167). So ist beispielsweise für das deutsche [a:] ein weiter Kieferwinkel wesentlich. Eine wesentliche Organeinstellung symbolisieren wir allgemein durch das Zeichen o.

Für die unwesentlich beteiligten Organe ist kennzeichnend, daß sie in ihrer Einstellung eine verhältnismäßig große Variationsbreite aufweisen. Um den Sprechbewegungsablauf so vollständig wie möglich zu erfassen, ist es wünschenswert, auch diese Organe in ihrer jeweiligen Einstellung festzuhalten, "aber in einer Form, die von vornherein mit einer großen Variationsbreite rechnet. In der Systematik wird dann diejenige Organeinstellung angegeben, die die unbeteiligten Organe wahrscheinlich einnehmen" (LINDNER 1969a, 453, vgl. auch LINDNER 1975, 128). Diese Art der Organeinstellung kennzeichnen wir allgemein durch das Symbol x. In Tabelle 1 sind für jeden Laut des von LINDNER der Analyse zugrundegelegten Inventars die wesentlichen und die nichtwesentlichen Organeinstellungen angegeben.

Die beschriebene Zweiteilung der Sprechorgane bezüglich eines Lautes ist die Grundlage für die Klassifikation der Bewegungsarten eines jeden Sprechorgans beim Übergang vom ersten zum

zweiten Laut. "Die Bewegungsarten werden von den Einstellungen, die die Sprechorgane am Anfang und am Ende der Lautverbindung einnehmen, abgeleitet" (LINDNER 1975, 150). Und zwar sind für jedes Organ insgesamt sechs Bewegungsarten bzw. Abläufe denkbar. Diese lassen sich schematisch wie folgt darstellen (der Pfeil symbolisiert den Übergang des Sprechorgans aus der ersten Einstellung in die zweite, die Indizes 1 und m dienen der Unterscheidung verschiedener Organeinstellungen):

1. *Präzise Bewegung*: $o_1 \longrightarrow o_m$ Das Organ bewegt sich aus einer wesentlichen Einstellung in eine von dieser verschiedene wesentliche Einstellung.
2. *Einseitig modifizierbare Bewegung* (auch *Zielbewegung* genannt): $o_1 \longrightarrow x_m$ oder $x_1 \longrightarrow o_m$ Das Organ bewegt sich aus einer wesentlichen Einstellung in eine von dieser verschiedene nichtwesentliche Einstellung oder umgekehrt.
3. *Modifizierbare Bewegung*: $x_1 \longrightarrow x_m$ Das Organ bewegt sich aus einer nichtwesentlichen Einstellung in eine von dieser verschiedene nichtwesentliche Einstellung.
4. *Präzise Übernahme*: $o_1 \longrightarrow o_1$ Das Organ befindet sich bei beiden Lauten der Lautfolge in der gleichen wesentlichen Einstellung.
5. *Einseitig modifizierbare Übernahme*: $o_1 \longrightarrow x_1$ oder $x_1 \longrightarrow o_1$ Die gleiche Organeinstellung ist für den ersten Laut wesentlich und für den zweiten nichtwesentlich oder umgekehrt.
6. *Modifizierbare Übernahme*: $x_1 \longrightarrow x_1$ Das Organ befindet sich bei beiden Lauten der Lautfolge in der gleichen nichtwesentlichen Einstellung (zu Einzelheiten vgl. LINDNER 1969a, 1975, 150-152).

Nach diesen Darlegungen können wir dazu übergehen, versuchsweise ein Maß der Schwierigkeit des Übergangs von einem Laut i

zu einem Laut j oder kürzer ein Maß der Übergangsschwierigkeit von i bezüglich j zu bestimmen. Wir sagen "versuchsweise", da das vorgeschlagene Maß in verschiedenen Hinsichten problematisch ist (s.u.) und daher vernünftigerweise vor allem als Ausgangspunkt für weiterführende Überlegungen und Analysen betrachtet werden sollte. Einige solcher Überlegungen wollen wir unten andeuten.

Wie bereits deutlich geworden ist, müssen wir den Übergang von einem Laut zum anderen als einen Komplex gleichzeitig sich vollziehender koordinierter Abläufe mehrerer Sprechorgane ansehen. In das Maß sollte daher zunächst die Zahl der verschiedenen Abläufe eingehen.

Angenommen, an der Bildung eines Lautes sind insgesamt S Sprechorgane wesentlich oder unwesentlich beteiligt (in unserem Falle ist S = 8). Dann können wir die Zahl der verschiedenen Abläufe pro Übergang wie folgt symbolisieren:

u_{ij} = Zahl der präzisen Bewegungen

v_{ij} = Zahl der einseitig modifizierbaren Bewegungen

w_{ij} = Zahl der modifizierbaren Bewegungen

x_{ij} = Zahl der präzisen Übernahmen

y_{ij} = Zahl der einseitig modifizierbaren Übernahmen

z_{ij} = Zahl der modifizierbaren Übernahmen

Da für jeden Übergang alle einmal festgelegten Sprechorgane berücksichtigt werden, ergibt sich trivialerweise, daß

$$u_{ij} + v_{ij} + w_{ij} + x_{ij} + y_{ij} + z_{ij} = S$$

immer gilt, für beliebige Laute i und j.

Des weiteren ist folgender Umstand in Rechnung zu stellen: Die einzelnen Abläufe sind verschieden schwierig, bzw. sie haben, wie LINDNER (1969a, 455) sich ausdrückt, "einen unterschiedlichen Grad an Verbindlichkeit". Aus diesem Grunde ist es erforderlich, jeden Ablauf mit einem Faktor zu gewichten, der seine Schwierigkeit im Verhältnis zu den anderen Abläufen zum Ausdruck bringt.

LINDNER spricht den Bewegungen insgesamt eine höhere Bedeutsamkeit für den Sprechbewegungsprozeß zu als den Übernahmen, da in diesen "keine motorische Aktivität enthalten zu sein braucht" (LINDNER 1969a, 456), nach ihm ist "die *Bewegung* im Grad der Verbindlichkeit in jedem Fall höher als die Übernahme einzuschätzen" (LINDNER 1975, 153; vgl. auch LINDNER 1969b, 178). Wir gewichten daher die Bewegungen insgesamt stärker als die Übernahmen. Der Vorschlag von MÜLLER (1973, 51), jede Übernahme jeweils nur um einen Schwierigkeitsgrad tiefer als die entsprechende Bewegung einzustufen, erscheint uns wenig plausibel.

Für die weiteren Überlegungen ist noch folgendes festzuhalten: Was wir im Moment tun können, ist, die einzelnen Abläufe in eine Rangordnung zu bringen, aus der hervorgeht, welcher von je zwei Abläufen dem jeweils anderen hinsichtlich der Schwierigkeit vorangeht. Diese Rangordnung entspricht in ihrer formalen Struktur einer sogenannten Quasireihe, die das Ergebnis der Anwendung eines komparativen oder topologischen Begriffes auf alle Paare von Elementen eines bestimmten Gegenstandsbereiches bildet (vgl. hierzu STEGMÜLLER 1970, 27-44). Das bedeutet insbesondere, daß die Rangordnung nichts über die Größe des Abstandes zwischen den Rängen aussagt. Hierin liegt ein Problem; denn die Schwierigkeit der Abläufe zu gewichten bedeutet, diesen bestimmte Zahlen zuzuschreiben. Diese Zahlen sagen aber nicht nur etwas über den Rang aus, sondern auch über die Größe des Abstandes zwischen den verschiedenen Rängen, obgleich wir über den dazu erforderlichen quantitativen Begriff noch gar nicht verfügen, obgleich noch gar

keine Metrik existiert. Wenn wir im weiteren die Gewichtung in der Weise vornehmen, daß die Abstände zwischen allen zwei Abläufen, die in der Rangordnung unmittelbar aufeinander folgen, als gleich groß aufgefaßt werden, so müssen wir uns darüber im klaren sein, daß diese Art des Vorgehens gewissermaßen eine Verlegenheitslösung ist, die nur vorläufig akzeptiert werden kann, lediglich solange, wie es nicht gelingt, eine geeignete Metrik zu entwicklen.

Von den Bewegungen sind die präzisen am schwierigsten, da sie "beim Sprechen höchste Präzision für ihre Ausführung" (LINDNER 1969a, 456) verlangen. Insgesamt unterscheiden wir sechs Schwierigkeitsstufen; von diesen weisen wir den präzisen Bewegungen die höchste mit dem Gewichtungsfaktor 6 zu.

Beispiel: Beim Übergang von [f] zu [ɑ:] in der Lautfolge [fɑ:] wie in dem Wort *fahren* geht der Unterkiefer aus einer wesentlichen engen Einstellung in eine wesentliche weite Einstellung über. Diesen Ablauf gewichten wir mit dem Faktor 6.

Weniger schwierig sind die einseitig modifizierbaren Bewegungen, da bei ihnen nur der Anfangs- oder der Endpunkt des Bewegungsvollzugs festliegt. Sie erhalten den Gewichtungsfaktor 5.

Beispiel: Beim Übergang von [l] zu [b] in der Lautfolge [lb] wie in dem Wort *Kälber* geht die Zungenspitze aus einer wesentlichen gehobenen Einstellung in eine nichtwesentliche gesenkte Einstellung über. Diesen Ablauf gewichten wir mit dem Faktor 5.

Den modifizierbaren Bewegungen, deren gesamter Ablauf koartikulatorisch modifiziert werden kann, wird der Gewichtungsfaktor 4 zugeteilt.

Beispiel: Beim Übergang von [ø:] zu [f] in der Lautfolge [ø:f] wie in dem Wort *Höfe* geht der Zungenrand aus einer nichtwesentlichen gehobenen Einstellung in eine nichtwesentlich gesenkte

Einstellung über. Diesen Ablauf gewichten wir mit dem Faktor 4.

In analoger Weise werden die Übernahmen gwichtet: Die präzise Übernahme erhält den Faktor 3, die einseitig modifizierbare Übernahme den Faktor 2 und die modifizierbare Übernahme den Faktor 1.

Beispiel: Beim Übergang von [k] zu [t] in der Lautfolge [kt] wie in dem Wort *Kaktus* verharrt das Velum in einer wesentlichen gehobenen Einstellung. Diese Übernahme gewichten wir mit dem Faktor 3.

Beispiel: Beim Übergang von [n] zu [s] in der Lautfolge [ns] wie in dem Wort *Hans* bleibt der Zungenrücken prädorsal gehoben, wobei diese Einstellung für [n] unwesentlich, für [s] wesentlich ist. Diese Übernahme gewichten wir mit dem Faktor 2.

Beispiel: Beim Übergang von [k] zu [t] in der Lautfolge [kt] verharren die Lippen in einer nichtwesentlichen unbeteiligten Einstellung. Diese Übernahme gewichten wir mit dem niedrigsten Faktor 1.

Das gesuchte vorläufige Maß der Übergangsschwierigkeit U_{ij} kann nunmehr bestimmt werden als die Summe der Anzahlen der einzelnen Abläufe pro Übergang, wobei jeder einzelne Ablauf mit dem ihm zugewiesenen Schwierigkeitsfaktor gewichtet wird:

$$U_{ij} = 6u_{ij} + 5v_{ij} + 4w_{ij} + 3x_{ij} + 2y_{ij} + 1z_{ij}$$

Da die Zahl der Sprechorgane, die bei einem Übergang berücksichtigt werden, für eine gegebene Sprache immer gleich ist, sind die Werte, die U_{ij} annehmen kann, unmittelbar miteinander vergleichbar, d.h., es ist nicht nötig, sie zu relativieren. Der Wertebereich von U_{ij} kann als $\langle 8,48 \rangle$ bestimmt werden. Natürlich sind die Intervallgrenzen nur theoretische Bezugsgrößen, da es weder einen Übergang gibt, der ausschließlich aus modifizierba-

ren Übernahme bestünde, noch einen solchen, der lediglich durch präzise Bewegungen konstituiert wäre.

Beispiel: Es soll der Wert des Maßes der Übergangsschwierigkeit zwischen den Lauten [a:] und [n] errechnet werden. Aus Tabelle 1 erhalten wir folgende Einzelwerte:

$$u_{na:} = 2$$
$$v_{na:} = 3$$
$$w_{na:} = 1$$
$$x_{na:} = 1$$
$$y_{na:} = 0$$
$$z_{na:} = 1$$

Es ergibt sich demnach:

$$U_{na:} = 6(2) + 5(3) + 4(1) + 3(1) + 2(0) + 1(1) = 35$$

Wie man sich leicht überzeugt, ist $U_{na:} = U_{a:n}$.

Man kann das vorgeschlagene Maß zum Ausgangspunkt nehmen, um die Schwierigkeit des Sprechbewegungsablaufes in bezug auf höhere Einheiten wie zum Beispiel Silben oder Wortformen zu berechnen. Die Überlegung ist einfach: Man summiert die Werte von U_{ij} für alle aufeinanderfolgenden Lautpaare und dividiert die Summe durch die Anzahl der Lautpaare. Wenn die in Rede stehende Einheit k Laute umfaßt, wobei nach der Analyse von LINDNER Diphthonge jeweils als zwei Laute gezählt werden müssen, dann haben wir insgesamt k-1 Lautpaare, so daß wir zu folgendem Maß gelangen:

$$U_E = \frac{\Sigma U_{ij}}{k-1} \qquad E = \text{Silbe, Allomorph, Wortform, ...}$$

Auch dieses Maß hat den Wertebereich <0;48>.

Beispiel: Es soll der Wert des Maßes der Schwierigkeit des Sprechbewegungsablaufes für das Wort [mɪlç] *Milch* berechnet werden. Zuerst sind die Werte für $U_{mɪ}$, $U_{ɪl}$ und $U_{lç}$ zu ermitteln. Aus Tabelle 1 erhalten wir folgende Einzelwerte:

$u_{mɪ} = 2$	$u_{ɪl} = 0$	$u_{lç} = 4$
$v_{mɪ} = 1$	$v_{ɪl} = 4$	$v_{lç} = 0$
$w_{mɪ} = 2$	$w_{ɪl} = 1$	$w_{lç} = 2$
$x_{mɪ} = 1$	$x_{ɪl} = 2$	$x_{lç} = 1$
$y_{mɪ} = 1$	$y_{ɪl} = 1$	$y_{lç} = 1$
$z_{mɪ} = 1$	$z_{ɪl} = 0$	$z_{lç} = 0$

Wir erhalten daraus:

$$U_{mɪ} = 6(2) + 5(1) + 4(2) + 3(1) + 2(1) + 1(1) = 31$$
$$U_{ɪl} = 6(0) + 5(4) + 4(1) + 3(2) + 2(1) + 1(0) = 32$$
$$U_{lç} = 6(4) + 5(0) + 4(2) + 3(1) + 2(1) + 1(0) = 37$$

Da im vorliegenden Fall k = 4, ergibt sich, daß

$$U_{mɪlç} = \frac{U_{mɪ} + U_{ɪl} + U_{lç}}{4 - 1} = \frac{31 + 32 + 37}{3} = \frac{100}{3} = 33.33$$

Es versteht sich, daß das Maß U_E mindestens genauso problematisch ist wie das ihm zugrundeliegende Maß U_{ij}. Einige der hier zu berücksichtigenden problematischen Punkte sind bereits angesprochen worden: Vernachlässigung der koartikulatorisch gesteuerten Abläufe, Gleichgewichtung aller Bewegungsarten und, damit zusammenhängend, "Vortäuschung" einer in Wirklichkeit noch nicht vorhandenen Metrik. Der Hinweis auf diese Punkte hat natürlich

im Zusammenhang unserer Analyse einzig die Funktion, die Aufmerksamkeit auf solche Probleme modellinterner wie -externer Natur zu lenken, die in Zukunft unbedingt gelöst werden müssen, wenn wir zu einer auch nur einigermaßen befriedigenden und brauchbaren Metrisierung der Schwierigkeit des Sprechbewegungsablaufes gelangen wollen. In genau diesen Zusammenhang wollen wir auch die Besprechung einiger weiterer problematischer Punkte eingeordnet wissen, auf die LINDNER z.T. schon kurz eingegangen ist.

In unserem Modell werden, wie ausgeführt, die Abstände zwischen allen zwei Abläufen, die in der Rangordnung unmittelbar aufeinanderfolgen, als gleich groß aufgefaßt. Dieses Vorgehen impliziert, daß wir jeden der sechs möglichen Abläufe jedesmal mit dem gleichen, ein für allemal festgelegten Faktor gewichten. Gegen diesen modellinternen Schematismus kann folgende Überlegung geltend gemacht werden: Da, wie LINDNER (1975, 35 f.) sagt, "die spezifische Art und Weise der Bewegung *eines bestimmten Organs* von der Art und Weise der Bewegung anderer Sprechorgane *beeinflußt* ... wird", da m.a.W. "der Bewegungsablauf eines jeden Organs immer *in Relation zu den übrigen* betrachtet werden" muß, sollte berücksichtigt werden, daß "ein und derselbe" Ablauf verschieden schwierig bzw. kompliziert sein kann, je nachdem, in welche Abläufe er eingebettet ist.

Ferner sollte bedacht werden, daß für Sprechorgane, "deren *Bewegungsinventar* sich in zwei möglichen Bewegungen erschöpft: Senkung oder Hebung" (LINDNER 1975, 176), der "gleiche" Ablauf u.U. weniger schwierig sein kann als für ein Organ, das "ein Inventar von mehreren Bewegungen beherrschen muß, um aus einer bestimmten Einstellung eine andere Einstellung exakt und ohne Umwege direkt zu erreichen" (das.). Schließlich sollte auch nicht

vergessen werden, daß solche Abläufe eines bestimmten Typs, die sehr häufig vollzogen werden (z.B. die präzisen Bewegungen der Glottis und des Gaumensegels), aufgrund ihres guten Eingeübtseins von den Sprechern einer Sprache als weniger schwierig eingestuft werden können als seltener vollzogene Abläufe desselben Typs (vgl. hierzu LINDNER 1975, 179).

Nach all dem Gesagten dürfte ohne weiteres klar sein, daß unser Maß U_{ij} eine in hohem Maße provisorische, vorläufige Größe ist, die nur unter Vorbehalt akzeptiert und verwendet werden darf. Dennoch ist es keineswegs überflüssig gewesen, sich mit ihm zu beschäftigen. So wie ganz allgemein komparative Begriffe historisch oft eine Übergangsstufe zwischen den primitiveren qualitativen und den komplizierteren quantitativen Begriffen bilden (vgl. STEGMÜLLER 1970, 29), so können wir U_{ij} als eine Art Zwischenergebnis auffassen. Sein methodologischer Wert liegt v.a. darin, daß die Beschäftigung mit ihm uns die Gelegenheit gab, in zusammenhändender Form einige der Probleme zu erörtern, die gelöst werden müssen, um ein besseres Maß zu gewinnen.

Literatur

ALTMANN, G., Zur Verwendung der Quotiente in der Textanalyse. In: ALTMANN, G. (Ed.), Glottometrika 1. Bochum, Brockmeyer 1978, 91-106.

LEHFELDT, W., Zur Messung der phonetischen Lautdifferenz - Eine begriffskritische Untersuchung. In: ALTMANN, G. (Ed.), Glottometrika 1, 26-45.

LINDNER, G., Untersuchungen über den zeitlichen Verlauf des Sprechens und ihre Anwendung auf die Pädagogik Hörgeschädigter. Habilschrift Humboldt-Universität Berlin, 1961.

LINDNER, G., Theoretische Analyse des Sprechbewegungsablaufs. Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 22, 1969(a), 450-473.

LINDNER, G., Einführung in die experimentelle Phonetik. München 1969(b).

LINDNER, G., Der Sprechbewegungsablauf - Eine phonetische Studie des Deutschen. Berlin 1975.

MEL'ČUK, I.A., Several types of linguistic meanings. In: AKHMANOVA, O.S. et al., Exact Methods in Linguistic Research. Berkeley and Los Angeles 1963(a), 36-43.

MEL'ČUK, I.A., O standartnoj forme i količestvennyx xarakteristikax nekotoryx lingvističeskix opisanij. In: Voprosy jazykoznanija 1963(b)/1, 113-123.

MEL'ČUK, I.A., Opyt razrabotki fragmenta sistemy ponjatij i terminov dlja morfologii (k formalizacii jazyka lingvistiki). In: Semiotika i informatika 6, 1975, 5-50.

MEL'ČUK, I.A., K postroeniju sistemy ponjatij dlja morfologii. In: Studia linguistica Alexandro Vasilii filio Issatschenko a collegis amicisque oblata. Lisse 1978, 267-287.

MÜLLER, U., Zur Bestimmung der Rangfolge des Schwierigkeitsgrades von Lauten der deutschen Sprache. In: Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 26, 1973, 49-60.

SOVA, L.Z., Analitičeskaja lingvistika. Moskva 1970.

STEGMÜLLER, W., Theorie und Erfahrung. Berlin 1970.

ŠELOV, S.D., O semantičeskom opisanii pri sozdanii informacionnogo jazyka (na materiale lingvistiki). In: Naučno-texničeskaja informacija. Serija 2: Informacionnye processy i sistemy, 1974/6, 10-16.

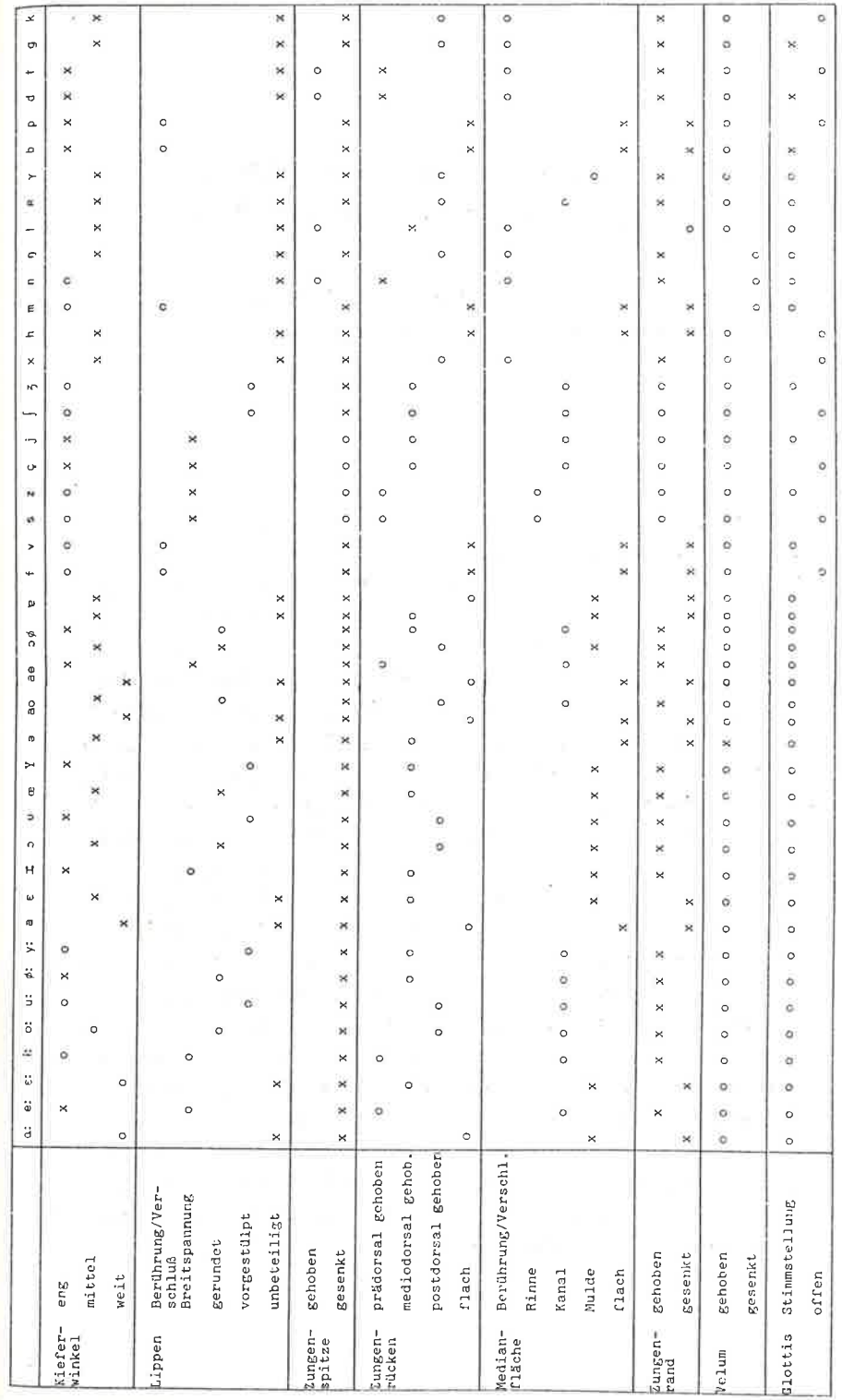


Tabelle 1: Einstellungen der Sprechorgane bei der Artikulation der Vokale und der Konsonanten der deutschen Hochlautung (nach LINDNER 1975).

On the Distance of Japanese Words Based on Distinctive Features and a Second-Order Model

Shochi YOKOYAMA and Shuichi ITAHASHI,
Electrotechnical Laboratory, Tokyo

INTRODUCTION

It is a big problem in the study of automatic speech recognition what sort of parameters should be used for matching with the standard patterns. In modeling the speech process including language level, it is difficult to determine the optimum model only from the speech analysis point of view.

One way would be to use various phonetic features which have been studied by phoneticians. The word level is assumed as the starting point here, and the features are used to describe words so that the relation among words can be investigated. In this area, there is a study [1] on the influence of phoneme confusion on the discrimination between words in a given set. However, on the discrete level the complete segmentation of a word into phonemes is usually presupposed. In order to take account of the segmentation error (addition or deletion of phonemes) it is necessary to adopt a different method such as dynamic programming on the phonemic level [2].

In this paper, Japanese word distances are investigated on the continous parameter level, in order to take account of the influence of the segmentation error automatically. That is, the parameters are obtained in the following sequence:

a word → phonemes → distinctive features → a model → continous parameters. The step response [3] of a second-order linear system is adopted as a model to transform distinctive features into continous parameters.

Part of the contents of this paper has already been published [4]. In this paper, the number of words concerned has been increased considerably (to about 4 times, or to nearly 16 times in the number of word pairs) and the model has been improved.

MATERIAL

It is desirable to select frequently used words. For this purpose the "Shin Meikai Kokugo Jiten" (New Concise Japanese Dictionary) [5] was chosen, which is currently being inserted into a data base in our laboratory [6]. The entry words are extracted from the most important words (marked with ** in the dictionary). However, the accent of each word is not used as a parameter at the moment. For the details of the data structure refer to (6).

Entry words and accents are both stored in the disk storage (TOSBAC-5600) in the form of KANJI (Chinese charakter) code, which is transformed into a form acceptable by the computer. Entry words are romanized from the corresponding KANA (Japanese syllabic alphabet) code. Regarding the phonemics of Japanese, the following rules are used in the transcription into alphabetic representation:

1) The Japanese method of Romanization (see Table 1) is used as a rule:
   Ex: あか → aka
2) Syllabic nasal ん is represented by 'N' to distinguish it from な (na)-line sound.
   Ex: こんなん → koNnaN
3) Contracted sound is represented by the Japanese method and treated in the same manner as や (ya)-line sound.
   Ex: じゃんじょ → zyuNzyo
4) Double consonant symbol っ is represented by 'q'.
   Ex: かって → kaqte

Table 1. A Romanization of Japanese according to the Japanese method.

あ	a	い	i	う	u	え	e	お	o
か	k a	き	k i	く	k u	け	k e	こ	k o
さ	s a	し	s i	す	s u	せ	s e	そ	s o
た	t a	ち	t i	つ	t u	て	t e	と	t o
な	n a	に	n i	ぬ	n u	ね	n e	の	n o
は	h a	ひ	h i	ふ	h u	へ	h e	ほ	h o
ま	m a	み	m i	む	m u	め	m e	も	m o
や	y a			ゆ	y u			よ	y o
ら	r a	り	r i	る	r u	れ	r e	ろ	r o
わ	w a	ゐ	w i			ゑ	w e	を	w o
ん	n								
が	g a	ぎ	g i	ぐ	g u	げ	g e	ご	g o
ざ	z a	じ	z i	ず	z u	ぜ	z e	ぞ	z o
だ	d a	ぢ	z i	づ	z u	で	d e	ど	d o
ば	b a	び	b i	ぶ	b u	べ	b e	ぼ	b o
ぱ	p a	ぴ	p i	ぷ	p u	ぺ	p e	ぽ	p o
きゃ	k y a		きゅ	k y u		きょ	k y o		
しゃ	s y a		しゅ	s y u		しょ	s y o		
ちゃ	t y a		ちゅ	t y u		ちょ	t y o		
にゃ	n y a		にゅ	n y u		にょ	n y o		
ひゃ	h y a		ひゅ	h y u		ひょ	h y o		
みゃ	m y a		みゅ	m y u		みょ	m y o		
りゃ	r y a		りゅ	r y u		りょ	r y o		
ぎゃ	g y a		ぎゅ	g y u		ぎょ	g y o		
じゃ	z y a		じゅ	z y u		じょ	z y o		
ぢゃ	z y a		ぢゅ	z y u		ぢょ	z y o		
びゃ	b y a		びゅ	b y u		びょ	b y o		
ぴゃ	p y a		ぴゅ	p y u		ぴょ	p y o		

5) /ou/, /ei/ sounds are represented by /o-/, /e-/ as elongated sounds. But those clearly pronounced /ou/ are represented as they are. Vowel duplications such as /oo/ and /ee/ are sometimes transcribed into elongated sounds to take account of the actual pronunciation.

Ex: かよう (Tuesday) → kayo-
Ex: かよう (to go) → kayou

6) Of the た(ta)-line sounds 'ち' and 'つ' are represented as /ci/ and /cu/ respectively. Their voiced sounds are represented as /zi/ and /zu/, not /di/ or /du/.

Table 2 shows a part of the entry words represented according to the above mentioned rules. Entries in the Table 2 are the serial number, the number of phonemes in each word, and the accent type. The number of the words is 1755 which excludes homonyms because the accent type is not used to distinguish them in this paper. Some words which are not classified as the most important are included in this set due to error in data base creation.

Of these 1755 words, Table 3 shows the number of words counted by the number of phonemes and the number of syllables. As seen from the table, 6-phoneme words are the most frequent in terms of the number of phonemes; 2- and 3-syllable words are the most frequent in terms of the number of syllables. The word having the maximum number of phonemes is 'じゅういちがつ'/zyu-icigacu/ (November) with 11 phonemes, and the shortest words are 'い'/i/ and 'え'/e/, consisting of one vowel.

DISTINCTIVE FEATURES

A phoneme sequence of a word thus obtained is then transformed into a matrix of distinctive features [7]. The distinctive features of the phonemes are shown in Fig. 1. That is, 12 distinctive features ranging from 'mora' to 'nasal' are assigned to each phoneme. The same distinctive features are assigned for /c/ and /t/ at present.

Table 2: A part of the wordlist

No.	*	**	words
122	5	2	itoko
123	7	3	itonamu
124	5	0	inaka
125	5	1	inoci
126	5	2	ibaru
127	4	0	ihaN
128	3	1	ima
129	3	1	imi
130	6	4	imo-to
131	3	1	iya
132	6	2	iyoiyo
133	4	0	irai
134	7	0	iriguci
135	3	0	iru
136	5	0	ireru
137	3	2	iro
138	6	0	iroiro
139	3	2	iwa
140	4	2	iwau
141	6	0	iNsacu
142	6	0	iNsyo-
143	2	0	ue
144	4	0	ueru
145	6	0	ukagau
146	5	0	ukabu
147	3	0	uku

* No. of phonemes, ** accent type

Table 3: Number of words counted with (a) phonemes and (b) syllables

(a)

1	2
2	30
3	87
4	295
5	372
6	536
7	256
8	142
9	28
10	6
11	1

(b)

1	49
2	740
3	716
4	225
5	24
6	1

	p	b	t	d	k	g	s	z	m	n	r	h	N	q	w	y	i	e	a	o	u
(mora)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+
(vocalic)	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	+	+	+	+	+
(consonantal)	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-
(high)	-	-	-	-	+	+	-	-	-	-	-	-	-	-	+	+	+	-	-	-	+
(back)	-	-	-	-	+	+	-	-	-	-	-	-	-	-	+	-	-	-	+	+	+
(low)	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	+	-	-
(anterior)	+	+	+	+	-	-	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-
(coronal)	-	-	+	+	-	-	+	+	-	+	+	-	-	-	-	-	-	-	-	-	-
(obstruent)	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-
(voiced)	-	+	-	+	-	+	-	+	+	+	+	-	+	-	+	+	+	+	+	+	+
(continuant)	-	-	-	-	-	-	+	+	-	-	-	+	-	-	+	+	+	+	+	+	+
(nasal)	-	-	-	-	-	-	-	-	+	+	-	-	+	-	-	-	-	-	-	-	-

Fig. 1 Distinctive features of phonemes

Some of the '+' and '-' in Fig. 1 could be derived from the others by phonological rules [7] (for example, if 'consonantal' is '-', 'obstruent' must be invariably '-'). A value '1' is assigned if its distinctive feature is '+' and '0' if '-', to obtain a 12-dimensional (12-bit) 1-0 pattern for one phoneme.

Each phoneme is assigned a duration as shown in Table 4. This duration is mostly based on the mean values of experimental data [8] and corresponds to the duration of the phoneme when a speech is uttered at an ordinary speed. As is apparent from Table 4, /t/ and /c/, which are given the same distinctive features, have different duration.

Since the phoneme at the initial position of a word is generally uttered short and the word final phoneme long, the duration of a word initial sound is 0.7 times as long as that in Table 4. The duration of a word final sound is 1.5 times as long as that in Table 4. The elongated sound symbol makes the duration of the preceding vowel 1.5 times as long.

Table 4: Duration of phonemes in ms

p	b	t	d	k	g	s	z	m	n	r
43	48	59	47	65	83	92	62	49	51	28
h	N	q	w	y	c	i	e	a	o	u
75	96	96	84	33	47	71	77	77	76	53

A 12-dimensional step sequence in the form of a time series is obtained in terms of these distinctive features and the duration for the phonemic representation of each word. Fig. 2 shows the example for /aka/ (red). In this figure, the hill corresponds to the distinctive feature '+' and the valley to '-'.

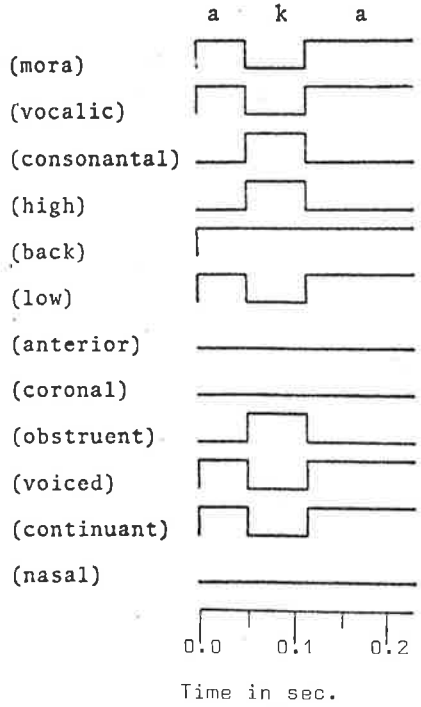


Fig. 2 Input sequence of /aka/ (red).

SECOND-ORDER LINEAR SYSTEM MODEL

The time series of Fig. 2 is used as input to the second-order linear system [3] of critical damping.

$$\left.\begin{aligned} f_k(t) &= f_k(t;t_0,\ldots,t_{k-1},F^{(1)}\ldots,F^{(k)}) \\ &= f_{k-1}(t) + \left(F^{(k)} - F^{(k-1)}\right)\left\{1 - \left(1 + \frac{t-t_{k-1}}{\tau}\right)\right. \\ &\qquad \left.\exp\left(-\frac{t-t_{k-1}}{\tau}\right)\right\} u_{-1}(t-t_{k-1}) \\ f_1(t) &= f_1(t;t_0,F^{(1)}) = F^{(1)},(t_0=0) \text{ for } k = 1 \end{aligned}\right\} \quad (1)$$

where f(t) is the output at time t and $F^{(k)}$ is the input (1 or 0) at time t_{k-1} (k = 2,3...,N), the input coming in succesively at each time period. τ is the time constant of the system and u_{-1} the unit step function.

Eq (1) may be represented in the form of Z-transform as follows:

$$\left.\begin{aligned} f_k(t) &= f(t) \\ f(t) &= (1-r)^2 F + 2rf(t-T) - r^2 f(t-2T) \\ F &= F^{(k)} \quad (t_{k-1} \leq t < t_k) \\ r &= \exp(-T/\tau) \end{aligned}\right\} \quad (2)$$

where T is the sampling rate.

The sampling time is chosen to be 10 ms. The feature 'mora' is given 5 ms as the time constant; 'nasal' is given 10 ms because it changes relatively fast experimentally, and the other 10 distinctive features are given 30 ms respectively. The latter value fits to the pattern very well when the formant frequency [3] is used as parameter.

Fig. 3 shows, for some words, the step input of the distinctive features and the output response of the second-order model. A small discrepancy between the rise of continuous parameter and that of step input is due to the difference between the sampling rate and the timing of the step input. No calcu-

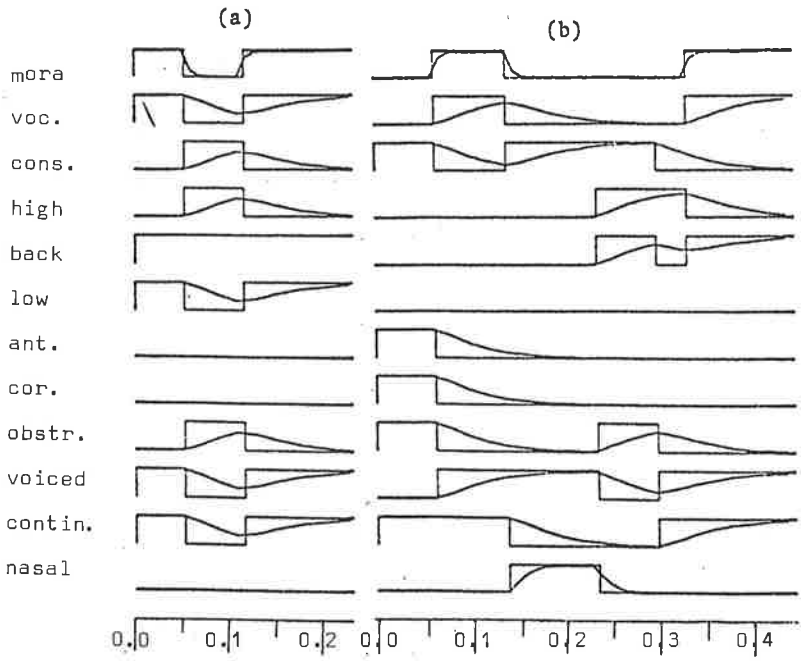
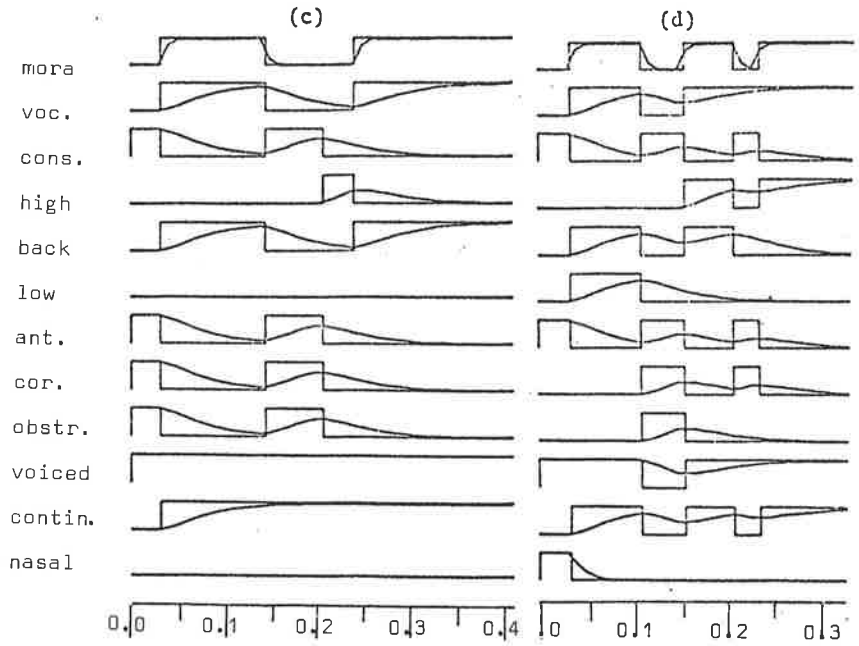


Fig. 3 Time patterns of distinctive features of the words (a) /aka/ (red), (b) /seNkyo/ (election), (c) /do-zyo-/ (sympathy) and (d) /macuri/ (festival)

Fig. 3 (continued)



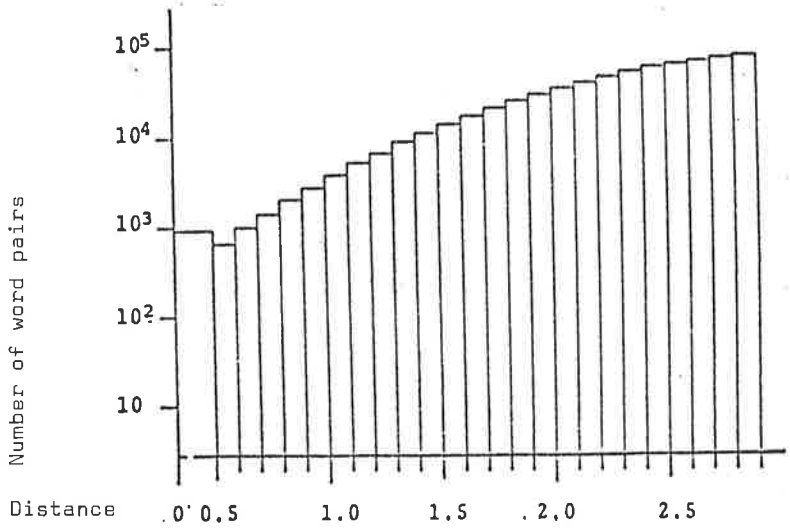


Fig. 4 Distribution of distances of word pairs

lation has been done for the discrepant portion. For example, the number of calculations (number of frames) for the word of duration 295 ms is 29.

DYNAMIC PROGRAMMING

The distance between the words is calculated using the continuous pattern thus obtained. Every word has the same dimension or 12 distinctive features, but its number of frames is different. The distance is calculated by a method of dynamic programming (abbreviated as DP), so that nonlinear matching of those words could be performed [9]. The details are in the reference. A brief description is given in the following.

Let W_1 be a word consisting of I frames and W_2 a word consisting of J frames. Let us denote By $\underline{a}_i$ the i-th frame 12-dimensional vector of W_1, and by $\underline{b}_j$ the j-th frame 12-dimensional vector of W_2. Let the distance between $\underline{a}_i$ and $\underline{b}_j$ be

$$d(i,j) = \| \underline{a}_i - \underline{b}_j \| \tag{3}$$

Let us denote the combination of the time by c_k

$$c_k = (i_k, j_k) \qquad k = 1,2\ldots,K \quad (K = I \text{ or } J) \tag{4}$$

where $c_1 = (1,1)$ and $c_K = (I,J)$
and $i_m - j_m \leqq 5 \quad (i_m \neq I,\ j_m \neq J)$

With these notations, the distance between two words becomes:

$$D(W_1,W_2) = \frac{1}{N} \min_c \left(\sum_{k=1}^{K} d(c_k) \cdot \Delta_k \right) \tag{5}$$

where

$$N = \sum_{k=1}^{K} \Delta_k = I + J - 1 \tag{6}$$

Δ_k is the distance along a path tracing the error and has the value one if proceeding by one in the direction of i_k, the value one if proceeding by one in the direction of j_k, the value two if proceeding by one both in direction of i_k and j_k (in the oblique direction). Other directions are not permitted with the function.

In the actual computation using the computer, a table for c_k is prepared and the following recursive expressions are used.

$$g(i,j) = \min \begin{pmatrix} d(i,j) + g(i-1,j) \\ d(i,j) + g(i,j-1) \\ 2d(i,j) + g(i-1,j-1) \end{pmatrix} \tag{7}$$

$$g(1,1) = d(1,1) \tag{8}$$

$$D(W_1,W_2) = \frac{1}{N} g(I,J) \tag{9}$$

RESULTS

The distances of pairs of words selected from the set of 1755 words are calculated based on eqs (5) and (9). Since the total number of the word pairs is

$$1755C_2 = 1755 \cdot 1754 \div 2 = 1539135 \tag{10}$$

the following criterion is used to reduce the number of calculations. Namely, the distances for only the pairs (W_1,W_2) satisfying the relation

$$\left.\begin{array}{ll} P_1 \leqq 2P_2 & (P_1 \geqq P_2) \\ P_2 \leqq 2P_1 & (P_1 < P_2) \end{array}\right\} \tag{11}$$

are calculated, where P_1 is the number of phonemes of word W_1 and P_2 the number of phonemes of word W_2. For example, the distances from a word having 4 phonemes are calculated for only those words having 2 to 8 phonemes. Distances from the words having one phoneme or over 8 phonemes are not calculated. The number of word pairs is thus reduced to 1447091 as seen from Table 3.

Fig. 4 shows the distribution of the number of word pairs up to the distance $D(W_1,W_2) \leqq 3.2$. The distribution is approximately normal.

Table 5 shows the calculated distances (less than 1.5) of the word /kioku/ (memory) (5 phonemes, 3 syllables) from the

Table 5: List of words with short distance (less than 1.5) from /kioku/ (memory)

No.	a	b	word	c	d	dist.	No.	a	b	word	c	d	dist.
283	7	3	kaikyu-	2	0	1.092	509	7	3	ke-kaku	2	0	1.192
284	6	3	kaike-	1	0	1.332	510	5	2	ke-ki	0	1	0.602
286	6	3	gaiko-	1	0	1.268	512	6	2	ke-ko-	1	1	0.802
287	7	4	gaikoku	2	1	1.499	519	7	3	ke-yaku	2	0	1.381
293	3	2	kau	2	1	1.466	520	4	2	kega	1	1	1.227
302	4	2	kagi	1	1	1.445	521	4	2	geki	1	1	1.054
305	4	2	kaku	1	1	1.362	529	8	3	keqkyoku	3	0	1.498
317	4	2	kako	1	1	1.486	530	6	2	keqko-	1	1	1.335
359	5	3	kayou	0	0	1.451	540	4	2	keru	1	1	1.306
396	5	3	kieru	0	0	1.189	546	7	2	keNkyu-	2	1	1.446
398	5	3	kikai	0	0	1.436	547	6	2	keNko-	1	1	1.495
399	4	2	kiku	1	1	1.137	559	3	2	koi	2	1	1.329
402	5	2	kiko-	0	1	1.260	560	3	2	kou	2	1	1.132
403	5	2	kigo-	0	1	1.250	561	4	2	ko-i	1	1	1.350
425	5	2	kibo-	0	1	1.490	563	5	2	ko-ka	0	1	0.814
433	5	2	kyaku	0	1	0.792	564	7	3	go-kaku	2	0	1.420
434	5	2	gyaku	0	1	1.012	566	7	2	ko-gyo-	2	1	1.098
435	4	1	kyu-	1	2	1.167	567	6	2	go-ke-	1	1	1.323
436	4	1	kyo-	1	2	1.235	568	7	3	ko-geki	2	0	1.482
437	7	3	kyo-iku	2	0	1.093	569	7	3	ko-koku	2	0	1.059
438	6	2	kyo-gi	1	1	0.791	579	6	2	ko-cu-	1	1	1.091
439	6	2	gyo-gi	1	1	1.113	580	6	2	ko-do-	1	1	1.431
442	7	2	kyo-cu-	2	1	1.031	582	5	2	ko-ba	0	1	1.423
443	7	2	kyo-do-	2	1	1.477	587	3	2	koe	2	1	1.471
444	6	2	kyo-mi	1	1	1.495	588	5	3	koeru	0	0	1.450
446	5	2	kyoka	0	1	0.779	590	5	2	ko-to	0	1	1.276
450	5	3	kirau	0	0	1.433	592	5	2	ko-ru	0	1	1.352
451	4	2	kiru	1	1	1.166	594	6	2	kokyu-	1	1	1.027
454	6	3	kiroku	1	0	0.438	595	6	2	kokyo-	1	1	1.333
463	3	2	kuu	2	1	1.052	596	6	3	kokugo	1	0	1.447
464	5	2	ku-ki	0	1	1.070	599	6	3	kogeru	1	0	1.359

Table 5 (continued)

No.	a	b	word	c	d	dist.	No.	a	b	word	c	d	dist.
477	6	3	kudaku	1	0	1.200	631	6	3	koyomi	1	0	1.432
494	4	2	kuru	1	1	1.389	632	4	2	koru	1	1	1.326
495	5	3	kurun	0	0	1.379	633	4	2	koro	1	1	1.425
500	4	2	kuro	1	1	1.433	635	8	4	korogaru	3	1	1.360
502	5	2	kuro-	0	1	1.491	637	6	3	korobu	1	0	1.215
507	5	2	ke-ka	0	1	0.778	1428	6	2	byo-ki	1	1	1.413
508	6	3	ke-kai	1	0	1.008	1497	6	3	bo-eki	1	0	1.490
397	5	3	kioku										

a: number of phonemes, b: number of syllables,
c: difference of phonemes, d: difference of syllables.

Table 6: Number of short-distance word pairs counted with (a) phonemes and (b) syllables

(a)

1	2	3	4	5	6	7	8	9	10	11	*
1	7										1
	58	170	305								2
		363	772	845	381						3
			3606	4615	4608	974	594				4
				3184	5973	1946	514	77	0		5
					6328	3905	1773	197	16	0	6
						1314	836	122	7	0	7
							510	130	20	2	8
								21	2	0	9
									0	1	10

(b)

1	125					
2	669	12654				
3	110	14080	10013			
4	2	1627	3554	963		
5	0	52	198	117	6	
6	0	0	4	3	0	0
**	1	2	3	4	5	6

* number of phonemes, ** number of syllables

other 1721 words (excluding the thirty-two 1- and 2-phoneme words and one 11-phoneme word). Entries in the table are, from left to the right, the word number, number of phonemes, number of syllables, phonemic transcription of the word, difference of the number of phonemes, difference of the number of syllables, and the distance. As seen from this table, the word nearest in the distance to /kioku/ is /kiroku/ (record). Next come /ke-ki/, /kyoka/, /kyaku/ in this order. Words in か(ka)-line appear very frequently. But words such as /byo-ki/ and /bo-eki/ are also seen.

Table 6 shows the number of short distance (less than 1.5) word pairs counted with phonemes and syllables for the 44177 word pairs in all. In Table 6(a) the entry 0 shows that none of those calculated in the range of eq (11) have a distance less than 1.5. Blank entries in (a) show that no calculation was performed due to the condition (11). The table shows that pairs with 6-phoneme words have the largest number. This is because 6-phoneme words are the most frequent. Pairs with 4-phoneme words and 5-phoneme words also have many short distance words. On the other hand, some of those with a fairly different number of phonemes (e.g. 4-phoneme words and 8-phoneme words) have quite a few short distance words.

In respect to the number of syllables in Table 6(b), pairs to 2-syllable words and 3-syllable words are frequent. This is because 83% of the total is occupied by the words of these syllables. In spite of a fairly large difference in the number of phonemes and the number of syllables, some have short distance (e.g. /iyoiyo/,/yo-/).

Table 7 shows addition and deletion of phonemes for word pairs of distance less than 0.5 (923 word pairs). They are classified into the following 8 types:

Av : deletion or addition of a vowel,
Ac : deletion or addition of a consonant,
Ae : deletion or addition of a syllabic nasal, a double consonant or a contracted sound,
B : replacement of vowels,

Table 7: Number of short-distance word pairs of eight error types with examples

*	**	examples
Av	51	/narau/-/naru/
Ac	15	/kaeru/-/kareru/
Ae	9	/kyo-cu-/-/ko-cu-/
B	282	/hiku/-/huku/
C	142	/takai/-/tagai/
D	41	/kiqto/-/kiqpu/
F	19	/kuro/-/kuro-/
B F	45	/sici/-/sire-/
B B	37	/tokeru/-/tokoro/
B Av	28	/ko-sai/-/ko-su/
AcAv	21	/oku/-/okuru/
F F	18	/ko-syu-/-/kosyo-/
C F	16	/ke-ki/-/geki/
C B	15	/sikaku/-/zikoku/
C C	14	/se-me-/-/te-ne-/
Other 2 typ.	89	/do-gu/-/doku/
B AcAv	10	/hazimaru/-/hazime/
B B B	9	/se-siN/-/soseN/
Others	62	/yaburu/-/yamu/

* error types
** number of error types
Av: deletion or addition of a vowel.
Ac: deletion or addotion of a consonant.
Ae: del. or add. of a syllabic nasal or a choked sound.
B : replacement of vowels.
C : replacement of consonants.
D : replacement of syllables.
E : rep. of syllabic nasal with choked sound.
F : del. or add. of an elongated sound.

C : replacement of consonants,
D : replacement of syllables,
E : replacement of a syllabic nasal or a double consonant with a contracted sound,
F : deletion or addition of an elongated sound.

The type E did not occur. The number of occurences of the other 7 types are shown in Table 7. Types B and C are particularly frequent. Of the 64 mixed types such as BC (BC and CB are of different caregories), 23 types are within the distance 0.5. BF (vowel replacement and elongated sound deletion), BB (duplicated vowel replavement), BAv (vowel replacement and vowel deletion) and AcAv (one-syllable deletion) are among the frequent occurences. Mixture of more than two types occur rarely (there are 35 such types in total), but some can be observed such as BAcAv (vowel replacement and one-syllable deletion) and BFB (double replacement of vowels and deletion or addition of an elongated sound).

DISCUSSION

Investigation of Japanese word distance on the continuous parameter level shows, as expected, that those which are close in the number of phonemes and syllables are found frequently among those having short distances. Particularly, the replacement of sounds within the same row or column in Table 1 (i.e. replacement of vowels or consonants) is observed frequently. This is natural from the structure of Japanese words. That is, some nouns are derived from infinitive forms of verbs as /simekiru/-/simekiri/ (close-closing), some intransitive verbs and transitive verbs have similar forms as /cumaru/-/cumeru/ (be full - fill).

Also interesting is the existence of word pairs with a fairly large difference in the number of phonemes and syllables but having a short distance. For example the pair /yabureru/ -/yamu/ (tear - stop) are doubly different in both the number

of phonemes and syllables, but its distance is within 0.4% of the lower end in the distribution of the whole distances. This seems to be related to our word perception, but needs further confirmation through hearing experiments.

The problem in this experiment is the modeling. The distinctive features of phonemes are taken into account. But these are treated as intrinsic to the phonemes. Some feature values must be changed depending on the preceding and/or following phonemes. This is a problem left to the future, as well as the choice of distinctive features and the time constant.

The word accent is not used here. However, the distinction of homonyms by accent [10] is common in Japanese. Description of word accent by continuous lines [11] should be adopted in the future.

Detailed data are obtained for the 44177 word pairs of the distance less than 1.5, and it is possible to construct a confusion matrix to indicate the similarity of words. Since, however, this requires a further analysis of the data, it remains to be investigated in the future.

ABSTRACT

Basic words selected from a Japanese dictionary are transcribed into a series of phonemes, which are then rewritten into dictinctive features with a value of 1 or 0. A series of distinctive feature vectors is used as inputs to a second-order linear system model which transforms input step functions into continuously changing parameter values. Distances between each word pair in the dictionary are calculated using dynamic programming based on continuous parameters. The distances can be used for the study of discriminability between words.

ACKNOWLEDGEMENT

The authors wish to express their gratitude to Mr. Kazuhiro Fuchi, the Chief of Machine Inference Section, Information Sciences Division, who granted support to our study.

REFERENCES

All references are written in Japanese. Some are available in English as indicated.

1 ITAHASHI, S., SUZUKI, H. and KIDO, K., Discrimination of some Consonants in Words with the Aid of Dictionary. The Transactions of the Institute of Electronics and Communication Engineers of Japan 54-C, 1 (1971), 10-17. Available in English in: Electronics and Communications in Japan, 54-C, 1 (1971), 107-114.

2 MAKINO, S. and KIDO, K., Properties of Phonemic Pairs for Distinguishing a Word Pair with a Short Distance. The Transactions of the Institute of Electronics and Communication Engineers of Japan, J62-D, 8 (1979), 507-514.

3 ITAHASHI, S. and YOKOYAMA, S., Description and Segmentation of Formant Trajectory with Second-Order Linear System Model. Bulletin of the Electrotechnical Laboratory 40, 6 (1976),530-541.

4 YOKOYAMA, S. and ITAHASHI, S., On the Distance of Words Based on Distinctive Features and the Second-Order Model. Preprint of Autumn Meeting of the Acoustical Society of Japan (Oct. 1977), Paper 3-1-12. Available in English in PIPS-Report No. 18, Electrotechnical Laboratory.

5 KINDAICHI, K., KINDAICHI, H., KENBO, H., SHIBATA, T. and YAMADA, T. (eds.), Shin Meikai Kokugo Jiten (New Concise Japanese Dictionary) 1st ed. (1972), 2nd ed. (1974) Sanseido, Tokyo.

6 YOKOYAMA, S., Preparation for the Database Management of a Japanese Dictionary. Bulletin of the Electrotechnical Laboratory 41, 11 (1977), 855-863.

7 HAYATA, T., Japanese Phonology. In: HIKI, S. (ed.), Speech Information Processing. Tokyo, Univ. Press 1973, 17-41.

8 HIKI, S., KANAMORI, Y. and OIZUMI, J., On the Duration of Phoneme in Running Speech. The Journal of the Institute of Electronics and Communication Engineers of Japan 50, 5 (1967), 849-856.

9 SAKOE, H. and CHIBA, S., Recognition of Continuously Spoken Words Based on Time-Normalization by Dynamic Programming. The Journal of the Acoustical Society of Japan 27, 9 (1971), 483-490.

10 YOKOYAMA, S. and ITAHASHI, S., On the Discrimination of Words with Different Accent. Preprint of Autumn Meeting of the Acoustical Society of Japan (Oct. 1976) Paper 2-4-18. Available in English in PIPS-Report No. 14, Electrotechnical Laboratory.

11 ITAHASHI, S., Description of Fundamental Frequency Trajectory with continuous lines. Preprint of Spring Meeting of the Acoustical Society of Japan (Apr. 1977) Paper 3-2-3. Available in English in PIPS-Report No. 18, Electrotechnical Laboratory.

Verteilungswörterbücher und Grundwortschatz

N.D. Andreev, Leningrad

Die Geschichte lehrt, daß den stabilsten Bestandteil des Sprachsystems der lexikalische Kern bildet, den wir terminologisch präziser als *Grundwortschatz der Sprache* bezeichnen. In der Tat weisen die Mengen der Flexionsmorpheme in der Grammatik des Hindi einerseits und in der Grammatik des modernen Englisch andererseits kein solches gemeinsames Element auf, von dem man sagen könnte, daß in ihm die Abbildung eines und desselben indogermanischen Archetypus enthalten sei. Ganz anders steht es um den Wortschatz. Es genügt, nur ein Beispiel anzuführen: Dem russischen нос (aus dem gemeinslavischen **nosŭ*) entsprechen das englische *nose* (das auf das urgermanische **nasu* zurückgeht) und im Hindi das Wort *nāk* (von der altindischen Form *nāsikā*, die von *nāsā* abgeleitet ist). Es zeigt sich also, daß sich das betrachtete Element des Grundwortschatzes über die Jahrhunderte und die Jahrtausende hinweg erhalten hat, und zwar mit einer erstaunlichen Stabilität, die um ein vielfaches größer ist als diejenige, die wir bei beliebigen grammatischen Morphemen feststellen können.

Der Grundwortschatz der Sprache weist nicht nur das Merkmal der diachronen Stabilität auf, er spielt auch eine wichtige Rolle bei der Organisation der sprachlichen Synchronie, denn er ist diejenige Invariante, die für beliebige Kommunikationsbedingungen charakteristisch ist, in den verschiedensten Anwendungsbereichen der Sprache benutzt wird und wie eine Klammer alle Träger einer Sprache vereinigt, indem sie deren soziolinguistische Gemeinsamkeit konstituiert.

All diese Behauptungen sind unbestritten, Gegenstand der Auseinandersetzung ist ein anderes Moment, die Frage nämlich, wo genau die Grenzen des lexikalischen Kerns einer Sprache verlaufen, anders ausgedrückt, worin die *objektiven* Kriterien bestehen, die es erlauben, den Grundwortschatz als das Zentrum des Wortschatzes einer Sprache abzugrenzen.

Das vollständige Scheitern aller Versuche, eindeutige Unterscheidungsmerkmale des Grundwortschatzes in einem System rein qualitativer Oppositionen zu finden, hat die Sprachwissenschaft gezwungen, auf quantitative Kriterien zurückzugreifen. Leider erhielt man auch hier nicht sofort eine Antwort: Der elementarstatistische Ansatz, wie er in der Schaffung von Frequenzwörterbüchern zum Ausdruck kommt, hat nicht nur das Problem nicht gelöst, sondern nährte im Gegenteil Zweifel an seiner Lösbarkeit, da sich letzten Endes herausstellte, daß einige Wörter, die eindeutig einer eng spezialisierten Subsprache zuzurechnen sind, einen hohen statistischen Rang aufweisen (vgl. beispielsweise das Substantiv боцман *Bootsmann* in dem Frequenzwörterbuch der russischen Sprache aus dem Jahre 1977 ["Частотный словарь русского языка", Москва: Русский язык, 1977]), während andere Wörter, die unzweifelhaft dem Grundwortschatz angehören, einen sehr viel niedrigeren statistischen Rang besitzen (vgl. beispielsweise das Verb заботиться *sich sorgen, sich kümmern* in demselben Wörterbuch mit einer Frequenz, die genau dreimal niedriger ist als die des Wortes боцман).

Einen Ausweg aus diesen Schwierigkeiten bietet die Konzeption des *Verteilungswörterbuches*. Diese Konzeption beruht auf der kla-

ren Erkenntnis der Tatsache, daß die Auftretenshäufigkeit eines Wortes in Texten eine Funktion zweier Argumente ist. Erstens hängt sie ab von der Anzahl der Autoren, die dieses oder jenes Lexem gebrauchen. Dieses Argument wollen wir als *Repräsentationsmaß* eines Wortes bezeichnen. Zweitens hängt die Auftretenshäufigkeit auch davon ab, mit welcher Intensität einzelne Autoren ein gegebenes Lexem verwenden. Wir wollen dieses Argument die *individuelle Gebrauchsvariabilität* eines Wortes nennen.

Das Repräsentationsmaß eines Wortes wird bestimmt durch die Wechselwirkung zwischen den innersprachlichen Systemrelationen und den außersprachlichen Bedingungen des Funktionierens der Sprache. Die individuelle Gebrauchsvariabilität hingegen hängt vor allem mit subjektiven Faktoren zusammen und muß als Rauschen behandelt werden, das das Signal überlagert. Entsprechend wird im Verteilungswörterbuch - abweichend von den Frequenzwörterbüchern - als *Grund*charakteristik des Wortes dessen Repräsentationsmaß angesehen. Dieses wird anhand eines Korpus von Textabschnitten ermittelt, die jeweils gleich lang sind und den Werken einer hinreichend großen Zahl verschiedener Autoren entstammen.

Neben dem *allgemeinsprachlichen* Repräsentationsmaß, das die systemhaften Beziehungen innerhalb des Wortschatzes einer Sprache widerspiegelt, gibt es auch *subsprachliche* Repräsentationsmaße, die durch die außersprachlichen Bedingungen des Funktionierens der Sprache in der Gesellschaft induziert werden. Beispielsweise ist das Verb брать *nehmen* in einem Korpus von Textstichproben, die den Werken von 500 Autoren entnommen wurden und jeweils 2000 Wortformen umfaßten, mit dem allgemeinsprachlichen Repräsentationsmaß 159 vertreten, d.h., innerhalb des betrachteten Textmassivs wurde dieses Verb von fast jedem dritten

Autor gebraucht. In Techniktexten hingegen ist sein subsprachliches Repräsentationsmaß um mehr als das Doppelte kleiner als das allgemeinsprachliche: In Stichproben von Texten, in denen technische Probleme behandelt werden, verwendet im Durchschnitt nur ein Autor von sieben das Verb брать. In Dramen hingegen läßt sich das gleiche Verb bei gleichem Stichprobenumfang durchschnittlich bei jedem zweiten Autor beobachten, d.h., hier übersteigt das subsprachliche Repräsentationsmaß das allgemeinsprachliche sehr deutlich.

So gelangen wir zu dem Begriff der Gleichmäßigkeit oder der Ungleichmäßigkeit der Verteilung des Wortes in Subsprachen, der einen objektiven Meßwert des Wortes hinsichtlich seines Verhältnisses zum Grundwortschatz darstellt.

Wir wollen nun die antonymen Begriffe der Neutralität und der Wertigkeit des Wortes in bezug auf einzelne Subsprachen einführen.

Wir sagen, eine lexikalische Einheit sei *neutral* bezüglich irgendeiner Subsprache, wenn ihr subsprachliches Repräsentationsmaß um nicht mehr als das Doppelte von dem allgemeinsprachlichen abweicht, d.h., wenn es dieses um mehr als das Doppelte überschreitet bzw. unterschreitet.

Entsprechend sagen wir, eine lexikalische Einheit sei *pluswertig* bezüglich irgendeiner Subsprache, wenn ihr subsprachliches Repräsentationsmaß das allgemeinsprachliche um mehr als das Doppelte übersteigt, bzw. es sei *minuswertig*, wenn es das allgemeinsprachliche um mehr als das Doppelte unterschreitet.

Ein gutes Beispiel für die Verwirklichung aller drei Situationen gleichzeitig bietet das Lexem вывод *Schlußfolgerung*, das bezüglich der wissenschaftlichen Subsprache pluswertig ist, mi-

nuswertig in bezug auf die künstlerische Prosa und das Drama und neutral in Hinsicht auf solche Subsprachen wie die der internationalen Chronik, des Sports, der Technik und der Ökonomie.

Es gibt Wörter, die nur hinsichtlich einer Minderzahl von Subsprachen neutral sind: агрегат *Aggregat*, купаться *baden*, пища *Nahrung*, юный *jung, jugendlich*. Andererseits existieren Wörter, die in bezug auf die Mehrzahl der Subsprachen neutral sind: безусловно *unbedingt*, план *Plan*, школа *Schule*, вернуться *zurückkehren*. Schließlich gibt es auch solche Wörter, die in bezug auf jede Subsprache weder plus- noch minuswertig sind, d.h., die den *höchsten Neutralitätsgrad* aufweisen: быть *sein*, вода *Wasser*, влево *nach links*, новый *neu*.

Es muß unterstrichen werden, daß sich zwischen dem Repräsentationsmaß und dem Neutralitätsgrad keine eindeutige Abhängigkeit feststellen läßt: Das Repräsentationsmaß liegt bei dem Wort быть *sein* fast dreißigmal höher als bei dem Wort влево *nach links*, doch ist der Neutralitätsgrad bei beiden gleichermaßen maximal. Das Repräsentationsmaß des Wortes вода *Wasser* liegt ein wenig unterhalb des Maßes des Wortes вид *Aussehen, Art*, doch der Neutralitätsgrad des ersteren liegt beträchtlich höher, da das Lexem вид in bezug auf zwei Subsprachen minuswertig ist.

Es folgt daraus, daß gerade ein hoher Neutralitätsgrad und nicht etwa ein hoher Wert des Repräsentationsmaßes das *Hauptkriterium* für die Zugehörigkeit eines Wortes zum Grundwortschatz ist.

Wenn im Rahmen eines Korpus von Textstichproben ein Wort in keiner einzigen Subsprache vertreten ist, dann haben wir es mit dem Extremfall der Minuswertigkeit zu tun, der als paradigmatische Null im Subsprachenraum anzusehen ist. Das Vorkommen oder das Fehlen solcher paradigmatischer Nullen verengt oder erweitert das Ausmaß der Verbreitung des Lexems. Je weiter dieses Ausmaß ist, mit um so größerer Berechtigung kann man das Wort dem Grundwortschatz zurechnen. Die hier in Rede stehende Charakteristik ist jedoch lediglich eine Ergänzung des Hauptkriteriums, d.h. des Neutralitätsgrades des Wortes. Am interessantesten sind die *systemhaften* Beziehungen zwischen den subsprachlichen Pluswertigkeiten. Jede dieser Wertigkeiten verfügt über eine auszeichnende Stärke im mehrdimensionalen Subsprachenraum. Wir werden alle Wörter, die bezüglich einer und derselben Subsprache pluswertig sind, als *subsprachliche Phratrie* betrachten. Eine subsprachliche Phratrie stellt ein Analogon des semantischen Feldes dar, unterscheidet sich indes von diesem in dreifacher Hinsicht: Erstens wird sie *objektiv* ausgegrenzt, unabhängig von den theoretischen Anschauungen des Linguisten. Zweitens geschieht ihre Ausgrenzung nicht auf begrifflich-logischer, sondern auf *soziofunktionaler* Grundlage. Drittens geschieht die Formierung subsprachlicher Phratrien aufgrund eines *quantitativ-qualitativen* Kriteriums, im Unterschied zu den qualitativen Klassifikationen in den früheren Theorien der semantischen Felder.

Bei einem solchen Vorgehen zerfallen die Wörter, die subsprachliche Wertigkeiten aufweisen, in zwei Gruppen, die Gruppe der *einwertigen* Wörter, die lediglich je einer subsprachlichen Phratrie angehören, bzw. die Gruppe der *mehrwertigen* Wörter, die jeweils gleichzeitig zu zwei oder zu drei subsprachlichen Phratrien gehören. Mehrwertige Wörter, die den Phratrien zweier verschiedener Subsprachen gemeinsam sind, bilden einen solchen Teil dieser Phratrien, der relativ klein ist, jedoch außerordentlich wichtig für das Verständnis der semantischen Relationen zwischen den beiden jeweils betrachteten Subsprachen.

Wir werden sagen, zwei Subsprachen seien *kowertig* in Hinsicht auf die Lexeme, die in den gemeinsamen Teil ihrer Phratrien gehören. Wir werden die *Stärke der Kowertigkeit* zweier Subsprachen in bezug aufeinander durch die Menge der mehrwertigen Wörter bestimmen, die in diesen gemeinsamen Teil eingehen. Verschiedene Paare von Subsprachen weisen eine *verschiedene* Stärke der sie miteinander verbindenden Kowertigkeit auf. Der gegebene Faktor gibt eine objektive Grundlage ab für den Aufbau einer funktional-distributionellen Semantiktheorie als eines gesonderten Zweiges der struktural-probabilistischen Sprachanalyse.

A Statistical Approach to Some Aspects of Style in Six Old English Poems: A Computer-Assisted Study

Sandra J. Harmatiuk, Notre Dame

This study is based on two fundamental assumptions. The first is that underlying the style of any literary work are basic linguistic patterns which help to discriminate between one writer and another. Secondly, these patterns are measurable and verifiable empirically through the sciences of linguistics and statistics. The specific poems analyzed are *The Fates of the Apostles, Christ II, Juliana, Elene, Seafarer,* and "Deor's Lament". The last two poems were included as a control group for purposes of statistical comparison. The text for these poems follows that of *The Vercelli Book* and *The Exeter Book,* Volumes II and III of *The Anglo-Saxon Poetic Records* edited by KRAPP and VAN KIRK DOBBIE. The linguistic data were compiled using a computer program written especially for this purpose.

The question raised is whether the proportionate distribution of selected linguistic items is really different, that is, do they differ more than one would expect according to the ordinary fluctuations of chance. One statistical procedure for testing such fluctuations is the Chi-square (χ^2) test which measures the discrepancy that exists between the observed and the expected frequencies. The formula used to calculate the value called Chi-square is:

$$X^2 = \Sigma \frac{(o-e)^2}{e}$$

In this formula, o = the observed frequencies, e = the expected frequencies, and Σ indicates that one should take the sum of the individual calculations. In order to calculate these values, it is necessary to set up a contingency table consisting of rows (r) and columns (k) containing the raw data. The expected frequency for any given cell in this table is computed by the formula

$$\frac{\text{row total} \times \text{column total}}{N}$$

where N = the total number of occurrences. A sample contingency table would appear as follows:

	Col. 1	Col. 2	
Row 1	A	C	A + C
Row 2	B	D	B + D
	A + B	C + D	N

N = A + B + C + D

Since one function of statistics is to test the truth or falsity of a given theory, a few words about the statement of the test hypothesis must be made. In the simplest terms, the experimenter begins with a theory which he wishes to test. This theory is the research hypothesis, usually designated by the symbol H_1. In order to test this theory, however, another hypothesis is formulated called the null hypothesis (H_o) or the hypothesis of no differences. In the Chi-square test, which is a test of association, the null hypothesis takes the general form of stating that the distributions being examined are not influenced by membership in a particular group and are, therefore, independent. Thus, for example, an examiner

might analyze the possible influence of sex on the choice of reading matter in a given population. A significant difference in such a comparison would mean that a definite relationship exists between sex and choice of reading matter. In the case of this study, distributions of form-class (that is, distribution of content words and function words) and certain syntactic groups were analyzed for possible influence of membership in a particular group (that is, appearance in a given poem) on the distribution considered. The null hypothesis states that the distributions are independent, that is, they could appear with equal likelihood in any of the poems. The research hypothesis, on the other hand, posits a relationship of interaction between the distribution examined and the group to which that distribution belongs. In order to reject the null hypothesis in favor of the research hypothesis, it is necessary that the calculated value of Chi-square be great enough to result in a reading at a predetermined significance level. The conventional level of significance for Chi-square is any value which renders the probability of an error occurring as equal to or less than five out of one hundred ($p \leq .05$). To determine whether the calculated value is significant, it is necessary to consult a standard Chi-square table. To find the correct value of the obtained statistic, it is necessary to compute the number of degrees of freedom for the test run. The formula for this is df = (r-1)(k-1) where r = the number of rows in the contingency table and k = the number of columns. For a 2 x 2 table, this becomes (2-1)(2-1) = (1)(1) = 1. For a 2 x 3 table, this becomes (2-1)(3-1) = (1)(2) = 2.

One additional statistical test was used for this study, the analysis of variance test, which was used to analyze the fluctuations in line length and sentence length for the six poems. The null hypothesis for this test states that the means of the different groups are equal, while the alternative hypothesis states that they are not equal. As with the Chi-square test, the level of significance, that is, that value of the statistic which allows rejection of the null hypothesis, is

any value ≤ .05.

To calculate the analysis of variance statistic, it is necessary to compute the individual means for each group, the deviation of members within the group from the mean, the combined means of all the groups, and the deviation of each member from the combined mean. The computation of the group mean and the deviations around that mean lead to the calculation of the within-groups variance, while the computation of the combined mean, usually called the grand mean, and the deviations around it lead to the between-groups variance. Each deviation around the group mean is squared and the sums of these squares are placed in the appropriate table. The same procedure follows for the deviation around the grand mean. The next step in the calculation is to divide each of these values by the appropriate degrees of freedom. The formula for the degrees of freedom is:

$df_{total} = n - 1$

$df_{bg} = k - 1$

$df_{wg} = r - 1.$

To compute the F statistic or variance ratio, the larger mean square is divided by the smaller:

$$F = \frac{\text{mean square}_{bg}}{\text{mean square}_{wg}}$$

To determine the level of significance, it is necessary to consult a standard F table.

The following linguistic distributions were considered in this study:

1. Overall occurrence of content words / function words in the poems (see Table 1)
2. Occurrence of content words / function words at the beginning of sentences (see Table 2)
3. Occurrence of content words / function words at the end of sentences (see Table 3)

4. Distribution of sentences according to number of clauses (see Table 4)
5. Distribution of sentences with and without subordination (see Table 5)
6. Distribution of independent / dependent clauses (see Table 6)
7. Distribution of relative / subordinate clauses (see Table 7)
8. Distribution of sentence connectivity (see Table 8)
9. Distribution of the connective *ða* (see Table 9)
10. Distribution of adverbs as head of subordinate clauses (see Table 10)
11. Distribution of determiners as dependent clause head (see Table 11)
12. Line length variance (see Tables 12, 13)
13. Sentence length variance (see Tables 14, 15)

The comparison of the results of tests on Items 1-3 above yielded the following results. In overall distribution of content words and function words for the six poems the calculated value of X^2 (15.78) exceeds the tabled value of χ^2 at the .05 level (see Table 1); therefore, the null hypothesis can be rejected in favor of the research or alternative hypothesis. The two poems that fall farthest outside the theoretical frequencies are *Fates of the Apostles* (Poem 1) and *Juliana* (Poem 3). The null hypothesis is also rejected in the third comparison (see Table 3). In this case, the poem falling farthest outside the theoretical frequencies is "Deor" (Poem 5). There is no significant difference in the distribution of function words and content words at the beginning of sentences (see Table 2).

Before continuing with a discussion of the statistical tests on syntactic units, it would appear worthwhile to discuss the determination of these units. The level of the syntactic unit analyzed at the outset of this discussion is

that of the sentence. It is here that the computer proved useful. A morphologically coded text for each poem was stored on tape and scanned for end-stop punctuation. A computer printout of the text is obtained which presents the text vertically rather than horizontally. One word is printed on a line while blank spaces indicate the beginning and end of a unit. In order to understand how the syntactic program operates, let us examine the first sentence of *Juliana* which comprises the first seven and one-half lines of the poem:

> Hwæt! We ðæt hyrdon hæleð eahtian,
> deman dædhwate, þætte in dagum gelamp
> Maximianes, se geond middangeard,
> arleas cyning, eahtnysse ahof
> cwealde cristne men, circan fylde,
> geat on græswong godhergendra,
> hæþen hildfruma, haligra blod,
> ryhtfremmendra. (ll. 1-8a)

The coded text for this same line appears as follows:

> FIOO! FSOO FDOO VTON VTAN,
> VTAN AOOE, FDOO FPOO NOUM VIOO
> NOES, FDOO FPOO NOOO,
> VTDE AONE NOOO, NOAN VTDE,
> VTOO FPOO NOOO NORA,
> AOOO NOOA, AORA NOOO,
> VTRA.

Once sentence boundaries are established on the basis of end-stop pnctuation (commas and semi-colons are ignored since there is some inconsistency in their use by the editors of the text), it is possible to analyze the structure for clausal relationship. The text may be diagrammed as the sample from *Juliana* (ll. 1-8a) demonstrates (see Figure 1):

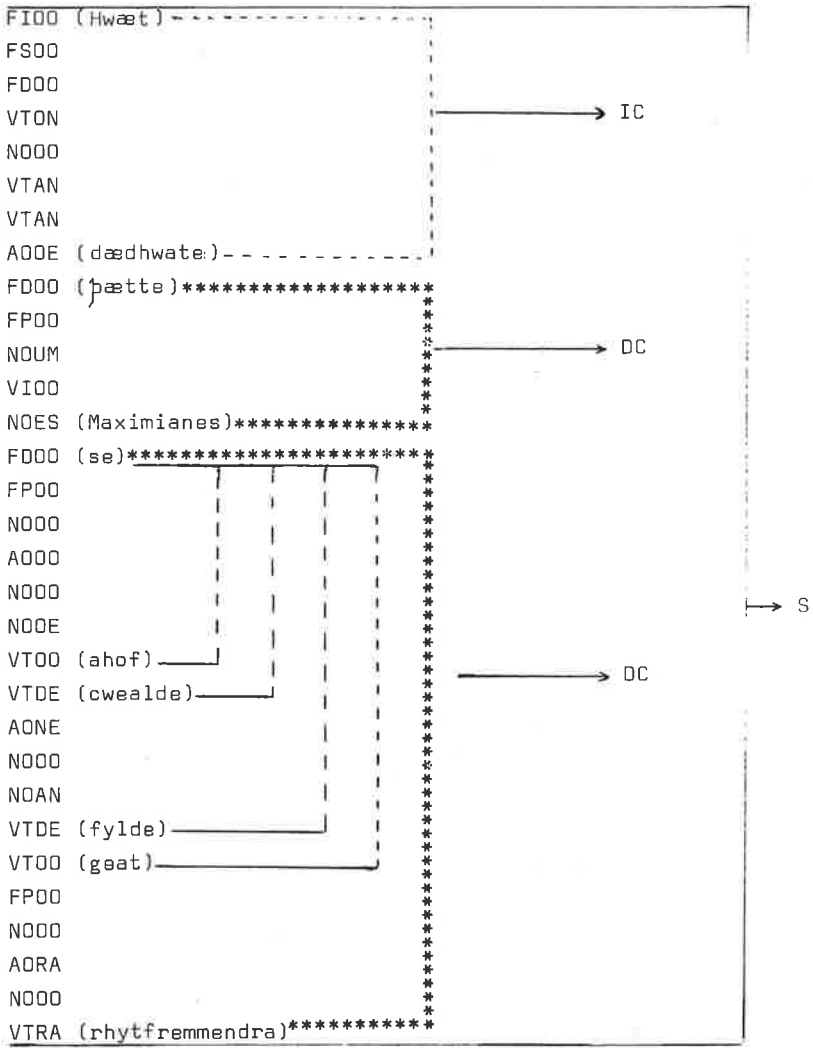


Figure 1: Clausal Relationship of the first sentence of *Juliana*

In diagramming the sentence the solid line designates the sentence unit. A broken line indicates the beginning and end of an independent clause (IC), while the line of asterisks sets off dependent clauses (DC). In Figure 1, it can be seen that the sentence is made up of one independent and two dependent clauses. The second dependent clause contains a series of four verbs which are logical predicates of the determiner which opens the clause:

se geond middangeard,
arleas cyning, eahtnysse *ahof*,
cwealde cristne men, circan *fylde*,
geat on græswong godhergendra,
hæþen hildfruma, haligra blod,
ryhtfremmendra.
(*who* throughout the earth,
/the/ cruel king, *waged* persecution,
slew Christian men, *destroyed* /the/ Church,
shed the holy blood of the worshippers of God,
/the/ heathen ruler.)

The determiner *se* in this clause functions as a relative pronoun referring to the noun *Maximianes* which immediately precedes it and further modified by the appositive phrases *arleas cyning* and *hæþen hildfruma*. In the diagram, the connection of these verbs to *se* is indicated by a series of lines joining each verb to the determiner.

After diagramming the sentences in the six poems in this manner, clauses were classified as either independent or dependent. Independent clauses are those clauses beginning with any of the following: noun, verb, adjective, or sentence connectives such as *forþon, þus*, etc. and functioning as the main clause of the sentence. Dependent clauses were classified as either relative (Type I) or subordinate (Type II) clauses.

After classifying clauses as independent or dependent, it became possible to determine the distribution of sentence length on the basis of the number of independent clauses contained. Three categories were established:

Class 1: one independent clause ± dependent clauses
Class 2: two independent clauses ± dependent clauses
Class 3: more than two independent clauses ± dependent clauses.

Since the use of subordination may be an indication of stylistic preference, the distribution of sentences with and without subordination was also analyzed.

In comparing sentences based on the number of independent clauses, the calculated value of X^2 (93.1) was great enough to reject the null hypothesis at the .001 level (see Table 4). In comparing sentences with and without subordination it was not possible to reject the null hypothesis (see Table 5).

The next series of comparisons performed dealt with the distribution of independent and dependent clauses and the distribution of relative and subordinate clauses. In neither case was the calculated value of X^2 great enough to result in rejection of the null hypothesis (see Tables 6,7).

The text features considered in the next series of comparisons are the use (or absence in independent clauses) of syntactic markers for the relationship of sentences and clauses. Sentence connectivity may be indicated by simple juxtaposition (parataxis) or on the basis of the use of a syntactic marker indicating relationship (hypotaxis). Generally speaking, connectives were found to fall into three categories: 1) sentence connectives such as *forþon, þus, swylce*; 2) subordinating conjunctions such as *gif, þeah*; and 3) adverbs such as *swa, þonne* and determiners such as *da, se* etc. Of the six poems studied, "Deor" uses only the third type, while *Fates of the Apostles* and *Seafarer* use types one and three only. The remaining poems make use of all three types.

In considering sentences for the distribution of parataxis / hypotaxis it was not possible to reject the null hypothesis (see Table 8). For those sentences making use of hypotaxis, the connective *da* would appear to be of some importance. In comparing the use of *da* with other sentence connectives (type 1, above) the calculated value of X^2 was great enough to reject the null hypothesis at the .001 level (see Table 9).

A word of caution should perhaps be interjected at this juncture. In the use of the connective *da* and in the distribution of the sentences according to the number of independent clauses, the null hypothesis was rejected at the .001 level of significance. In the former case, the fact that the connective *da* is not used in Poems 1, 5 and 6 may simply be the result of the relative shortness of the poems themselves. In the latter case, the number of clauses indicated was arrived at following the editorial practice of the texts used and may not necessarily reflect the intentions of the original poets.

The use of syntactic markers for dependent clauses is much more clearly evident for these poems. Relative clauses may begin with an adverb, a determiner, or a pronoun. Subordinate clauses may begin with an adverb, a conjunction, or a determiner. There is a significant difference in choice of markers for both relative and subordinate clauses in the six poems (see Tables 10,11).

The final items analyzed were the distribution of line length and sentence length for the six poems. As was stated earlier, the statistical procedure used for these comparisons is the analysis of variance test. The calculated F-ratio allows for rejection of the null hypothesis at the .05 level (see Tables 12, 13 and Tables 14, 15).

Significant differences among the six poems are evident, then, in eight of the thirteen comparisons made for this study. The data may be summarized as follows:

DISTRIBUTION ANALYZED	CALCULATED STATISTIC	RESULT
Overall distribution of Content Words / Function Words	$X^2 = 15.78$, $p<.01$	Reject H_o
Distribution of CW/FW at the beginning of sentences	$X^2 = 9.71$, $p>.05$	Retain H_o
Distribution of CW/FW at end of sentences	$X^2 = 51.6$, $p<.001$	Reject H_o
Distribution of Sentence Type: Class 1, 2, and 3	$X^2 = 93.1$, $p<.001$	Reject H_o
Distribution of Sentences with and without subordination	$X^2 = 7.02$, $p>.05$	Retain H_o
Distribution of IC/DC	$X^2 = 8.22$, $p>.05$	Retain H_o
Distribution of Relative / Subordinate Clauses	$X^2 = 8.99$, $p>.05$	Retain H_o
Sentence connectives	$X^2 = 8.42$, $p>.05$	Retain H_o
Use of *da* as a connective	$X^2 = 37.68$, $p<.001$	Reject H_o
Distribution of relative clause markers	$X^2 = 56.92$, $p<.001$	Reject H_o
Distribution of subordinate clause markers	$X^2 = 20.34$, $p<.05$	Reject H_o
Distribution of line length	$X^2 = 7.496$, $p<.05$	Reject H_o
Distribution of sentence length	$X^2 = 2.896$, $p<.05$	Reject H_o

Where significant differences occur, the control poems ("Deor" and ***Seafarer***) most probably account for these differences in three of the eight instances (items 3, 4, and 10). When the contributing factor appears to be one or more of the Cynewulfian poems, ***Juliana*** tends most often to fall outside the expected frequencies (items 1, 9, and 11).

In conclusion, although it must be stated that the data analyzed here are in no way exhaustive, they do , indeed cover a number of distinctive linguistic vectors for the poems analyzed. It should be noted that in the area of syntax only very broad categories were treated. Using the data amassed for this initial study, however, a more sensitive measuring device may be possible and is presently, in fact, being developed. It is to be hoped that the further information such refinements will provide will prove a reliable basis for further stylistic analysis of the Old English poetic corpus.

SUMMARY

This study attempts to measure some linguistic features of six Old English poems in an attempt to arrive at an initial statement regarding the authorship of the Cynewulfian poems. The linguistic features considered are: form-class distribution, distribution of sentence and clause types, the use or absence of syntactic markers, and the distribution of line and sentence length.

This paper summarizes the results of a series of statistical comparisons made on the existing data. On the basis of the results of these tests, the following statement would appear valid. Although not, at this point, conclusive, there is a strong evidence supporting a consideration of an authorial difference between ***Juliana*** and the other Cynewulfian poems. Further investigation with more sensitive measuring categories would appear justified.

Table 1: Overall Distribution of Content Words/Function Words

	Poem 1	Poem 2	Poem 3	Poem 4	Poem 5	Poem 6	
CW	457 (425.5)	1584 (1563.5)	2545 (2631.3)	4700 (4666.2)	145 (142.9)	487 (488.6)	9918
FW	210 (241.5)	867 (887.5)	1580 (1493.7)	2615 (2648.8)	79 (81.1)	279 (277.4)	5630
	667	2451	4125	7315	224	766	15548

X^2 = 15.78 df = 5 p<.01

Table 2: Distribution of CW/FW; Sentence Begin

	Poem 1	Poem 2	Poem 3	Poem 4	Poem 5	Poem 6	
CW	25 (19.4)	65 (61.6)	94 (107.8)	189 (185.6)	6 (8.3)	19 (15.3)	398
FW	17 (22.6)	68 (71.4)	139 (125.2)	212 (215.4)	12 (9.7)	14 (17.7)	462
	42	133	233	401	18	33	860

X^2 = 9.61 df = 5 p>.05

Table 3: Distribution of CW/FW, Sentence End

	Poem 1	Poem 2	Poem 3	Poem 4	Poem 5	Poem 6	
CW	42 (39.5)	127 (125.0)	217 (218.9)	382 (376.8)	10 (16.9)	30 (31.0)	808
FW	0 (2.5)	6 (8.0)	16 (14.1)	19 (24.2)	8 (1.1)	3 (2.0)	52
	42	133	233	401	18	33	860

X^2 = 51.6 df = 5 p<.001

Table 4: Distribution of Sentences According to Number of Independent Clauses (Type 1, 2, 3)

	Poem 1	Poem 2	Poem 3	Poem 4	Poem 5	Poem 6	
1	35 (36.0)	117 (114.0)	212 (199.7)	351 (343.7)	7 (15.4)	15 (28.3)	737
2	6 (4.6)	12 (14.5)	17 (25.4)	39 (43.8)	9 (2.0)	11 (3.6)	94
3	1 (1.4)	4 (4.5)	4 (7.9)	11 (13.5)	2 (0.6)	7 (1.1)	29
	42	133	233	401	18	33	860

X^2 = 93.1 df = 10 p<.001

Table 5: Distribution of Sentences with Subordination (WS) and without Subordination (WO)

	Poem 1	Poem 2	Poem 3	Poem 4	Poem 5	Poem 6	
WS	23 (21.5)	70 (68.2)	124 (119.5)	192 (205.6)	9 (9.2)	23 (16.9)	441
WO	19 (20.5)	63 (64.8)	109 (113.5)	209 (195.4)	9 (8.8)	10 (16.1)	419
	42	133	233	401	18	33	860

X^2 = 7.02 df = 5 p>.05

Table 6: Distribution of Independent / Dependent Clauses

	Poem 1	Poem 2	Poem 3	Poem 4	Poem 5	Poem 6	
IC	53 (48.8)	162 (163.1)	266 (283.9)	462 (459.4)	32 (26.8)	70 (63.1)	1045
DC	29 (33.2)	112 (110.9)	211 (193.1)	310 (312.6)	13 (18.2)	36 (42.9)	711
	82	274	477	772	45	106	1756

X^2 = 8.22 df = 5 p>.05

Table 7: Distribution of Relative and Subordinate Clauses

	Poem 1	Poem 2	Poem 3	Poem 4	Poem 5	Poem 6	
RC	16 (14.8)	56 (57.3)	91 (108.0)	174 (158.7)	7 (6.7)	20 (18.4)	364
SC	13 (14.2)	56 (54.7)	120 (103.0)	136 (151.3)	6 (6.3)	16 (17.6)	347
	29	112	211	310	13	36	711

X^2 = 8.99 df = 5 p>.05

Table 8: Distribution of Sentences with Syntactic Markers (WM) and Without Syntactic Markers (WOM)

	Poem 1	Poem 2	Poem 3	Poem 4	Poem 5	Poem 6	
WM	11 (10.1)	31 (32.0)	69 (56.1)	81 (96.5)	6 (4.3)	9 (7.9)	207
WOM	31 (31.9)	102 (101.0)	164 (176.9)	320 (304.5)	12 (13.7)	24 (25.1)	653
	42	133	233	401	18	33	860

X^2 = 8.42 df = 5 p>.05

Table 9: Distribution of the Connective *da*

	Poem 1	Poem 2	Poem 3	Poem 4	Poem 5	Poem 6	
da	0 (4.5)	6 (13.5)	32 (26.4)	46 (33.5)	0 (2.7)	0 (3.6)	84
Other	10 (5.5)	24 (16.6)	27 (32.6)	29 (41.5)	6 (3.3)	8 (4.4)	104
	10	30	59	75	6	8	188

X^2 = 37.68 df = 5 p<.001

Table 10: Relative Clause: Word Begin

	Poem 1	Poem 2	Poem 3	Poem 4	Poem 5	Poem 6	
Adv.	9 (1.9)	6 (6.6)	5 (10.5)	15 (20.6)	0 (0.8)	8 (2.4)	43
Deter.	5 (12.6)	47 (44.0)	80 (71.5)	137 (136.7)	7 (5.5)	10 (15.7)	286
Other	2 (1.5)	3 (5.4)	6 (8.8)	22 (16.7)	0 (0.7)	2 (1.9)	35
	16	56	91	174	7	20	364

X^2 = 56.92 df = 10 p<.001

Table 11: Subordinate Clause: Word Begin

	Poem 1	Poem 2	Poem 3	Poem 4	Poem 5	Poem 6	
Adv.	5 (6.5)	30 (26.6)	43 (57.1)	76 (64.7)	4 (2.9)	7 (7.6)	165
Conj.	2 (2.5)	6 (10.7)	25 (22.8)	30 (25.9)	0 (1.1)	3 (3.0)	66
Deter.	6 (4.3)	20 (18.7)	52 (40.1)	30 (45.5)	2 (2.0)	6 (5.4)	116
	13	56	120	136	6	16	347

$X^2 = 20.34$ df = 10, $p<.05$

Table 12: Line Length Data

POEM	No. of Lines	No. of Words	Mean	Mode	S.D.
Poem 1	122	667	5.46	5	± 1.30
Poem 2	427	2451	5.74	6	± 1.13
Poem 3	731	4129	5.65	6	± 1.20
Poem 4	1321	7320	5.54	6	± 1.29
Poem 5	42	224	5.33	5	± 1.03
Poem 6	124	768	6.19	6	± 1.90

Table 13: Analysis of Variance, Line Length

SOURCE	df	SUM OF SQUARES	MEAN SQUARES
Between	5	63.6250	12.7250
Within	2761	4687.8125	1.6976
Total	2766	4750.8125	

$\hat{F} = 7.496$ $p<.05$

Table 14: Sentence Length Data

POEM	No. of Sent.	No. of Words	Mean	Mode	S.D.
Poem 1	42	666	15.86	10	± 17.38
Poem 2	133	2451	18.42	20	± 20.99
Poem 3	232	4128	17.78	13	± 21.07
Poem 4	400	7316	18.29	11	± 21.77
Poem 5	18	224	12.44	5	± 15.34
Poem 6	32	767	23.94	13	± 26.49

Table 15: Analysis of Variance, Sentence Length

SOURCE	df	SUM OF SQUARES	MEAN SQUARES
Between	5	1938.8125	387.7625
Within	851	113939.8125	133.8895
Total	856	115878.6250	

$\hat{F} = 2.896$ $p<.05$

Aberrant Frequency Words: Their Identification and Uses*

Alastair McKinnon, Montreal, McGill University

An aberrant frequency or "abfreq" word is one whose frequency in some specified subset of a corpus represents a statistically significant departure from its frequency in the whole corpus or some other appropriate comparison text. In this paper we describe a method for producing lists of such words for subsets of a corpus and indicate some of their possible uses.

Most of our work in connection with this problem has been done on Kierkegaard's *Samlede Værker*, a corpus of 54,259 types and 1,942,032 tokens in 34 separate works ranging in size from 3,034 to 214,013 tokens. However, our method is not specifically tied to this corpus and we mention it only to provide the context for some of our remarks and to permit the reader to check our results should he wish to do so[1].

The concept of aberrant frequency can be defined more precisely. Briefly, we say that a word shows aberrant frequency in a particular book (or subset of a corpus) if its observed absolute frequency in that book is statistically significantly different, at some level α, from its expected frequency. The null hypothesis assumes that the word is randomly distributed throughout the corpus, and the expected frequency is hence equal to the

* This research was done under a grant from the Canada Council (now the Social Sciences and Humanities Research Council of Canada) to which I am most grateful. (A.McK.)

proportion of the total corpus represented by the book in question multiplied by the number of tokens of the word in the whole corpus. One way of operationalizing this definition is to make use of the fact that under the null hypothesis, the observed frequency of a word in a book will be approximately normally distributed with mean and standard deviation determined by book and corpus size and overall frequency of the word. Then, if the actual observed frequency lies more *than 1.96 standard deviations* (i.e. α = .05), say, from its expected value, we may characterize the word as abfreq for that book.

This conception can be illustrated in terms of several of the words shown in Table 1. Assuming an even spread, *Abraham* should occur 5.38 times in *Fear and Trembling* and its actual 231 occurrences represent 97.26 standard deviations from this expected frequency. Similarly, *Isaak* should occur 2.46 times in this work and its 111 occurrences represent 69.212 standard deviations. So, too, for the rest of the words in this abbreviated list and indeed the longer one of which it is but a small part.

This concept can also be illustrated in terms of a single word as in Figure 1. For example, since *Du* occurs 8.716 times in the corpus its expected frequency is 592 in EE2 (*Either/Or*, vol. 2), 960 in AE *(Concluding Unscientific Postscript)*, and 20 in FV *(On My Work as an Author)*. In fact, its actual frequencies in these three works are 1,931,139 and 0. All these frequencies are significant departures from the expected ones and represent 54.749, -26.507, and -4.527 standard deviations, respectively. Each of these three values therefore falls outside the dotted lines representing +1.96 and -1.96 standard deviations in this figure. In fact, the first of these works is characterized by a

significant excess of this pronoun and the last two by a significant deficiency; indeed, this deficiency (and all that it implies) is an important characteristic which these two works have in common. Incidentally, this Figure shows all those works in which *Du* is abberrant and the extent and way in which it is so.

Our original method[2] for identifying aberrant frequency words excluded from consideration all words having corpus frequency < 5 as well as all words showing frequency 0 in the work in question. These two decisions were due in part to doubts about the distribution of such relatively rare words but it is clear that both precluded the identification of certain aberrant words and were therefore unfortunate. More importantly, this method assumed that randomly chosen samples of all words treated by it would show a normal frequency distribution, though the normal approximation is only valid for words with relatively high expected frequency in a book. As a consequence, it almost certainly identified as aberrant some words which were not actually so and failed to identify as aberrant some which actually were. Finally, as Table 3 shows, the proportion of words identified as aberrant was much larger in the case of our smallest texts than in that of our largest ones and our results therefore seemed overly dependent upon the size of the work in question. All these appeared as objections to this method as originally implemented and motivated the refinements which follow.

The most obvious difference in our present method is that it distinguishes between those cases in which the normal approximation clearly holds, i.e. in which the expected frequency of the word in a particular book is > 20, say, and those in which

it is less justified, i.e., the expected frequency ≤ 20. In the former cases the frequency distribution of the word is assumed normal with:

$$\hat{\lambda} = \frac{w_c \cdot n_i}{N}$$

$$s = \sqrt{\hat{\lambda}}$$

and standardized score for the word in the book

$$z = \frac{w_i - \hat{\lambda}}{\sqrt{\hat{\lambda}}}$$

where

N = number of tokens in the corpus
n_i = number of tokens in text i
w_c = frequency of the given word in the corpus
w_i = frequency of the given word in text i

Again, such words were identified as aberrant if the standardized score was > 1.96 or <-1.96, i.e., if the observed frequency did not lie within 1.96 standard deviations of the mean or, more precisely, of the expected frequency for the work in question[3].

Under the null hypothesis, if the corpus is large enough, the distribution of the observed frequency may be considered to be a Poisson one with mean parameter $\frac{w_c \cdot n_i}{N}$. Words having a value ≤ 20 for this mean in a particular book were not assumed to satisfy the normal approximation[4] to the Poisson and were treated according to a more exact method. They were discarded as non-

aberrant unless they met the conditions laid down in Table 4. This table was constructed from a table of Poisson probabilities in such a way as to assure that the probability of rejection of the null hypothesis is about 5% evenly distributed between the two tails of the distribution.

Note that the conditions tabulated here are all more or less equivalent to the 5% significance level specified for words with mean > 20. Of course, it is possible for such words to differ even more significantly from their expected value. As already indicated, though *Abraham* has an expected frequency of 5.38 in *Fear and Trembling*, its observed frequency represents 97.26 standard deviations from this value and its significance level is also very much higher. Hence, a standardized score for words meeting these conditions was calculated according to the formulae cited above, though its interpretations is less clear cut for the asymmetric Poisson than for the normal distribution which is symmetric.

It is of interest that the number of cases having an expected frequency > 20 constitutes only a very small proportion of the total. We have not actually counted such cases, but it is approximately 34 times the number of words in the corpus having sufficient occurrences to yield an expected frequency > 20 in an average size work. Since that requisite frequency is 680, and since there are only 303 such words in the corpus, the number of such cases must be approximately 34 x 303 or 10,302. Since this study dealt with 34 x 54,259 or 1,844,806 cases, it is obvious that the vast majority were treated by the more exact option in the present method, i.e., according to the Poisson distribution.

The choice of 20 as a lower threshold for the suitability of

the normal distribution is admittedly arbitrary. Examination of the tabulated conditions, however, shows that near 20 the use of the exact values coincided for all practical purposes with the use of the normal approximation.

There are a number of reasons for accepting this present method as a satisfactory way of identifying aberrant frequency words. By considering words having corpus frequency < 5 it is rightly able to identify as aberrant a word occurring, say, 4 times but always in a particular work. Similarly, by treating words showing frequency 0 in a particular work it rightly identifies as aberrant usually common words which are in fact totally absent from the work in question. It has deleted certain doubtful words identified by the earlier method and included other revealing ones not selected by it[5]. It makes the reasonable assumption that in certain cases the relevant distribution is approximately normal but that in the vast majority of cases it is necessary to have recourse to the more exact Poisson. All the words which it identifies as aberrant are significant at at least approximately the 5% level whereas the majority are significant at a level so high that from a statistical point of view their status must be regarded as virtually unquestionable. Finally, as Table 3 shows, while it did not increase the proportion of word types identified as aberrant in the largest works, it substantially reduced the proportion thus identified in the smallest works and to that extent achieves greater independence of work size. We therefore recommend this method as a simple and satisfactory way of identifying words which show an aberrant frequency in a particular work and which may therefore be taken as characterizing that work and, equally, as distinguishing it from others in the same corpus.

This method can be used to identify all abfreq words in any actual work or artificially constructed subset of a corpus. Depending upon their intended use these words can be listed alphabetically or in terms of their number of standard deviations, preferably in descending order. Such lists have a variety of uses.

One of the most obvious and helpful of these uses is to provide an objective and accurate abstract or summary of the subject matter of a particular work. This is clear from Table 1 which lists all words showing ≥ 20.00 standard deviations in *Fear and Trembling*. In fact, it would be impossible to recount this work without constant recourse to all these words and indeed to many others much further down this ordered list. Naturally, those words showing the highest number of standard deviations are generally most obviously revealing and important but, for reasons already indicated, the vast majority of words in the entire list may and must be regarded as expressing some characteristic of this work. Indeed, even the fact that certain words are aberrantly "low" is revealing and important. Thus, for example, the relative scarcity of *Du*, *Dig*, *Din*, and *Dit* reminds us that, despite its passion, the tone of this work is, in an important sense, much less intimate than that of, say, *Either/Or*, vol. 2 or *Christian Discourses*.

This method can also be used to produce a reliable and informative abstract of a particular sub- or mini-text dealing, for example, with some figure or concept. This can be done by extracting all sentences containing the key or search word in question, determining the frequency of each word type in this new mini-text, and comparing this with its expected frequency according to the formulae described above. Thus one can readily discover those words in Kierkegaard's writings which are particularly associated with, for example, figures such as Hegel and Socrates and concepts such as the System or, as shown in Table 2, philosophy[6].

It is perhaps worth noting that both these approaches can be taken one step further. As already indicated, an aberrant frequency list shows those words which are most central in a particular work or most closely tied with the search word of a mini-text. In fact, it is possible to discover the way in which many of these words are associated *with one another* by determining their corrected cooccurrence scores and inputting these into some computer program to produce a two dimensional map or three dimensional model giving a spatial representation of the relations of these terms to each other within the text in question[7]. Thus one can identify the principal associations or ties between the various abfreq words of a particular work or mini-text. Of course, these further procedures can be used with a set of subjectively chosen terms but, given our method, it seems more reasonable to begin with those words showing the largest number of positive standard deviations.

One can also use the various aberrant frequency lists and the same sort of program to determine and represent the overall relations between the various works of a corpus[8]. Briefly, this involves comparing the aberrant frequency list for each pair of titles, computing a similarity index showing the number of abfreq words common to both titles as a percentage of all abfreq words in these titles less those which are common, and inputting these results into a program capable of producing a spatial representation of the overall relations of all the titles to one another. This is possible because each abfreq word is some kind

of expression of a characteristic of the title in which it occurs.

It follows of course that one can use this same general approach to plot the overall relations of, say, 20 authors from a given period or school. This can be done by comparing the abfreq list of each author with that of every other, computing their similarity index and inputting these values into an appropriate program. Note that this procedure permits the comparison of authors with respect to a large proportion of their vocabulary and that it does so while at the same time excluding from consideration those words whose frequency is such that their use might be due to chance. Put another way, it permits large scale vocabulary comparisons precisely in respect of those words whose frequency is such that they must be regarded as characteristic of the authors in question.

The above are but a few obvious examples of possible uses of what we have called abfreq words. We believe that our method for the identification of such words is now adequate and leave it to the reader to imagine other possible uses.

Table 1. Words showing ≥ 20.00 standard deviations in Kierkegaard's Fear and Trembling

Word		Absolute Frequency	Standard Deviations
Abraham	Abraham	231	97.260
Isaak	Isaac	111	69.212
Almene	the universal	104	44.570
Agnete	Agnes	40	42.159
Sara	Sarah	35	38.805
Havmanden	the Merman	29	36.363
Abrahams	Abraham's	35	35.863
Ridder	knight	50	32.858
Helt	hero	70	30.560
Kniven	the knife	18	26.428
Agamemnon	Agamemnon	14	25.265
Troens	of faith (gen.)	76	24.877
tragiske	tragic	39	24.634
Iphigenia	Iphigenia	12	23.391
Troen	faith	109	22.518
Morija	Moriah	10	21.353
Absurde	the absurd	31	21.297
han	he	1018	21.077
Enkelte	the individual	103	20.835
Morijabjerget	Mount Moriah	9	20.257

Table 2. Words showing ≥ 10.00 standard deviations in philosophy mini-text

Word		Absolute Frequency	Standard Deviations
hegelske	Hegelian	23	44.34
nyeste	newest	14	41.37
nyere	newer	27	36.97
uphilosophisk	unphilosophical	7	33.23
Smuler	fragments	10	26.58
Candidatus	candidate	3	21.75
Værkstederne	workshops	3	21.75
Geschichte	history	6	19.97
Theologien	theology	6	19.27
Hegel	Hegel	24	17.58
græske	Greek	15	15.33
Theolog	theologian	4	14.32
moderne	modern	10	13.24
Tidsskrift	periodical	5	11.99
Seminarist	seminarist	3	11.74
Philosoph	philosopher	8	11.52
Tænken	thought	11	11.29
kommende	coming	3	11.17
dum	stupid	7	11.01
1844	1844	3	10.67
Identiteten	identity	4	10.67
begribes	be understood	3	10.23
Nødvendighedens	of necessity (gen.)	3	10.23
Lære	doctrine	15	10.05

Table 3. Number and proportion of aberrant frequency words in smallest and largest texts using earlier and present method

Title Code	Word Tokens	% Corpus	Word Types	Aberrant Frequency Words							
				Earlier Method				Present Method			
				Pos.	Neg.	Tot.	%	Pos.	Neg.	Tot.	%
HCD	3034	0.16	895	441	10	451	50.39	304	16	320	35.75
EOT	4019	0.21	977	458	12	470	48.11	277	15	292	29.88
FV	4557	0.23	1300	600	12	612	47.08	433	20	453	34.85
EE1	154518	7.96	14269	2301	558	2859	20.03	1840	881	2721	19.06
SV	173007	8.91	15996	1900	561	2461	15.39	1496	766	2262	14.14
AE	214013	11.02	15481	2002	820	2822	18.23	1608	1076	2684	17.34

Table 4. Frequency criteria for the identification of abfreq words

$\hat{\lambda}$ > and w_i <		$\hat{\lambda}$ < and w_i >			
3.7	1	0.025	0	9.1	15
5.6	2	0.2	1	9.9	16
7.3	3	0.6	2	10.5	17
8.8	4	1.1	3	11.5	18
10.3	5	1.6	4	12.0	19
12.0	6	2.2	5	13.0	20
13.0	7	2.8	6	14.0	21
14.5	8	3.4	7	14.5	22
16.0	9	4.1	8	15.3	23
17.0	10	4.8	9	16.0	24
18.5	11	5.5	10	17.0	25
20.0	12	6.2	11	18.0	26
		6.9	12	18.5	27
		7.6	13	19.3	28
		8.4	14	20.0	29

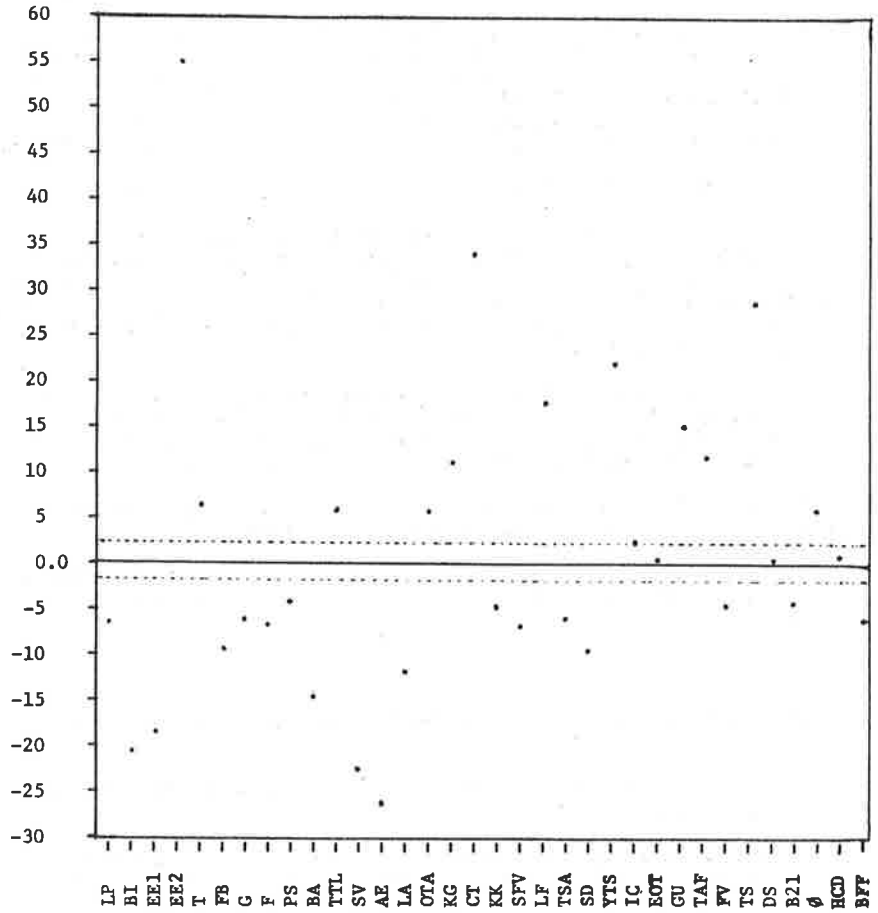


Figure 1. Number of standard deviations of Du in works of Corpus

NOTES

1 All the necessary information may be found in McKINNON (1975).

2 Cf. McKINNON (1976). The formulae described there were devised and implemented by Michael TO to whom I am most grateful.

3 The statistical formulae and the distinction between cases having expected frequency >20 and ≤20 was suggested by Dr. David SANKOFF to whom I am indebted and grateful. He has also suggested a number of helpful improvements to the text though I am of course wholly responsible for any errors which remain. The programming for this method was done by Charles MORRIS.

4 For a justification of this assumption see, e.g., FREUND (1971).

5 For example, in the case of *Fear and Trembling*, it omitted Ahnelse (inkling), allerlatterligste (most ludicrous), Amor (Cupid), Anlæg (plan), etc., but included Achilleus (Achilles), Ararat (Ararat), Augur (augur), Aulis (Aulis), etc.. Of course, it is only fair to add that many of the same words are identified by both methods.

6 A list of all words showing ≥10.00 standard deviations in the philosophy mini-text is shown in Table 2. This list is from McKINNON (Forthcoming). This same paper also provides a fuller account of the method sketched out in this paragraph.

7 A more detailed description of this method, though not using abfreq terms as such, may be found in McKINNON (1977).

8 See note 2 above. In fact, this study used our earlier method for determining aberrant frequency and so presumably should be repeated.

REFERENCES

FREUND, J.E., Mathematical Statistics. Engelwood Cliffs, Prentice-Hall 1971.

McKINNON, A., Computational Analysis of Kierkegaard's Samlede Værker. Leiden, Brill 1975.

McKINNON, A., Aberrant Frequencies as a Basis for Clustering the Works of a Corpus. Revue CIRPHO review 3, 1976, 33-52.

McKINNON, A., From Co-occurences to Concepts. Computers and the Humanities 11, 1977, 147-155.

McKINNON, A., Kierkegaard on Philosophy: The Geography of a Concept. In: CORTESE, A. (Ed.), Liber Academiae Kierkegaardiensis Annuaris (Forthcoming).

APPENDIX

Title codes used in this study

LP	Af en endnu Levendes Papirer	(From the Papers of One Still Living)
BI	Om Begrebet Ironi	*The Concept of Irony*
EE1	Enten-Eller, Forste halvbind	*Either/Or*, vol. I
EE2	Enten-Eller, Andet halvbind	*Either/Or*, vol. II
FB	Frygt og Bæven	*Fear and Trembling*
G	Gjentagelsen	*Repetition*
T	Atten opbyggelige Taler	*Edifying Discourses*, vol. I & II
BA	Begrebet Angest	*The Concept of Dread*
PS	Philosophiske Smuler	*Philosophical Fragments*
F	Forord	(Prefaces)
SV	Stadier paa Livets Vei	*Stages on Life's Way*
TTL	Tre Taler ved tænkte Leiligheder	*Thoughts on Crucial Situations in Human Life*
BFF	Bladartikler, der staar i Forhold til "Forfatterskabet"	A small part available as "A Passing Comment on a Detail in Don Juan", in *Crisis in the Life of an Actress*
AE	Afsluttende uvidenskabelig Efterskrift	*Concluding Unscientific Postscript*
LA	En literair Anmeldelse	*Two Ages*
OTA	Opbyggelige Taler i forskjellig Aand	In English as two separate volumes: *Purity of Heart* and *The Gospel of Suffering*
KK	Krisen og en Krise i en Skuespillerindes Liv	*Crisis in the Life of an Actress*
KG	Kjerlighedens Gjerninger	*Works of Love*
TSA	Tvende ethisk-religieuse Smaa-Afhandlinger	"Two Minor Ethico-Religious Treatises", in *The Present Age*
CT	Christelige Taler	*Christian Discourses*
SD	Sygdommen til Døden	*The Sickness Unto Death*
IC	Indøvelse i Christendom	*Training in Christianity*
SFV	Synspunktet for min Forfatter Virksomhed	*The Point of View for my Work as an Author*
LF	Lilien paa Marken og Fuglen under Himlen	"The Lilies of the Field and the Birds of the Air", in *Christian Discourses*
FV	Om min Forfatter-Virksomhed	"On My Work as an Author", in *The Point of View for my Work as an Author*
YTS	"Ypperstepræsten" - "Tolderen" - "Synderinden"	"'The High Priest' - 'The Publican'-'The Woman that was a Sinner'", in *Christian Discourses*

TAF	To Taler ved Altergangen om Fredagen	"Two Discourses at the Communion on Fridays", in *For Self-Examination ...*
EOT	En opbyggelig Tale	"An Edifying Discourse", in *Training in Christianity*
GU	Guds Uforanderlighed	"The Unchangeableness of God", in *For Self-Examination & Judge for Yourselves!*
TS	Til Selvprøvelse, Samtiden anbefalet	"For Self-Examination", in *For Self-Examination and Judge for Yourselves!*
DS	Dømmer selv!	"Judge for Yourselves!", in *For Self-Examination and Judge for Yourselves!*
B21	Bladartikler 1854-55 [I-XXI]	"Articles in the *Fatherland*", in *Attack upon "Christendom" 1854-55*
Ø	Øieblikket	"The Instant", Nos. I-X, in *Attack upon "Christendom" 1854-55*
HCD	Hvad Christus dømmer om officiel Christendom	"What Christ's Judgement is about Official Christianity", in *Attack ...*

Measuring Vocabulary Richness in Literary Works: A New Proposal and a Re-Assessment of Some Earlier Measures

D.A. Ratkowsky, M.H. Halstead and Linda Hantrais

SUMMARY

An expression originally derived for relating the number of elements in a computer program with the program's software vocabulary is shown to be a valuable indicator of vocabulary richness defined in terms of variety for literary works. For any text of known length with a known number of different vocabulary items it is possible to calculate a measure which we refer to as its 'richness ratio'. Examination of a group of 41 heterogeneous literary works revealed richness ratios varying from that of an average computer program to values that are interpreted as indicating exceptional richness of vocabulary. An examination was made of the 38 plays of Shakespeare, and one result was the finding that the comedies were significantly less rich in vocabulary than either the tragedies or historical plays. From analysis of covariance, a value 0.45 was estimated to be a significant difference between two values of the richness ratio. Exa-

---

D.A. Ratkowsky is a Principal Research Scientist with CSIRO, Division of Mathematics and Statistics, Tasmanian Regional Laboratory, Stowell Avenue, Hobart, Tasmania, 7000, Australia. M.H. Halstead is Professor, Department of Computer Sciences, Purdue University, West Lafayette, Indiana 47907, United States of America. Linda Hantrais is Lecturer, Department of Modern Languages, University of Aston, Gosta Green, Birmingham B4 7ET, England.

mination of a number of other richness measures showed that the 'logarithmic type-token ratio' gave very similar results to, and was very highly correlated with, the richness ratio. Both measures are influenced by the length of the text; for example the richness ratio tends to decrease by 0.192 for every increase in text length of ten thousand words. Examples are given which illustrate the use of the richness ratio for comparing the vocabulary richness of different texts.

INTRODUCTION

If V is the number of different vocabulary items in a text of total number of words N, then an obvious measure of vocabulary richness in terms of variety, but one which does not necessarily imply literary merit, is the ratio V/N. However, since this quantity is dependent upon text length, being large for small N and decreasing steadily with increasing N, it cannot be exploited in comparative studies of vocabulary richness for works of different lengths. To overcome this, GUIRAUD (1954) proposed instead that the ratio V/√N was constant, but he later admitted that the value of the "constant" changed with the size of the sample (GUIRAUD, 1959). A further attempt to establish a richness ratio using V and N was made by HERDAN (1960), who proposed that the so-called "logarithmic type-token ratio", log V/log N, was constant. MULLER (1964) showed that both the GUIRAUD (1954) and HERDAN (1960) ratios are markedly influenced by the size of the text. WEITZMAN (1971) strongly questioned the utility of the HERDAN (1960) ratio. We will return to a consideration of these ratios later in this study, but first we propose

a new measure of vocabulary richness.

THE SOFTWARE RELATIONSHIP

From first principles, HALSTEAD (1972) has derived a relationship between the total number of elements in a computer program and its software vocabulary. Details of the derivation, which is founded on combinatorial arithmetic, are given in a book on software science (HALSTEAD, 1977), where experimental evidence is presented confirming the relationship for computer software. Some simplifications are necessary (see Appendix) to apply the software relation to literary texts, resulting in the following relationship between the total text length and unique vocabulary:

$$N = V \log_2 (V/2). \qquad (1)$$

This equation can be used with natural (Naperian) logarithms as follows:

$$N = 1.4427\ V \log_e (V/2), \qquad (1a)$$

or with base 10 logarithms as follows:

$$N = 3.322\ V \log_{10} (V/2). \qquad (1b)$$

TESTING THE SOFTWARE RELATIONSHIP ON 41 LITERARY TEXTS

We examined a number of prose and poetry works, both ancient and modern, for which N and V values were available. The sources of these data were YULE (1944), HERDAN (1960), HERDAN (1964), MULLER (1967), GUIRAUD (1954), GUIRAUD (1959) and HANTRAIS (1976). The works in these sources form a very heterogeneous group, as some authors, such as YULE (1944), counted only nouns or some other specific part of speech. Various languages, ancient and modern, are represented. There are inconsistencies in the manner in which the various forms of different parts of speech are dealt with, as some quantitative linguists group all forms of a single vocabulary item, while others count each form as a separate item. Nevertheless, these works constitute a broad base upon which to examine the validity of the software relationship.

Values of N and V for 41 texts having N values between 3000 and 35000 are given in Table 1. The software relationship can be used in one of two ways. From the known values of V and N for a text, either V can be inserted into equation (1) and a predicted value of text length denoted by $\hat{N}$ obtained from which a 'richness ratio' defined as $\hat{N}/N$ can be calculated, or alternatively, N can be inserted into equation (1) to predict $\hat{V}$, which requires a trial-and-error or iterative procedure, after which a 'richness ratio' defined as $V/\hat{V}$ can be calculated. It is obvious that $\hat{N}/N$ is much easier to calculate than $V/\hat{V}$ because no iteration is necessary.

Table 1 gives results for both procedures and it can be seen that although the two definitions of richness give somewhat dif-

Table 1. Length, vocabulary sizes and richness ratios for each of 41 diverse literary works

Author	Title of work	Source of data	N	V	$\hat{N}$	$\hat{V}$	$\hat{N}/N$	$V/\hat{V}$
Baudelaire	Les fleurs du mal	Guiraud (1954)	16704	4009	43975	1714	2.633	2.338
Rimbaud	Les illuminations	do.	4150	2175	21939	518	5.286	4.201
Mallarmé	Les poesies	do.	4550	2170	21881	560	4.809	3.877
Valéry	Les poesies	do.	9200	2595	26483	1022	2.879	2.509
Claudel	Les cinq grandes odes	do.	10000	3115	33035	1099	3.303	2.835
Apollinaire	Alcools	do.	8200	2830	29629	926	3.612	3.056
Rimbaud	Vers 1869-1870	Guiraud (1959)	6000	1590	15319	708	2.553	2.244
Rimbaud	Vers 1870-1871	do.	6000	1630	15763	708	2.627	2.301
Rimbaud	Vers 1872	do.	3000	1050	9488	394	3.163	2.667
Rimbaud	Saison en enfer	do.	7225	1855	18285	831	2.531	2.233
Macaulay	Essay on Bacon	Yule (1944)	8045	2048	20480	911	2.546	2.248
Macaulay	Four essays	do.	20178	3543	38232	2022	1.895	1.753
Bunyan	Pilgrim's Progress, Part I	do.	4016	1020	9174	504	2.284	2.026
Bunyan	Pilgrim's Progress, Part II	do.	4047	1005	9018	507	2.228	1.983
Bunyan	Mr. Badman	do.	3992	1030	9279	501	2.324	2.056
Bunyan	Holy War	do.	4001	996	8924	502	2.230	1.984
Pushkin	Captain's Daughter	Herdan (1964)	28591	4783	53683	2743	1.878	1.743
Aelfric	Homilies	Herdan (1960)	9752	1834	18048	1075	1.851	1.706
A Kempis	Miscellaneous works	Yule (1944)	8203	1406	13297	926	1.621	1.518
A Kempis	Vallis Liliorum	do.	3143	748	6393	409	2.034	1.827
A Kempis	Soliloquium animae	do.	4811	1025	9226	587	1.918	1.746

Table 1 (continued). Length, vocabulary sizes and richness ratios for each of 41 diverse literary works

Author	Title of work	Source of data	N	V	$\hat{N}$	$\hat{V}$	$\hat{N}/N$	$V/\hat{V}$
Gerson	Theological works	do.	8196	1754	17148	926	2.092	1.895
(Disputed)	De imitatione christi (nouns)	do.	8225	1168	10734	928	1.305	1.258
(Disputed)	do. (adjectives)	do.	5053	529	4257	612	0.842	0.864
(Disputed)	do. (verbs)	do.	8116	1157	10617	918	1.308	1.261
Brassens	Chansons	Hantrais (1976)	29583	3567	38525	2827	1.302	1.262
Brel	Chansons	do.	17488	1827	17969	1784	1.028	1.024
Ferré	Chansons	do.	12729	2163	21800	1354	1.713	1.598
Corneille	Sertorius	Muller (1967)	17710	1552	14899	1804	0.841	0.860
	New Testament - Matthew	Herdan (1960)	18278	1691	16443	1854	0.900	0.912
	New Testament - Mark	do.	11229	1345	12634	1214	1.125	1.107
	New Testament - Luke	do.	19404	2055	20560	1954	1.060	1.052
	New Testament - John	do.	15420	1011	9080	1599	0.589	0.632
	New Testament - Acts	do.	18374	2038	20366	1863	1.108	1.094
	New Testament - Paul	do.	32303	2648	27462	3054	0.850	0.867
	New Testament - Hebrews	do.	4942	1038	9362	600	1.894	1.729
	New Testament - Revelations	do.	9818	917	8107	1081	0.826	0.848
St. Paul	Letters - Romans	do.	7094	1068	9677	818	1.364	1.306
St. Paul	Letters - 1 Corinthians	do.	6807	967	8623	789	1.267	1.225
St. Paul	Letters - 2 Corinthians	do.	4448	792	6834	549	1.537	1.442
St. Paul	Gospel - Basic English Version	Yule (1944)	3088	296	2134	403	0.691	0.734

ferent values, they are very highly correlated ($r = 0.999$, $df = 39$, $p < 0.001$). The highest values of the richness ratio are obtained by the French symbolist poets studied by GUIRAUD (1954), the compactness of poetry compared with prose generally results in a more varied vocabulary and the poet symbolists were particularly noted for their rich vocabulary, although the ratios are probably artifically inflated because of the counting procedure that GUIRAUD (1954) used. Low values for the richness ratio, of the order of unity or below unity, characterize biblical works such as the New Testament and the gospels, in which repetition of words and phrases is frequent. The three modern poet singers studied by HANTRAIS (1976) were also relatively low in richness, since songs are characterized by repetition of words and phrases, or even of entire stanzas, as in choruses. Some other early works such as *De imitatione Christi*, of uncertain authorship, also have low richness ratios.

When considering the validity of applying a relationship derived for a computer program in the assessment of the richness of literary works, it must be remembered that the object of a computer programmer is quite different from that of a literary author. A programmer who used synonyms in a computer program where they were not necessary would justifiably be viewed as a inefficient programmer. In view of this, it must be seen as an interesting fact that the richness ratios of some literary works are no greater than that of the average computer program.

We examined whether the richness ratio depends upon the length of text by calculating the correlation coefficient between $V/\hat{V}$ and N for the 41 works in Table 1. Its value was -0.39, which although not high, is nonetheless significant ($d.f. = 39$, $p < 0.02$). An almost identical result is obtained for the correla-

tion coefficient of $\hat{N}/N$ with N. Thus, in common with the relationships proposed by GUIRAUD (1954) and HERDAN (1960), the software relationship is not completely free of some dependence upon text length. Inspection of Table 1 shows that, within a group of related works, there is a tendency for short works to have a relatively high richness ratio, and, correspondingly, for long works to have a relatively low richness ratio, which explains the negative correlation coefficient.

TESTING THE SOFTWARE RELATIONSHIP ON THE PLAYS OF SHAKESPEARE

As already mentioned, the works examined above form a very heterogeneous group since the vocabulary counts are not based on a coherent set of criteria and because the works examined are taken from different periods, genres and languages. In order to eliminate any bias resulting from discrepancies in criteria, we have examined the 38 plays of a single author, Shakespeare, the vocabulary of which was treated in a consistent manner in all respects. A complete concordance of the works of Shakespeare, including the 38 plays, was published by SPEVACK (1968). However, that work did not summarize, for each play separately, the breakdown of the vocabulary size V into the frequency distribution of V_i, i.e. the vocabulary items appearing exactly i times. Professor Spevack has subsequently carried out the required calculations and has kindly made the information available to us.

Values of N and V for the 38 plays are given in Table 2. The V values are slightly larger than those published in SPEVACK (1968) as a result of corrections made to some earlier errors

Table 2. Length, vocabulary sizes and richness measures for each of the plays of William Shakespeare

	N	V	$\hat{N}$	$\hat{V}$	$\hat{N}/N$	$V/\hat{V}$	V_f	log V/log N	Genre*
Macbeth	16436	3322	35538	1690	2.162	1.965	3008	.8353	T
King Lear	25221	4797	53860	2457	2.136	1.952	3284	.8362	T
The Tempest	16036	3163	33613	1655	2.096	1.912	2913	.8323	C
Richard II	21809	4073	44770	2164	2.053	1.882	3123	.8320	H
Henry VI, Part I	20515	3824	41685	2051	2.032	1.864	3051	.8308	H
Henry V	25577	4617	51585	2488	2.017	1.856	3188	.8313	H
Titus Andronicus	19790	3662	39690	1988	2.006	1.842	2981	.8295	T
Henry IV, Part II	25706	4566	50942	2499	1.982	1.827	3114	.8298	H
Loves Labours Lost	21033	3825	41697	2096	1.982	1.825	3002	.8288	C
Troilus and Cressida	25516	4528	50463	2483	1.978	1.824	3094	.8296	T
Timon of Athens	17748	3289	35138	1807	1.980	1.820	2857	.8277	T
Midsummer Nights Dream	16087	3012	31796	1659	1.977	1.815	2775	.8270	C
Pericles	17723	3270	34907	1805	1.970	1.811	2833	.8272	T
Hamlet	29551	5046	57024	2824	1.930	1.787	3117	.8283	T
King John	20386	3596	38881	2040	1.907	1.763	2870	.8251	H
Henry VI, Part II	24450	4110	45230	2392	1.850	1.719	2954	.8235	H
Two Noble Kinsmen	23403	3926	42946	2302	1.835	1.706	2853	.8226	T
Comedy of Errors	14369	2549	26295	1504	1.830	1.695	2510	.8193	C
Anthony and Cleopatra	23742	3943	43156	2331	1.818	1.692	2845	.8218	T

*Genre: C = comedy; T = tragedy; H = historical play

Table 2 (continued). Length, vocabulary sizes and richness measures for each of the plays of William Shakespeare

	N	V	$\hat{N}$	$\hat{V}$	$\hat{N}/N$	$V/\hat{V}$	V_f	log V/log N	Genre*
Cymbeline	26778	4260	47101	2590	1.759	1.645	2874	.8197	T
A Winters Tale	24543	3924	42921	2399	1.749	1.635	2791	.8186	C
Henry IV, Part I	23955	3830	41759	2349	1.743	1.630	2768	.8182	H
Romeo and Juliet	23913	3789	41253	2345	1.725	1.615	2752	.8173	T
Taming of the Shrew	20411	3276	34980	2042	1.714	1.604	2623	.8157	C
Twelfth Night	19401	3115	33035	1953	1.703	1.595	2559	.8147	C
Alls Well That Ends Well	22550	3543	38232	2228	1.695	1.590	2655	.8154	C
Merry Wives of Windsor	21119	3335	35696	2104	1.690	1.585	2591	.8147	C
Two Gentlemen of Verona	16883	2736	28503	1730	1.688	1.581	2457	.8130	C
Measure for Measure	21269	3336	35708	2117	1.679	1.576	2588	.8141	C
Merchant of Venice	20921	3282	35053	2086	1.675	1.573	2586	.8138	C
Henry VI, Part III	23295	3605	38991	2292	1.674	1.573	2698	.8144	H
Coriolanus	26579	4047	44447	2573	1.672	1.573	2737	.8153	T
Richard III	28309	4268	47201	2720	1.667	1.569	2833	.8154	H
Othello	25887	3926	42946	2514	1.659	1.561	2711	.8144	T
Henry VIII	23325	3571	38574	2295	1.654	1.556	2655	.8134	H
As You Like It	21305	3282	35053	2120	1.645	1.548	2565	.8123	C
Julius Caesar	19110	2879	30205	1928	1.581	1.493	2413	.8080	T
Much Ado About Nothing	20768	2965	31233	2073	1.504	1.430	2355	.8042	C

*Genre: C = comedy; T = tragedy; H = historical play

and the fact that unique words appearing in square brackets in the original text of Shakespeare are now counted as separate items in determining the total vocabulary size.

Table 2 gives values fo $\hat{N}$, $\hat{V}$, $\hat{N}/N$ and $V/\hat{V}$ which were calculated using the software relation (equation (1) or (1a) or (1b)) in an identical fashion to that used for the 41 works in Table 1. The correlation coefficient between $V/\hat{V}$ and $\hat{N}/N$ for these plays is 0.9998, which indicates that these two quantities can be used interchangeably as a measure of richness. The plays in Table 2 are listed in order of decreasing richness and it appears that *Macbeth* has the richest vocabulary and *Much Ado About Nothing* the poorest.

As part of a study of the behaviour of the software relation, we reduced the text length of each of Shakespeare's plays to a common size N_f, employing a probabilistic method described in detail in RATKOWSKY and HANTRAIS (1975), which requires the frequency distribution of V_i for each play. The number of different words V_f in the reduced text then becomes another possible measure of vocabulary richness of the play. Since the shortest play of Shakespeare has an N of 14369, we chose the round number 14000 to be the common text length after text reduction. Values of V_f are presented in Table 2.

The correlation coefficient between $\hat{N}/N$ and V_f is 0.842 and that between $V/\hat{V}$ and V_f is 0.851, and as both coefficients are highly significant ($p < 0.001$) this supports the contention that these richness ratios are indeed measuring richness. An examination of the values of V_f does not reveal any great anomalies, except that the positions of the longest plays (*Hamlet* and *Richard III*) seem too low on the richness list, whereas the position of the shortest play (*Comedy of Errors*) seems too high.

This undoubtedly reflects the fact that the software relationship is somewhat dependent upon text length, which will be investigated more fully in the next section, but in general there are texts of all lengths both near the top and near the bottom of Table 2.

TEXT LENGTH DEPENDENCE AND THE VARIANCE OF THE RICHNESS RATIO

Comparing the order of the plays in Table 2, which is based on decreasing richness of vocabulary, with the approximate chronological order of the writing of these plays, as given by WILLIAMS (1970, Table 27), there is no apparent correlation between richness and time of writing. This lack of dependence was later confirmed by an analysis of covariance of the richness ratio, using the serial order of time of writing as a covariate, which showed that the covariate was not significant. Plays from Shakespeare's early, middle and later periods can be found distributed throughout the range of richness of Table 2 and the question arises whether this range may represent nothing more than random variation. This possibility can be explored in several ways, one of which involves examining the frequency distribution of the $\hat{N}/N$ values divided into class intervals differing by (say) 0.05 units. The frequencies are given in the left-hand half of Table 3 and indicate that the distribution of $\hat{N}/N$ is distinctly bi-modal, with one peak being located at a value of approximately 1.98, and the other peak at approximately 1.68 (by reference to the actual data in Table 2).

In order to assess whether the bi-modal distribution could be explained by genre and text length N, an analysis of covariance was employed. The genre of each of the plays, given by the compilers of the First Folio as comedy, tragedy or history, is listed in Table 2. For each genre, Table 4 presents mean values of N, the original unadjusted $\hat{N}/N$, and $\hat{N}/N$ adjusted for the effect of N used as a covariate. The mean values of N show that comedies tend to be shorter on the average than either tragedies or historical plays. The mean value for $\hat{N}/N$ for comedy is lower

Table 3. Frequency distribution of the richness ratio $\hat{N}/N$ and the type-token ratio log V/log N for the plays of Shakespeare

$\hat{N}/N$		log V/log N	
Class intervals	Frequency	Class intervals	Frequency
1.50 - 1.55	1	.8025 - .8050	1
1.55 - 1.60	1	.8050 - .8075	0
1.60 - 1.65	1	.8075 - .8100	1
1.65 - 1.70	10	.8100 - .8125	1
1.70 - 1.75	5	.8125 - .8150	8
1.75 - 1.80	1	.8150 - .8175	5
1.80 - 1.85	3	.8175 - .8200	4
1.85 - 1.90	1	.8200 - .8225	1
1.90 - 1.95	2	.8225 - .8250	2
1.95 - 2.00	6	.8250 - .8275	3
2.00 - 2.05	3	.8275 - .8300	6
2.05 - 2.10	2	.8300 - .8325	4
2.10 - 2.15	1	.8325 - .8350	0
2.15 - 2.20	1	.8350 - .8375	2

Table 4. Mean values of N and $\hat{N}/N$ (adjusted and unadjusted) for each genre

	N	$\hat{N}/N$ (unadjusted)	$\hat{N}/N$ (adjusted for N)
Comedy	19764	1.759	1.716
Tragedy	22957	1.872	1.891
History	23733	1.858	1.891

than that for the other genres, but analysis of variance shows the difference to be non-significant [$F(2,35) = 1.83$; $p > 0.05$]. However, when text length N is used as a covariate, a significant difference emerges between the adjusted mean values of $\hat{N}/N$, indicating a lower richness for comedy than for tragedy or history [$F(2,34) = 4.24$, $p < 0.05$]. This is explained by the fact that the regression coefficient, pooled within genres, of $\hat{N}/N$ upon N is negative, being -0.0000192 (which means that the richness ratio tends to diminish by 0.192 for every increase in text size of ten thousand words). As both tragedy and history have bigger mean text lengths than comedy, the difference in richness tends to be masked somewhat unless covariate analysis is used.

The above results, however, do not explain the bi-modality of the richness ratio distribution. A histogram of the adjusted values of $\hat{N}/N$ for each play exhibits a similar bi-modality to that of the unadjusted values. An additional useful feature of the analysis of covariance is that it provides an estimate of the variance of a single value of the richness ratio, from

which one can obtain an estimate of the least significant difference between two values of $\hat{N}/N$. The estimated residual variance from the analysis of covariance was 0.0247, so that for a 5 % significance level, the estimated least significant difference is $2.04[2(0.0247)]^{\frac{1}{2}} = 0.45$. Due to the bi-modality, this figure is likely to be an over-estimate, but it does provide an objective basis for comparing the richnesses of different texts.

APPLICATIONS OF THE RICHNESS RATIO

The results of the previous section can be used to re-examine some of the works in Table 1 and test the validity of the criterion that 0.45 may be a significant difference between two values of $\hat{N}/N$. It is of interest to compare the miscellaneous works of Thomas à Kempis with the theological works of Jean Charlier de Gerson and with the work *De imitatione christi*. An extensive study was made by YULE (1944) to determine the authorship of *De imitatione christi*, which has been most generally ascribed to à Kempis, but other scholars have attributed it to Gerson. YULE (1944) confined himself in this comparison to the nouns, as was his custom, and since the text length of the work of disputed authorship was 8225 words, he selected undisputed works of à Kempis and of Gerson such that the total text length would closely approximate this figure and thereby minimize bias due to different text length. The works are listed in Table 5 for ease of comparison.

YULE (1944) concluded, after using a comprehensive variety of statistical procedures, that Gerson was a highly unlikely candidate for authorship of *De imitatione christi*, but that à Kempis

Table 5.

Author	Title	N	$\hat{N}/N$
Thomas à Kempis	Miscellaneous works	8203	1.621
Gerson	Theological works	8196	2.092
(Disputed)	De imitatione christi	8225	1.305

could not be excluded, although he noted that the latter author's vocabulary was generally much richer than in the disputed work. The results we have obtained here for $\hat{N}/N$ (Table 5) support this conclusion. Because the values of N are so close to each other, no adjustment for the effect of text length was necessary. The difference in the richness ratio between Gerson's theological works and the disputed work, i.e. 0.79, is too large compared with 0.45 for Gerson to be the likely author. The corresponding difference for the works of à Kempis and the disputed work, i.e. 0.32, is less than 0.45, meaning that à Kempis cannot be excluded on the evidence of the noun-distribution.

The three poetic singers studied by HANTRAIS (1976) provide another interesting comparison. Table 6 presents values of N, unadjusted values of $\hat{N}/N$, and values of $\hat{N}/N$ adjusted to a common text length of N = 20000. The adjustment is made using the result obtained in the previous section, namely that the richness ratio decreases by 0.192 for every increase in text length of 10000 words. For example, since Brassens used 9583 words in excess of 20000, his $\hat{N}/N$ value is corrected upwards by 0.192(9583)/10000 = 0.184 to give him an adjusted value of 1.302 + 0.184 = 1.486. On the other hand, the values for Brel and Ferré will

Table 6.

	N	$\hat{N}/N$ (unadjusted)	$\hat{N}/N$ (adjusted to N = 20000)
Brassens	29583	1.302	1.486
Brel	17488	1.028	0.980
Ferré	12729	1.713	1.573

be corrected downwards by amounts equal to 0.192(17488-20000)/10000 and 0.192(12729-20000)/10000 respectively. The adjusted values presented in Table 6 show clearly, using the criterion of 0.45, that the difference in richness ratio of 0.087 between Brassens and Ferré is non-significant, whereas the differences between the values for each of these singers and Brel, being 0.506 and 0.593 respectively, indicate that the latter had a significantly poorer vocabulary than the former two singers. The value N = 20000 used above as the common text length for adjustment was merely a convenient round figure. Exactly the same differences between singers would have been obtained for any other common N value chosen.

A RE-EXAMINATION OF OTHER PROPOSED RICHNESS MEASURES

As mentioned in the Introduction, vocabulary richness measures that have previously been proposed include V/N, $V/\sqrt{N}$ and the logarithmic type-token ratio log V/log N. The effect of text

length N upon these measures was tested using the 38 plays of Shakespeare. Correlation coefficients were calculated between each of these measures and the text length N. The measures V/N and $V/\sqrt{N}$ were significantly dependent upon text length and need not be considered further. A somewhat surprising result, in view of its criticism by MULLER (1964) and WEITZMAN (1971), is the fact, that log V/log N has only a small dependence upon N. Its values for each of the 38 plays are given in Table 2. The correlation coefficient between log V/log N and $\hat{N}/N$ is 0.991, this high value explaining why these two measures give very similar orderings of the plays in terms of vocabulary richness. The frequency distribution of the values of log V/log N is given in the right-hand half of Table 3 and exhibits a bimodal pattern very similar to that for $\hat{N}/N$. An analysis of covariance of log V/log N in terms of genre, using N as a covariate, was carried out in an identical fashion to that used for $\hat{N}/N$, with closely parallel results. For example, genre was non-significant before adjustment [$F(2,35) = 2.84$, $p > 0.05$] but became significant when the covariate N was included [$F(2,34) = 4.24$, $p < 0.05$]. The pooled regression coefficient was -0.0063 per 10000 words, and the least significant difference between two values of log V/log N is estimated from the residual variance to be 0.022.

DISCUSSION AND CONCLUSIONS

It is clear that there is a very close relationship between the richness ratio $\hat{N}/N$, obtained from the theoretically derived software relation, and the empirical logarithmic type-token ratio log V/log N. Both ratios exhibit a dependence upon text length, but this is insignificant if the lengths of the texts under study

are of comparable magnitude. For larger differences, adjustments can be made to the richness measures by the method used in the example given in Table 6. Nevertheless, it seems imprudent to compare text sizes that differ very greatly from each other. The use of log V/log N was criticized by MULLER (1964) and WEITZMAN (1971), but those authors used the ratio to compare texts that were of widely different lengths (e.g. a factor of 5.2 in MULLER's study of some texts of Corneille, and a factor of 21.6 in WEITZMANN's study of the thirteen Pauline letters). Provided that the texts under consideration are of comparable length either log V/log N or the richness ratio $\hat{N}/N$ can be useful indicators of the relative vocabulary richness of those works. As log V/log N is a strictly empirical ratio, we recommend the use of $\hat{N}/N$, which has been derived from first principles, albeit with some simplifying approximations.

When using the ratio $\hat{N}/N$ as a measure of the relative richness of two or more literary works of comparable size, a difference of 0.45 can be used as a guideline for assessing significance. Differences greater than this probably indicate that the works differ in richness of vocabulary, whereas works with values differing by much less than this are likely to have comparable vocabulary richnesses.

An unexpected result of the study of the plays of Shakespeare was the emergence of the bi-modal distribution of vocabulary richness. We are unable to offer any explanation for this bimodality, but the search for the explanation could provide an interesting challenge for Shakespeare scholars.

ACKNOWLEDGMENT

The authors are grateful to Prof.Dr. Marvin Spevack, Westfälische Wilhelms-Universität, Englisches Seminar, Johannisstrasse 12-20, 4400 Münster, Federal Republic of Germany, for preparing the word frequency distributions for each of Shakespeare's plays, and for permitting us to use them in the present study.

References

GUIRAUD, P. (1954). Les caractères statistiques du vocabulaire: essai de méthodologie, Paris: Presses universitaires de France.

GUIRAUD, P. (1959). Problèmes et méthodes de la statistique linguistique, Dordrecht-Holland: D. Reidel.

HALSTEAD, M.H. (1972). "Natural laws controlling algorithm structure?" Association for Computing Machinery SIGPLAN Notices, 7(2), 19-26.

HALSTEAD, M.H. (1977). Elements of software science. New York: Elsevier North-Holland.

HANTRAIS, L. (1976). Le vocabulaire de Georges Brassens. Paris: Editions Klincksieck.

HERDAN, G. (1960). Type-token mathematics: a textbook of mathematical linguistics, S'Gravenhage: Mouton et Cie.

HERDAN, G. (1964). Quantitative linguistics, London: Butterworths.

MULLER, Ch. (1964). Essai de statistique lexicale: L'illusion comique de Pierre Corneille, Paris: Klincksieck.

MULLER, Ch. (1967). Etude de statistique lexicale: le vocabulaire du théâtre de Pierre Corneille, Paris: Larousse.

RATKOWSKY, D.A. and HANTRAIS, L. (1975). "Tables for comparing the richness and structure of vocabulary in texts of different lengths", Computers and the humanities, 9, 69-75.

SPEVACK, M. (1968). A complete and systematic concordance to the works of Shakespeare, Vols. 1-6, Hildesheim: George Olms.

WEITZMAN, M. (1971). "How useful is the logarithmic type/token ratio?", Journal of Linguistics, 7, 237-243.

WILLIAMS, C.B. (1970). Style and vocabulary: numerical studies, London: Griffin.

YULE, G.U. (1944). The statistical study of literary vocabulary, Cambridge: Cambridge University Press.

Appendix: Some considerations needed in applying the software relationship for computer programs to literary texts.

If N_c is the total length of a computer program (or literary text), and η_1 the count of unique operators (e.g. +, -, x, /, =, ;, IF, SQRT, GØ TØ A, etc, for computer programs, and articles, pronouns, auxiliary verbs, conjunctions, prepositions, exclamations, punctuation, and fonts for poetry and prose), and η_2 the count of unique operands (variables and constants in programs, "content words" in literary texts) then the following relationship may be derived (HALSTEAD, 1977), which results from the use of combinatorial arithmetic:

$$N_c = \eta_1 \log_2 \eta_1 + \eta_2 \log_2 \eta_2 \qquad \text{(A1)}$$

As the breakdown into operators and operands is generally unavailable for literary works, we seek to replace η_1 and η_2 by their sum η. When $\eta_1 = \eta_2$, as it might be for short computer programs or literary texts, equation (A1) reduces to the following:

$$N_c = \eta \log_2(\eta/2) \qquad \text{(A2)}$$

Even if $\eta_1 \ll \eta_2$, as it might be for very long programs or texts, the error in equation (A2) must always be less than $1/\log_2\eta$, or a few percent at most.

A further problem arises in applying equation (A2) to literary works, because N_c consists not only of words, but also of punctuation marks, symbols and type font changes. Similarly, η includes the count of unique punctuation marks, symbols and

fonts as well as words. Quantitative linguists usually ignore punctuation when they break down a literary work into vocabulary units. Also, in some poetry the author may dispense with punctuation. Moreover, symbols and font changes normally do not occur in literary works. Preliminary estimates of the effect of ignoring these factors indicate a maximum error of about six percent. Thus, we suggest that equation (A2) may be used for literary texts after replacing N_c and η by the total text length N and unique vocabulary V, respectively, giving the following:

$$N = V \log_2 (V/2) \qquad \text{(A3)}$$

which is equation (1) of the main body of this work.

Wahrscheinlichkeitstheoretisches Modell für das Zusammentreffen der Personen im Drama

H. Sterner, Bochum

Um die Beziehungen zwischen den Personen eines Dramas quantitativ zu erfassen, wurden in der mathematisch-statistischen Literatur zum Dramenstudium bereits verschiedene Maße vorgeschlagen, die auf der Häufigkeit basieren, mit der die Personen eines Dramas im Verlauf der Handlung miteinander zusammentreffen (vgl. DINU 1968, 1977, MARCUS 1973).

Ein beträchtlicher Nachteil der bisherigen Methoden bestand darin, daß sie es lediglich erlaubten, eine Rangfolge der möglichen Personenpaare (nach der Größe der paarweise gemessenen Distanz- oder Ähnlichkeitsmaße) aufzustellen. Man mußte sich auf einer solchen Ordinalskala also darauf beschränken, festzustellen, daß die Beziehung zwischen zwei bestimmten Personen eines Dramas enger (oder weniger eng) als zwischen einem anderen Personenpaar ist. Eine Methode, die es ermöglichte zu überprüfen, ob die berechneten Maßzahlen überhaupt statistisch signifikant waren, wurde nicht angegeben. Das Fehlen einer solchen Prüfungsmethode führte gelegentlich zu recht gewagten Interpretationen. So wurde etwa aus der Tatsache, daß zwei Ehegatten in 9 Szenen - statt wie erwartet in 9.65 (!) Szenen (sh. Berechnung unten) - zusammen auftreten, der Schluß gezogen, "mit gutem Recht an ihrer (der Frau) ehelichen Treue zweifeln (zu) dürfen" (MARCUS 1973: 312). Die Wahrscheinlichkeit für eine so geringfügige Abweichung der beobachteten Zahl vom Erwartungswert beträgt .4592, ist also so groß, daß man keinesfalls den Schluß ziehen kann, daß die Begegnung der beiden Personen vermieden wird.

Im folgenden soll gezeigt werden, wie sich solche Wahrscheinlichkeiten berechnen lassen. Bezeichnet man mit N die Anzahl der Szenen eines Dramas, mit n_i die Zahl der Szenen, in denen eine bestimmte Person P_i auftritt, so kann man das Verhältnis

$$p_i = \frac{n_i}{N}.$$

als die Wahrscheinlichkeit betrachten, daß die betreffende Person in einer zufällig gewählten Szene auftritt.

Die entsprechende Wahrscheinlichkeit für eine andere Person P_j sei

$$p_j = \frac{n_j}{N}.$$

Die Wahrscheinlichkeit für das Auftreten zweier unabhängiger Ereignisse ist gleich dem Produkt der Einzelwahrscheinlichkeiten. Somit ergibt sich die Wahrscheinlichkeit, daß zwei Personen P_i und P_j in irgendeiner Szene zusammenkommen - unter der Annahme, daß ihre Auftritte unabhängig voneinander sind -

$$p_{ij} = \frac{n_i n_j}{N^2}.$$

Multipliziert man diese Wahrscheinlichkeit mit der Gesamtzahl N der Szenen, so erhält man die zu erwartende Anzahl $E(n_{ij})$ der Szenen, in denen die beiden Personen zusammentreffen müßten:

$$E(n_{ij}) = \frac{n_i n_j}{N} \qquad (1)$$

Tatsächlich sind aber die Begegnungen der Personen in einem Drama wohl kaum zufällig. Durch Vergleich der beobachteten Zahl gemeinsamer Szenen mit der bei Unabhängigkeit der Auftritte zweier Personen zu erwartenden kann man feststellen, ob bestimmte Personen besonders häufig vom Dramenautor zusammengeführt werden oder ob im Gegenteil ihre Begegnung vermieden wird. Um eine Entscheidung darüber treffen zu können, ist es notwendig, die Wahrscheinlichkeit der beobachteten oder noch extremeren Anzahl der gemeinsamen Auftritte zu kennen.

Das Problem läßt sich durch eine elementare Überlegung lösen. Gesucht ist nach der Wahrscheinlichkeit, daß eine Person P_j in genau n_{ij} Szenen mit einer anderen Person P_i zusam-

mentrifft und in den restlichen $n_j - n_{ij}$ Szenen nicht mit P_i zusammenkommt.

Diese Wahrscheinlichkeit berechnet sich folgendermaßen: Die Zahl der Möglichkeiten, n_{ij} Szenen aus den n_i Szenen der Person P_i auszuwählen, multipliziert mit der Zahl der Möglichkeiten, $n_j - n_{ij}$ Szenen aus den restlichen $N - n_i$ Szenen auszuwählen, dividiert man durch die Gesamtzahl der Möglichkeiten, aus N Szenen n_j auszuwählen. Die gesuchte Wahrscheinlichkeit ist also

$$P(X = n_{ij}) = \frac{\binom{n_i}{n_{ij}}\binom{N - n_i}{n_j - n_{ij}}}{\binom{N}{n_j}}. \qquad (2)$$

Dies ist die Wahrscheinlichkeitsfunktion der hypergeometrischen Verteilung. Außer dieser Wahrscheinlichkeit müssen aber auch noch die Wahrscheinlichkeiten für alle extremeren Werte als n_{ij} berücksichtigt werden. Für den Fall, daß der beobachtete Wert n_{ij} unter dem Erwartungswert liegt, müssen also die Wahrscheinlichkeiten für einen so kleinen Wert und alle noch kleineren Werte summiert werden:

$$P(X \leq n_{ij}) = \sum_{x=0}^{n_{ij}} \frac{\binom{n_i}{x}\binom{N - n_i}{n_j - x}}{\binom{N}{n_j}} \qquad (3)$$

Ist der beobachtete Wert n_{ij} dagegen größer als der Erwartungswert, so muß die Summe der Wahrscheinlichkeiten für einen so großen Wert n_{ij} und alle möglichen noch größeren Werte gebildet werden, d.h. die obere Grenze ist die Zahl der Szenen, in denen die Person auftritt, die in weniger Szenen als die andere Person vorkommt. In dem Fall lautet die Formel

$$P(X \geq n_{ij}) = \sum_{x=n_{ij}}^{\min(n_i, n_j)} \frac{\binom{n_i}{x}\binom{N - n_i}{n_j - x}}{\binom{N}{n_j}} \qquad (4)$$

Wie man leicht feststellt, ist (1) nichts anderes als E(X) der hypergeometrischen Verteilung (2).

Die Berechnung soll am Beispiel der Personen Sara und Marwood aus Lessings "Miss Sara Sampson" veranschaulicht werden. Sara hat $n_i = 25$, die Marwood $n_j = 16$ Auftritte. Man würde also (wenn beide Personen unabhängig voneinander wären) - bei insgesamt N = 47 Szenen - erwarten, daß die beiden in

$$E(n_{ij}) = \frac{25 \cdot 16}{47} = 8.51$$

Szenen zusammentreffen. Tatsächlich findet eine Begegnung aber nur in $n_{ij} = 5$ Szenen statt. Die gesuchte Wahrscheinlichkeit ist also

$$P(X \leq 5) = \sum_{x=0}^{5} \frac{\binom{25}{x}\binom{47 - 25}{16 - x}}{\binom{47}{16}}$$

Am einfachsten ermittelt man diese Wahrscheinlichkeit aus vorhandenen Tabellen (vgl. Liebermann/Owen). Wenn nicht vorhanden, so kann man die Rekursionsformel benutzen. Es ist

$$P(X = 0) = \frac{\binom{N - n_i}{n_j}}{\binom{N}{n_j}} \qquad (5)$$

und

$$P(X = x+1) = \frac{(n_i-x)(n_j-x)}{(x+1)(N-n_i-n_j+x+1)} P(X = x). \qquad (6)$$

In unserem Beispiel ergibt sich

$$P(X = 0) = \frac{\binom{47-25}{16}}{\binom{47}{16}} = .0000000496$$

$$P(X = 1) = \frac{(25-0)(16-0)}{1(47-25-16+1)} P(X = 0) = .0000028343$$

$$P(X = 2) = \frac{(25-1)(16-1)}{2(47-25-16+2)} P(X = 1) = .0000637714$$

$$P(X = 3) = \frac{(25-2)(16-2)}{3(47-25-16+3)} P(X = 2) = .0007605333$$

$$P(X = 4) = \frac{(25-3)(16-3)}{4(47-25-16+4)} P(X = 3) = .0054378133$$

$$P(X = 5) = \frac{(25-4)(16-4)}{5(47-25-16+5)} P(X = 4) = .0249150720$$

Als Summe ergibt sich die gesuchte Wahrscheinlichkeit, daß Sara und Marwood in nur 5 oder noch weniger Szenen zusammen treffen, zu $P(X \leqq 5) = .0312$. Bei einem Signifikanzniveau von $\alpha = .05$ wäre damit die Aussage: "Die Begegnung von Sara und Marwood wird vermieden." statistisch gesichert.

Die Ergebnisse dieser Berechnungen für alle Personenpaare aus "Miss Sara Sampson" sind in den folgenden Tabellen I - III zusammengestellt.

Tabelle I. Zahl der Auftritte (n_i) der Personen aus "Miss Sara Sampson"

William	7
Sara	25
Mellefont	25
Marwood	16
Arabella	3
Waitwell	10
Norton	13
Betty	10
Hannah	7
Wirt	1

Tabelle II. Realisierte (n_{ij}) und erwartete ($E(n_{ij})$) Paar-Konfigurationen in "Miss Sara Sampson"

	Wi	Sa	Me	Ma	Ar	Wa	No	Be	Ha	Wt
Wi	—	2	1	0	0	7	1	0	0	1
Sa	3.72	—	13	5	0	5	7	9	0	0
Me	3.72	13.30	—	10	2	1	9	5	4	0
Ma	2.38	8.51	8.51	—	3	0	0	1	7	0
Ar	.45	1.60	1.60	1.02	—	0	0	0	3	0
Wa	1.49	5.32	5.32	3.40	.64	—	2	1	0	1
No	1.94	6.91	6.91	4.43	.83	2.77	—	6	0	0
Be	1.49	5.32	5.32	3.40	.64	2.13	2.77	—	0	0
Ha	1.04	3.72	3.72	2.38	.45	1.49	1.94	1.49	—	0
Wt	.15	.53	.53	.34	.06	.21	.28	.21	.15	—

Tabelle III. Wahrscheinlichkeit der realisierten Paar-Konfigurationen und Differenz zwischen realisierten und erwarteten Paar-Konfigurationen in "Miss Sara Sampson"

	Wi	Sa	Me	Ma	Ar	Wa	No	Be	Ha	Wt
Wi	——	.1580	.0324	.0418	.6093	$2 \cdot 10^{-6}$	.3635	.1637	.2964	.1489
Sa	-1.72	——	.5472	.0312	.0950	.5493	.6076	.0093	.0027	.4681
Me	-2.72	- .30	——	.2717	.5489	.0025	.1502	.5493	.5745	.4681
Ma	-2.38	-3.51	1.49	——	.0345	.0086	.0015	.0709	.0002	.6596
Ar	- .45	-1.60	.40	1.98	——	.4792	.3690	.4792	.0022	.9362
Wa	5.51	- .32	-4.32	-3.40	- .64	——	.4305	.3075	.1637	.2128
No	- .94	.09	2.09	-4.43	- .83	- .77	——	.0175	.0855	.7234
Be	-1.49	3.68	- .32	-2.40	- .64	-1.13	3.23	——	.1637	.7872
Ha	-1.04	-3.72	.28	4.62	2.55	-1.49	-1.94	-1.49	——	.8511
Wt	.85	- .53	- .53	- .34	- .06	.79	- .28	- .21	- .15	——

In Tabelle IV sind die Personen-Paare nach der Größe der Wahrscheinlichkeit geordnet, und zwar so, daß auf den ersten Platz das Paar mit der kleinsten Wahrschienlichkeit kommt, für das gilt: $n_{ij} > E(n_{ij})$, auf den zweiten Platz das Paar mit der zweitkleinsten Wahrscheinlichkeit und $n_{ij} > E(n_{ij})$ etc. Entsprechend steht an letzter Stelle das Paar mit der kleinsten Wahrscheinlichkeit und $n_{ij} < E(n_{ij})$ etc.
Wählt man als Signifikanzniveau $\alpha = .05$, so erhält man 6 Personen-Paare für die gilt: $n_{ij} > E(n_{ij})$ und 7 Paare mit $n_{ij} < E(n_{ij})$.
Eine Hilfe bei der Beschreibung und Interpretation der Beziehungen zwischen den Personen kann auch die graphische Darstellung der signifikanten Werte aus Tabelle III bzw. IV sein.
In Abbildung I verbindet eine durchgezogene Linie die Personen-Paare, die signifikant häufiger als erwartet zusammenkommen ($n_{ij} > E(n_{ij})$), eine gestrichelte Linie die Paare, die signifikant seltener zusammen auftreten als erwartet ($n_{ij} < E(n_{ij})$).

Tabelle IV. Rangfolge der Personen-Paare nach der Größe der Wahrscheinlichkeit p_{ij} in "Miss Sara Sampson" (Ab Rang 14 ist $n_{ij} < E(n_{ij})$).

Rang	Personen-Paar	p_{ij}	Rang	Personen-Paar	p_{ij}	Rang	Personen-Paar	p_{ij}
1	Wi-Wa	$2 \cdot 10^{-6}$	16	Be-Wt	.7872	31	Wi-Ha	.2964
2	Ma-Ha	.0002	17	No-Wt	.7234	33	Wi-Be	.1637
3	Ar-Ha	.0022	18	Ma-Wt	.6596	33	Wa-Ha	.1637
4	Sa-Be	.0093	19	Wi-Ar	.6093	33	Be-Ha	.1637
5	No-Be	.0175	20.5	Sa-Wa	.5493	35	Wi-Sa	.1580
6	Ma-Ar	.0345	20.5	Me-Be	.5493	36	Sa-Ar	.0950
7	Wi-Wt	.1489	22	Sa-Me	.5472	37	No-Ha	.0855
8	Me-No	.1502	23.5	Ar-Wa	.4792	38	Ma-Be	.0709
9	Wa-Wt	.2128	23.5	Ar-Be	.4792	39	Wi-Ma	.0418
10	Me-Ma	.2717	25.5	Sa-Wt	.4681	40	Wi-Me	.0324
11	Me-Ar	.5489	25.5	Me-Wt	.4681	41	Sa-Ma	.0312
12	Me-Ha	.5745	27	Wa-No	.4305	42	Ma-Wa	.0086
13	Sa-No	.6076	28	Ar-No	.3690	43	Sa-Ha	.0027
14	Ar-Wt	.9362	29	Wi-No	.3635	44	Me-Wa	.0025
15	Ha-Wt	.8511	30	Wa-Be	.3075	45	Ma-No	.0015

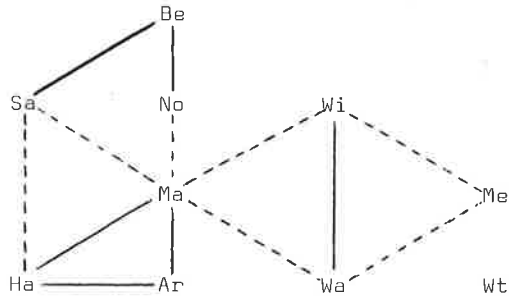


Abbildung I. Graph der Personen-Paare, die signifikant häufiger (——) bzw. seltener (-----) als erwartet auftreten

Bei der Festsetzung der Signifikanzgrenzen wurde davon ausgegangen, daß die Wahrscheinlichkeiten p_{ij} (Tabelle III und IV) für jedes Personen-Paar getrennt berechnet sind. Das so erhaltene Wahrscheinlichkeitsmaß ist symmetrisch bezüglich i und j, d.h. die Werte n_i und n_j in den Formeln (1) bis (6) sind vertauschbar. Eine Berechnung der bedingten Wahrscheinlichkeiten über eine verallgemeinerte hypergeometrische Wahrscheinlichkeintsverteilung wäre denkbar, hätte aber (neben einem äußerst großen Rechenaufwand) den Nachteil, daß die resultierende hypergeometrische Formel nicht mehr symmetrisch wäre.

Betrachtet man im vorliegenden Fall die Berechnungen der einzelnen Wahrscheinlichkeiten p_{ij} als simultan durchgeführte Tests, so kann man eine Korrektur des Signifikanzniveaus durchführen. Diese Alphaadjustierung (KRAUTH/LIENERT 1973: 40 ff.; SACHS 1978: 93 f.) ergäbe für r simultane Tests und $\alpha = .05$

$$\alpha' \leq \frac{\alpha}{r} \qquad \text{und}$$

in unserem Fall bei r = 45 möglichen Paaren

$$\alpha' = \frac{.05}{45} = .0011 .$$

Damit würde sich ein signifikanter Wert nur noch für zwei Paare ergeben (Tabelle IV, Rang 1 und 2), so daß wir vom heuristischen Gesichtspunkt aus eine Alphaadjustierung nicht vorschlagen.

Literatur

DINU, M., Structures linguistiques probabilistes dans l'étude du théâtre. Cahiers de linguistique théorique et appliquée 5, 1968, 29-46.

DINU, M., How to estimate the weight of stage relations. Poetics 6, 1977, 209-227.

KRAUTH, J., LIENERT, G.A., Die Konfigurationsfrequenzanalyse (KFA). Freiburg, Alber 1973.

LIEBERMAN, G.J., OWEN, D.B., Tables of the hypergeometric probability distribution. Stanford, Stanford Univ. Press 1961.

MARCUS, S., Mathematische Methoden im Theaterstudium. In: MARCUS, S., Mathematische Poetik. Frankfurt/Main, Athenäum 1973, 287-370.

SACHS, L., Angewandte Statistik. Berlin, Springer 1978.

Associative Meaning and its Analysis[+]

Gerhard Strube, Munich

Since the late fifties, when psychology of language became 'psycholinguistics', the field has been dominated by structural (i.e. largely syntactical) analysis. Behaviorism, once favoured by American linguists like Bloomfield, seemed to be stuck up the dead end of a once promising (though, we were told, principally wrong) road. 'Verbal Behavior' (SKINNER 1957) is one of the famous books which scarcely anyone knows more than the title about.

Recent years witnessed a revival if not in behaviorism itself, but in its quest for more than purely formal analysis of language per se, in short: for pragmatics. In Germany, for instance, Skinnerian 'tacts' and 'mands' are under investigation, re-titled 'Objektbenennung' (naming of objects) and 'Handlungsaufforderungen' (orders and requests), with behaviorism never mentioned in these reports (cf. HERRMANN & DEUTSCH 1977; LAUCHT & HERRMANN 1978). Instead cognitively oriented 'theories of action' have been stressed in these and related studies.

The present situation, then, tempts one to go a step further and look for new developments of the half-forgotten behavioristic approach to language. We shall study semantics rather than pragmatics, and review the concept of associative meaning, as well as present modern techniques of nonmetric multidimensional analysis for the study of associatively related words. Finally, illustration and corroboration will be given by means of experimental data.

[+] Much of this work is a revised version of ch. iii of my doctoral dissertation (STRUBE 1977a).

Part 1: THEORY AND EARLIER INVESTIGATIONS

'Association' could well be called the central concept of behavioristic theories. Even Pavlovian theorizing relies heavily on the 'incontestably oldest achievement of psychology' (PAWLOW 1973: 159). Indeed 'association', treated by as early a philosopher as Aristotle, is one of the oldest psychological terms, extensively discussed by British empiricists in the 18th and 19th centuries (cf. ch.2 in ANDERSON & BOWER 1973).

About 1880, association theory gained an experimental foothold through Galton's invention of the word-association task (GALTON 1880; 1883), and subsequent improvements thereof by Trautscholdt and J. McKeen Cattell. Word association, then, stands as one of the first 'mental tests', which clearly implies that it pertains to psychometrics rather than learning theory (STRUBE 1977 b). This is confirmed by early applications of word association, which lie mainly in the clinical domain (e.g., JUNG & RIKLIN 1904; KENT & ROSANOFF 1910; even HULL & LUGOFF 1921).

Thus behaviorism inherited not the task, but the concept of association, and put this to a rather restricted use. Learning theorists like Thorndike or Pavlov, as well as Watson and other exponents of behaviorism proper, stressed but one of the many 'laws of association', making contiguity the sole principle of association.

Experimental instigation of associations by contiguity is even older, and nowhere better exemplified than in the famous series of experiments done by EBBINGHAUS (1885). Here, associations between nonsense syllables are the outcome of experimental treatment, while Galton's association task intends to show the 'natural', pre-experimental associations already established in the subject.

DEESE (1962) has pointed out the two traditions, and noted that "recently, these two traditions have been combined" (op.cit.: 161). Examples of such merging of traditions are the 1952 study by JENKINS & RUSSELL and the study by JENKINS et al. (1958), assessing effects on S-R and R-S sequences in free recall due to pre-experimental word associations (calibrated by the Kent-Rosanoff lists).

Use of the association task, naturally, presented behaviorists with the problem of explaining word association in terms of conditioning theory. Of course, a simple S-R model will explain nothing. So-called 'mediation theory' has been presented in order to give a learning-theory notion of 'meaning' by relating verbal behavior to non-verbal behavior, and so to subsequently explain the process of word association. To show what mediation theory is like, we turn to the most illustrative discussions at the conference on 'Verbal Learning and Verbal Behavior' (autumn 1959; report: COFER 1961). Both Bousfield and Osgood, well-known mediation theorists, presented their respective versions of the theory. Moreover Osgood at that time had not restricted his form of mediation theory to apply to the semantic differential only (cf. OSGOOD 1962).

BOUSFIELD (1961: 81 ff.), in the opening of his lecture "The problem of meaning in verbal learning", shows uneasiness concerning the confusing, though common usage of 'meaning'. He even shuns the term 'mediation theory', being not sure "whether the conventional mediation paradigm may clearly be applied, and it does not apply to a simple first-order conditioning" (p. 82). He therefore talks of a 'representational response' when referring to the meaning of a word. Learning of representational responses is either direct (i.e., through first-order conditioning) or indirect, 'mediated' (second-order conditioning). Bousfield's first example concerns direct learning of the meaning of *bad*:

"Suppose a naive and naughty child hears the word *bad* and is spanked. The US is a pain stimulus which elicits its conditionable representational response. The CS is the word *bad* which elicits its representational sequence. This includes the representational response which involves the child's saying *bad* either implicitly or out loud. Its consequent representational stimulus is what becomes conditioned to elicit the representational response to the pain stimulus"(op.cit.: 82, see fig. 1a).

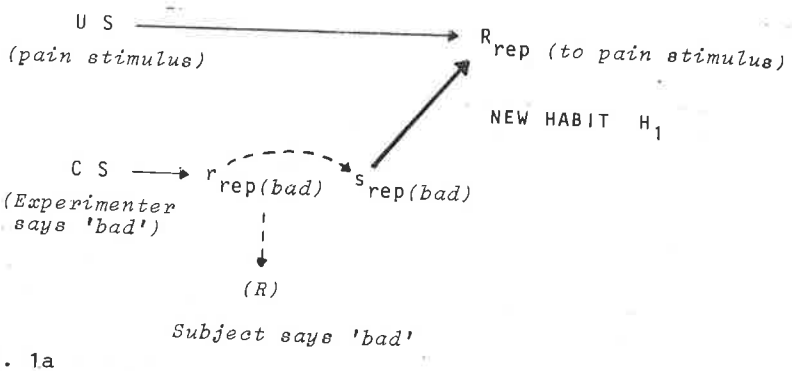


Fig. 1a

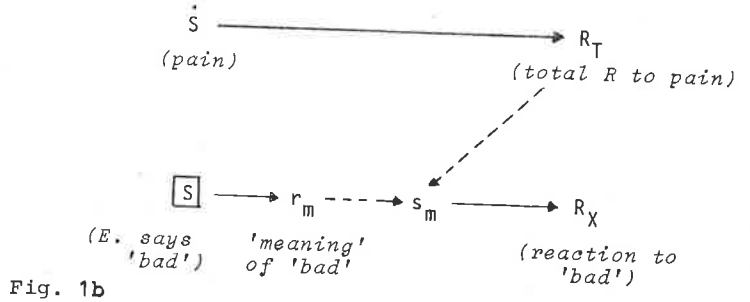


Fig. 1b

Fig. 1. First-order conditioning of *bad* (a) after BOUSFIELD (1961: 83) (b) after OSGOOD (1961: 93)

This view contains two weak points which are genuine to traditional behaviorism. The first is the questionable if not utterly falsified notion of *saying* the word. Though proposed by Watson and maintained as recently as Deese, as we shall see later, "there is not one shred of positive evidence in that whole literature", OSGOOD (1961: 106) comments. Even more serious is Bousfield's erroneous identification of conditioned and unconditioned response, which merges them into one 'R_{rep}'. Surely the boy won't cry in response to *bad*, as he is likely to do when spanked!

Osgood, then, is careful to evade both these fallacies. His model of 'primary sign learning' (fig. 1b) discriminates conditioned and unconditioned responses:

"... the representational process is assumed to be some reduced portion of the total behavior made to the thing signified; under the condition of repeated pairing of the sign stimulus (the word "bad") with the significate stimulus (painful spanking), the sign stimulus comes to elicit those most readily conditionable and least interfering components of the total behavior to the significate (e.g. autonomic anxiety reactions, avoidant postural sets, and the like). It is precisely because this representational process *is* part of the actual behavior made to the thing signified that the sign has its discriminative semantic or symbolic capacity. The important thing to note about this conception is that it specifies that the form of the representational response is not the same as the overt vocal production of the word". (OSGOOD 1961: 92f.)

Avoiding the pitfalls of Bousfield's behaviorism obviously costs Osgood a good deal of precision: What exactly *is* meaning, defined as an r_m, a 'mediating response' rather loosely connected with the observable reaction R_x? Osgood tends to equate r_m with emotional reactions and with the well-known 'affective meaning system' of his semantic differential (cf. OSGOOD 1962). Even 'affective meaning', however, refers to verbal behavior, of which semantic differentiation gives but one possible measurement.

We may specify our task now as defining the meaning of a word such that it be largely independent of specific reactions to the object being referred to. Thus, confusion of conditioned with unconditioned response will be avoided. On the other hand, the concept of meaning should retain as much concreteness of content as possible without confining 'meaning' to the latent or overt uttering of the word. - Looking for a solution, we turn to meaning as defined by word association techniques. Noble and Deese have pioneered in this field.

NOBLE (1952) presents us with the less sophisticated variant of associative meaning. For him, the meaning of a word acting as stimulus for association is directly defined by the variety of responses it elicits. Since the responses made in an association task may be thought of as organized in a Hullian habit hierarchy, the most frequent being the responses learnt best, association norm tables (those by Kent and Rosanoff, for instance) could well claim to give an operational definition of the stimulus word's meaning. It follows that the copiousness of meaning, i.e. the 'meaningfulness' of a word, directly depends on the number of different associative responses it elicits. Thus, meaningfulness is defined as

$$m = \frac{1}{N} \Sigma R_s,$$

with R_s as a response to the Stimulus S, and N as the number of subjects. The time interval for association is usually set at one minute. Meaningfulness, in the course of years, has proved to be of fundamental importance in verbal learning (cf. COFER 1969: 323 f.).

DEESE (1962) takes up the idea of associative meaning in a broader sense, concluding "that the way to discover the associative meaning of words is not to classify associations according to other senses of meaning but to discover the relationships associations have with one another" (p. 162).

Before we turn to Deese's experimental work, however, a theoretical link has to be provided between 'association' as the foremost principle in learning theory and the concept of associative meaning. Since word association has been developed as a psychometric device, it has not been accounted for by behavioristic learning theory.

How can the principle of association explain word association and consequentially account for associative meaning? Our point of departure shall be Bousfield's model of 'second-order conditioning' of meanings. Bousfield resumes his previous example, i.e. learning the meaning of *bad*, by demonstrating a possible way of learning the meaning of *evil* (cf. Fig. 2):

"At a later time our child hears the word *evil* followed by the word *bad*. Perhaps he asks his parents the meaning of *evil*, and he is told, '*Evil* means *bad*'. The word *evil*, which is the CS, elicits its representational sequence. Now, however, the word *bad*, functioning as the US, not only elicits its representational sequence but also the earlier established H 1, so that the representational response to the pain stimulus also occurs. As a result of this learning, we should expect that when the child subsequently hears the word *evil*, he should say *bad* either implicitly or out loud." (BOUSFIELD 1961: 83)

Though we may note that the criticism directed against this learning model for *bad* applies as well to the new example, Bousfield's theory may nevertheless be useful to explain the R-S sequence in word association. Figure 2 shows the graphic representation of the model, with the essential path from *evil* (as a stimulus) to *bad* (as a response) in bold lines.

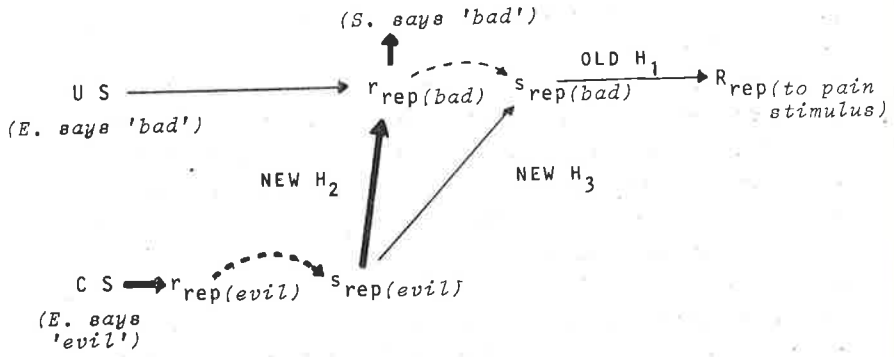


Fig. 2. Second-order conditioning of meaning (after BOUSFIELD 1961: 83)

Osgood has improved on this conceptualization. Though not contained in his reply to Bousfield (OSGOOD 1961), the crucial amplification can be found in his earlier book on the semantic differential (OSGOOD et al. 1957). His model, contrary to Bousfield's, is not restricted to sequences (like *evil* to *bad*) having actually occurred in learning, but may account for the far more numerous cases when /S/, the later stimulus word, has been presented in a context that merely contains relevant associatives. The word /S/, then, gains its meaning through the meaning of other words, S_1 to S_n, whereby "this new stimulus pattern ..., /S/, acquires portions of the mediating reactions already associated with the primary signs" (OSGOOD et al. 1957: 8). The result of this learning process is shown in Fig. 3 a. It may further be assumed that the partial meanings (symbolized by dotted lines) differ in associative habit strength according to the frequency of the connections made when a word has been explained (presumably this frequency is correlated with relevance or, as found for categories by ROSCH, 1973, 'typicality').

From here it is but a small step to explain performance in a word-association task. It is assumed that after r_{m_x} has acquired parts of r_{m_1} to r_{m_n}, in turn s_{m_x} elicits the mediated responses s_{m_1} to s_{m_n} (in an order largely dependent upon habit-strength). This assumption moreover conforms to the well-known asymmetry of word-association sequences (a frequent associative of a stimulus word, when presented as a stimulus itself, may not elicit the former stimulus word as its associative). The whole process of word association may be schematized as in Fig. 3 b.

This is the general model for learning the meaning of words. It fits word association as well as the semantic differential (for which it was designed), or even ordinary language use. Associative meaning, then, is not something vaguely connected with other types of meaning, but is the only concept of meaning conceivable in a behavioristic spirit.

Operational definition, however, of associative meaning is not quite as easy as it would seem. Even long and repeated trials at the word-association task cannot guarantee completeness of a word's associatives. DEESE (1962) therefore suggests

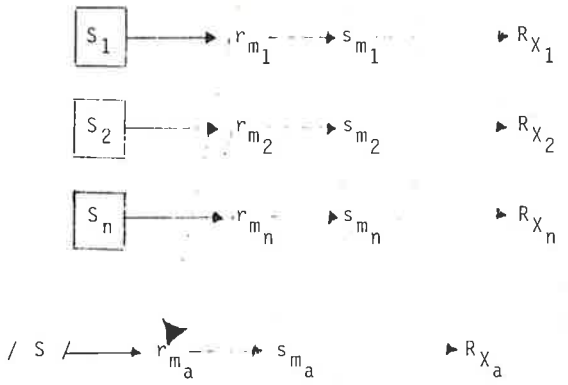


Fig. 3 a

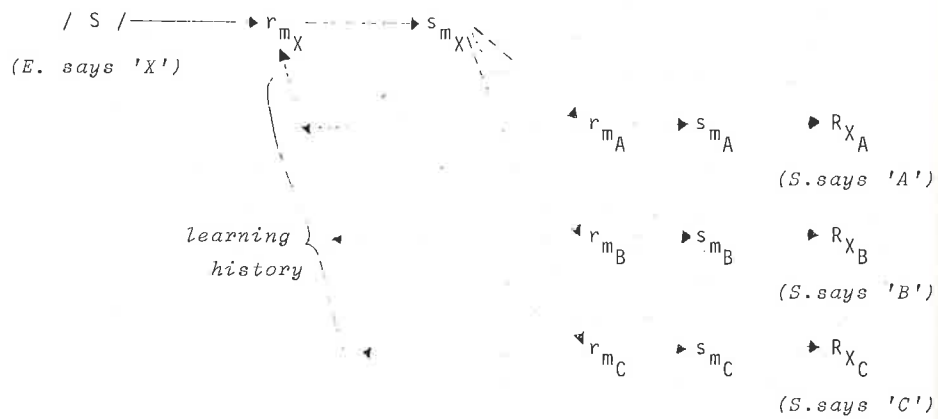


Fig. 3 b

Fig. 3. The learning of meaning and the word-association task
(a) Results of learning: r_m has acquired parts of r_{m_1} to r_{m_n}
(after OSGOOD et al. 1957: 7)
(b) Model of performance in a word-association task (The tiny dotted lines refer to previous learning history)

assessing the associative 'overlap' between two stimulus words, thus gaining better insight into a word's associative meaning by means of another word's associatives. He gives an example, taking "piano" and "symphony": "In a sample of Johns Hopkins University undergraduates, neither one elicited the other, but both elicited Note, Song, Sound, Noise, Music, and Orchestra in varying frequencies." (op. cit.: 163).

Deese's method provides a tool for simultaneous analysis of word fields; although it is not a tool to discover new word fields, such fields as are recognized by linguists may be put to empirical testing.

'On the Structure of Associative Meaning' is the title of DEESE's 1962 pioneering work, to be summarized and appraised below. In this paper, the basic ideas of associative semantic analysis are outlined. Concerning its technique and results, however, it is handicapped by the use of statistical methods hardly suited to the task, more appropriate methods not having been available at that time.

The word field taken up by Deese is that of "butterfly" and its eighteen most common associates as determined by the norms of RUSSELL & JENKINS (1954). These 19 words served as stimuli in a free association task, limited to one response per stimulus; 50 Ss participated in the experiment.

Discussing the measurement of associative overlap between each two of the 19 response distributions, Deese rejects Pearson's correlation coefficient. The Pearson correlation, it is to be understood, is sensitive only to the shape of the distribution but not to the portion of common associative responses which is, by definition, the associative overlap. This property of the usual correlation coefficient may even produce gross errors:

"Often, responses that are high in frequency to one stimulus will be low in frequency to another. This happens very often because of the typically steep rank-frequency distribution of free association responses to any given stimulus. Because the forms of the distribution of *common* responses may

be inverse to one another, the obtained product-moment correlations will often be negative." (op. cit.: 165)

Instead of correlation, Deese proposes his own formula, an 'overlap coefficient' (OC) which, shortening Deese's exhaustive treatment, can be given as

$$OC_{ij} = \frac{1}{n} \Sigma f_{ij},$$

where i and j are two stimuli, Σf_{ij} the number of common responses, and n the total number of responses to i or j.

A technical problem arises with 'direct' responses, i.e. when one or both of the stimulus words is an associative of the other. Like JENKINS & COFER (1957) and BOUSFIELD et al. (1958), Deese solves the problem by assuming that every stimulus elicits itself as a covert response (which reminds us of Bousfield's theory). The OC thus is a standardized measure, ranging from zero (i.e. no common responses) to between 0.5 and 1 (in the case of identical response distributions, depending on the extent of direct responses). Apart from this 'bias' in favour of direct responses, Deese's OC has the shortcoming of assuming equal n for both response distributions.

The OCs are then arranged in a matrix analogous to a matrix of correlation coefficients - which means that the matrix is symmetric, contains the value 1 in each cell of the main diagonal, and may be factorized. This is exactly what Deese does. Unfortunately he only superficially informs the reader of the specific kind of method used. There is only one sentence, stating that "these relative frequencies were factored by the centroid method, and after a very few rotations (pairs of factors were rotated through approximately 45 degrees) the factor loadings ... were obtained." (op. cit.: 168).

About other parameters employed, e.g. the criterion used to determine the number of factors, nothing is said. Yet careful inspection of the matrix of rotated factor loadings (Table 1) yields some important conclusions:

1. The first and greatest factor has an eigenvalue of 1.91 (sum of squares in the first column), which means that it accounts for roughly 10% of total variance at most (if it is an orthogonal solution, which Deese says nothing about). This is a fair, if moderate proportion of variance explained by so large a number of factors, relative to the small number of variables.

Table 1 Factor loadings (rotated) of stimulus words (from DEESE 1962: 169)

WORDS	FACTORS					
	I	II	III	IV	V	VI
moth	.44	.03	-.27	-.01	-.03	-.32
insect	.50	.01	-.33	.01	-.34	.11
wing	.52	.01	.45	.01	.29	-.07
bird	.52	.02	.46	.01	.29	-.07
fly	.48	.03	.32	.01	-.28	-.03
yellow	.01	.44	-.03	.34	-.32	-.02
flower	.01	.39	-.03	-.32	.03	.44
bug	.41	.01	-.34	.00	-.14	.37
cocoon	.40	.01	-.35	.00	.25	.02
color	-.02	.42	-.04	.44	.04	-.04
blue	-.02	.57	-.04	.52	.23	-.04
bees	.36	.04	.34	-.02	-.30	.00
summer	-.01	.31	-.03	-.34	-.02	-.34
sunshine	.02	.27	-.04	-.33	-.03	-.35
garden	.00	.35	-.02	-.34	-.03	.44
sky	-.01	.41	-.03	.43	.38	-.07
nature	.04	.31	.29	-.02	.01	.34
spring	-.01	.35	-.03	-.37	-.02	-.36
butterfly	.48	.06	-.29	-.01	.26	.01

2. The pattern of factor loadings is of "convenient simple structure" (in Thurstonian sense; op. cit.: 168); it does not, however, conform to the kind of hierarchical organization Deese tries to establish in his interpretation. If this type of factor interpretation is at all appropriate, it certainly cannot be applied to the factors V and VI. The first four factors yield something like the following classification:

A animals (positive loadings in factor I)
- A a Wing, Bird, Fly, Bees, and Nature (all positive in III)
- A b Butterfly, Cocoon, Insect, Moth, Bug (negative in III)

B other words (positive loadings in factor II)
- B a Yellow, Blue, Sky, Color (all positive in factor IV)
- B b Garden, Flower, Spring, Summer, Sunshine (all negative in IV)

Both the last factors, V and VI, do not fit into this general pattern, as the reader may readily see for himself. (Maybe that is why Deese does not mention these factors in his discussion.)

Though a quick classification into four categories is obtained, there remains the problem of deciding what the categories *mean*. A number of questions arise, such as

- Why is "nature" grouped with the animals (and not with "sunshine" and "summer")?
- Why are "fly" and "bees" grouped with "bird" instead of being grouped with "butterfly" and the rest of the "insects" which they surely are?

Under close inspection, an interpretation in terms of logical, hierarchical classification seems to melt away. More factors than needed account for a less than satisfying portion of total variance, and the strategy proposed for interpretation has failed.

An even more fundamental issue hitherto undiscussed concerns whether factor analysis can justifiably be applied to

overlap coefficients. Indeed, the answer must be negative, because OCs, being based on common frequencies, are not the kind of multiplicative distance (or similarity) measure which factor analysis has been designed for. This, of course, is not Deese's fault: what else should he have done, without other techniques of dimensional analysis available? But these difficulties (and the ensueing interpretational ones) may have brought an untimely end to this kind of work; anyway, Deese later joined forces with psycholinguists influenced by Chomsky. Other, so-called structural (i.e. thoroughly non-associationistic) theories of word association have been proposed (CLARK 1970) and largely been found to fail (e.g. LEUNINGER et al. 1972). The renewed interest in associationism, however, demands that Deese's illuminating ideas on semantic analysis be taken up and furnished with modern techniques of statistical analysis (e.g., SZALAY & DEESE 1978).

Part II: HOW TO ANALYSE ASSOCIATIVE MEANING

There are certain requirements to be met by any method in order to provide a practical as well as theoretically sufficient means for analyzing associative meaning.

(1) In a homogeneous population, response probabilities with respect to one subject only, assumed to indicate habit strength of the responses, may be approximated by response frequencies taken across subjects (one response per subject). Average response frequencies thus become an index of habit strength in an idealized, or indeed any member of the population. Yet a far broader data base would be provided by recording more than one response per subject. Therefore a generalized coefficient of associative overlap has to be designed.

(2) Factor analysis has been shown to be an inadequate method for the analysis of matrices composed of overlap coefficients. Since the main problem concerns the assumption of linearity (i.e. that the OC value be proportional to geometric distance in the structural model), this assumption is

to be "reduced" (sensu GIGERENZER & STRUBE 1978) to assuming ordinality. A nonmetric technique for multidimensional scaling is therefore required.

(3) Results of structural analysis (e.g., a factor structure) are usually interpreted in terms of dimensions; a meaningful geometric structure, however, does not necessarily imply the meaningfulness of the coordinates (see DEGERMAN 1972). Rather, an interpretation should be guided by the structure itself, by the content of the associative responses, and by external criterion variables.

(4) A purely descriptive analysis (like Deese's) is not sufficient, because techniques for dimensional analysis are designed to reproduce the data matrix with minimal aberrations. To ensure that the structural solution obtained through multidimensional analysis is in fact an associative structure, validation through further experiments is needed.

Generalizing the Overlap Coefficient

How are these requirements met? The first, i.e. generalization of the overlap coefficient, has been solved through the work of MARX (1976 a, b). He differs from Deese in basing his coefficient (named ÜK = Überlappungskoeffizient, the German word for overlap coefficient) on relative instead of absolute frequencies of common responses. His ÜK is just the sum of common response percentages. For illustration consider the case of stimulus A, eliciting a total of 30 responses (12 subjects, each associating for 30 sec), three of which, i.e. 10%, consist of the word X. Another stimulus word, B, elicits a total of 35 responses, seven of which are X, equalling 20 percent. The ÜK value, so far, is ten percent, i.e. the smaller of the two percentages. Let the relative frequencies for Y, another response common to both A and B, be .8 and .11, respectively. This adds another 8 percent to the ÜK, now amounting to .18. This procedure is continued until all common responses are dealt with.

Direct associations are treated in the following way: If B occurs twice as response to A, the percentage 2/30 = .067 is directly added to the value of the ÜK. This procedure eliminates the bias inherent in Deese's OC. (For an extensive discussion, see MARX 1976 a.) MARX (1976 b) provides critical values to test the significance of ÜK.

Nonmetric Multi-Dimensional Scaling of Associative Similarities

An overlap coefficient like the ÜK is merely an indicator of similarity of associative meaning, i.e., we assume an ordinal relationship between the ÜK value and the associative similarity of two stimulus words. Representing similarity by distance between points in a geometric space, the rank order of point distances given through the ÜKs implies an 'ordered metric' (COOMBS 1964) of the stimuli themselves. (Note that no assumption of linearity is needed).

In a similar way, the structural model employed to frame the dimensional analysis need not be of Euclidean metric. Factor analysis, relying upon Euclidean models, has brought about the customary view that structures of attitudes, values, judgements, and even personality, are adequately represented in a Cartesianized Euclidean space (BOYD 1972). Studies as early as ATTNEAVE (1950), however, and even more those by HYMAN & WELL (1967; 1968) and WENDER (1969) have made metric itself a variable relevant to the psychology of judgment. So it seems wise to dispense with any specific assumption of metric and use a technique which results in an appropriate metric rather than proceeds on the assumption of a questionable one.

Suitable techniques for the structural analysis of overlap matrices have been in existence since SHEPARD's basic work of 1962. Important improvements are due to KRUSKAL (1964 a, b). A similar approach, though differing in detail, is the mallest Space Analysis program series by GUTTMAN & LINGOES (s. GUTTMAN 1968; LINGOES 1972). Important readings have been collected in SHEPARD et al. (1972). In spite of various names being in use, we shall call the whole group of techniques

NMDS, i.e. Nonmetric Multi-Dimensional Scaling.

In the field of language, psychologists have not made much use of NMDS (the work of RAPOPORT & FILLENBAUM, 1972, provides an exceptional example), yet NMDS seems uniquely suited to the analysis of associative structures.

Description NMDS will be restricted to the basic principles and some of the specifities of POLYCON-2, a computer program by YOUNG (1972) expanding the scope of Kruskal's procedure. (This is the program we used in our experimental work.) Overlap coefficients form a quadratic similarities matrix (COOMBS 1964: 'quadrant-IV data'), and no other features of POLYCON-2 will be described.

The basic model underlying NMDS is represented by the equation

$$S \overset{m}{=} \hat{D} \hat{=} D = h(X).$$

S (similarities) is the matrix of overlap coefficients. It is transformed into $\hat{D}$ (disparities), a matrix being in strictly monotonic relation to S. This is the 'data side' of the equation. On the right side, the 'model side', there is the resulting matrix X, containing the Cartesian coordinates of points representing the stimulus words in a space of optional though usually minimal dimensionality. Upon this space, a freely defined metric may be imposed, h(X), yielding the matrix of distances, D, in an unequivocal way.

The most relevant connection is the one between $\hat{D}$ and D, i.e. the disparities and the distance matrix, both of which are alike in dimensionality. The aim of NMDS is to provide the model with an optimal fit (designated by $\hat{=}$) to the empirical data. Kruskal and Young use a least-squares approximation.

The rationale of the technique may summarized as follows: A matrix of similarities (e.g., overlap coefficients) is monotonically transformed into a matrix of disparities. In a geometric space of previously chosen dimensionality and metric, a configuration of points (e.g., representing stimulus words) is established such that the matrix of point distances becomes

a least-squares approximation to the disparities matrix.

To do so, however, is not quite as easy as it sounds. For instance, the disparities matrix $\hat{D}$ has to be determined such that a least-squares approximation is possible. The approximation process itself is an iterative algorithm, time-consuming even with the most modern computers. The process starts with an arbitrary initial configuration, which is changed stepwise according to the partial derivatives of the least-squares function L^2 (with respect to the metric of the space), until L^2 is minimized. Though Young employs a better algorithm than Kruskal, 20 to 30 iterations may be required.

The least-squares deviation L^2 is given by

$$L^2 = \frac{\sum_{i,j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i,j}(\hat{d}_{ij} - \hat{d})^2},$$

where d_{ij} denotes the distance between points i and j, $\hat{d}_{ij}$ the corresponding disparity, while $\hat{d}$ is the mean disparity (for details, see YOUNG, op. cit.: 79 -101).

Another deviation measure is 'stress', which has been introduced by Kruskal, and is defined ('stress-2') as

$$S^2 = \frac{\sum_{i,j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i,j}(d_{ij} - d)^2}.$$

Instead of the disparities sum of squares, a term denoting distances-SS enters the denominator, for technical reasons.

A problem commonly encountered with NMDS is that of a stress function having more than one minimum. The iteration process may then end up in a local minimum, possibly far from the absolute one. Young reduces the danger of reaching a merely local minimum by establishing a 'best' initial con-

figuration through metric multidimensional scaling (TORGERSON 1958: 247 - 293), before NMDS proper begins.

Finally, metric and dimensionality have to be discussed, these two being parameters of the model to be chosen freely. Neither Youngs's POLYCON technique nor any other available at the time being, is capable of optimizing both or at least one of these parameters. The only way out is in computing a multitude of NMDS solutions with various dimensionality and metric, looking which one gives best results.

Metric, of course, will be restricted to a class of possible metrics adequate to the problem posed. For analyzing overlap coefficients, the class of Minkowskian r metric is suggested, which gives the inter-point distance d_{ij} in a space of k dimensions by

$$d_{ij} = \left(\sum_{k} \left(|x_{ik} - x_{jk}|\right)^{r}\right)^{\frac{1}{r}} .$$

This results in Euclidean metric for r = 2, or city-block metric for r = 1. The larger the r, the greater is the contribution of the largest difference $|x_{ik} - x_{jk}|$ to the distance as a whole. For r growing infinitely large, d_{ij} becomes equal to $\max(|x_{ik} - x_{jk}|)$. The Minkowski r resulting in a NMDS solution with minimal stress will be selected and considered an adequate model for subjective judgment metric. (GIGERENZER, 1978, and WENDER, 1969, discuss the quality of metric as a model of human judgment.)

Considering dimensionality, looking for minimal stress seems less unequivocal: Provided the number of dimensions is large and not limited, further increase in the number of dimensions will result in even smaller stress, but this, of course, is a trivial solution to the problem. A structure of three or even less dimensions seems preferable, since NMDS aims at structures to be easily interpreted without reference to coordinates.

In our experiment (see below), NMDS solutions of three, two, and one dimension were obtained, with three Minkowski parameters each (r = 1, 2, and 8 for approximating $r \longrightarrow \infty$). The solutions were compared in order to find a best solution, yielding readily to visual inspection and a meaningful psychological interpretation.

Discussion of our requirements (3) and (4), i.e. interpretation and validation of NMDS solutions, will be postponed until the experimental data have been presented.

Part III: THE TECHNIQUE EXEMPLIFIED

Stimulus material: A dozen nouns referring to housing conditions were chosen and presented in the following randomized order:

APP Appartement (apartment. Note that the German word chiefly refers to small, one-room apartments, the general German term for apartment being "Wohnung"),

SOZ Sozialwohnung (An apartment built with government or other public subvention, mainly for people of low income),

WGM Wohngemeinschaft (communal living in a house or an apartment commonly shared by a small group, often students),

NEU Neubau (newly erected building),

BUN Bungalow (bungalow),

HHS Hochhaus (a building usually containing six to twenty or even more storeys),

ALT Altbau (a building erected before World War II),

UNT Untermiete (subletting, renting rooms),

BAU Bauernhaus (farmhouse),

STH Studentenheim (students' dormitory),

RHS Reihenhaus (row-house, semi-detached),

WBL Wohnblock (block of flats, usually a large building).

Method: 79 students (beginning students of psychology) were shown the above stimuli, 25 seconds each, on an overhead projector. During exposure, Ss wrote down their responses on seperate sheets of paper per stimulus. They had been carefully

instructed to respond to the given stimulus (and not to their last-written response).

Further measurements consisted of a semantic differential, by means of which the same 12 concepts were judged (presented in a randomized but different order), and of a paired-comparison judgement of the same concepts as to personal preference in living accomodations ('I would better like to live in a ... than in a ...').

Results of word association. The stimulus words elicited from 167 (WBL) to 317 (BAU) responses, i.e. between 2.11 and 4.01 responses per subject, the mean being 3.11 responses, and the standard deviation being less than 0.46. Noble's meaningfulness value m (gauged to a time-span of 60 sec instead of the 25 sec used here) should amount to about twice as much, the increase of total responses in time being not linear but negatively accelerated, according to the laws of the association process.

The responses to each stimulus word were listed alphabetically, and their relative frequencies were computed. The method of Marx, described above, was used to obtain "ÜK" overlap coefficients for all pairs of stimuli. The overlap matrix is shown in Table 2.

Young's POLYCON-2 procedure was applied to these data nine times, yielding NMDS solutions with varying dimensions (1, 2, and 3) and Minkowski metrics (r=1,2,8). Euclidean metric showed less stress than both the other metrics. The one-dimensional solutions showed too much stress and were not considered further. Stress was not considerably lower for three than for two dimensions (Stress-2 = 0.29), so the two-dimensional Euclidean solution was chosen for further inspection. This solution (columns x and y of Table 3 contain the coordinates) may be judged a 'fair' one (cf. KRUSKAL 1964 a).

Since the coordinates were not quite independent of one another (r_{xy}=.23), the coordinates were orthogonalized (columns

Table 2 Matrix of overlap coefficients

	SOZ	WGM	NEU	BUN	HHS	ALT	UNT	BAU	STH	RHS	WBL
APP	.132	.044	.230	.154	.193	.052	.165	.057	.152	.151	.150
SOZ		.136	.153	.094	.148	.195	.161	.070	.175	.166	.190
WGM			.082	.064	.082	.116	.077	.130	.224	.066	.074
NEU				.136	.329	.105	.112	.061	.125	.133	.218
BUN					.056	.060	.052	.159	.056	.163	.052
HHS						.048	.095	.038	.175	.122	.300
ALT							.048	.168	.092	.053	.030
UNT								.002	.176	.106	.130
BAU									.055	.060	.039
STH										.154	.230
RHS											.181

Note: All ÜKS but one (UNT/BAU) are significant, cf. MARX, 1976 b: 269.

z and z' of Table 3) and normalized (columns z_n and z'_n of Table 3). The necessary computations were done by hand (with the aid of a small electronic calculator). In addition, the configuration was slightly rotated (the reason for which is given below). Fig. 4 shows the orthogonalized and rotated sulution.

Interpreting the NMDS structure. Any adequate interpretation has to provide answers to the following questions:
- What is the geometric shape of the solution structure?
- What if anything should the dimensions be called, and are there optimal directions of the dimensional vectors concerning interpretation?
- Are there external variables useful for interpreting the NMDS structure?
- Is there a way towards more detailed interpretations, e.g. by doing a content analysis of the responses?

Table 3 NMDS solution: POLYCON coordinates, orthogonalized coordinates, and normalized rotated coordinates. Preference scales: BTL model and percentages.

Stimulus	FREE ASSOCIATION & NMDS SCALING						PREFERENCE	
	POLYCON		orthog.		norm./rot.		BTL	%
	x	y	z	z'	z_n	z'_n		
APPartement	-.011	-.559	.062	.553	.111	1.228	-.180	.546
SOZialwohnung	-.023	.233	-.053	-.234	-.095	-.519	.858	.280
WohnGeMeinschaft	.160	.765	.059	-.739	.106	-1.640	-1.034	.790
NEUbau	.016	-.224	.045	.224	.081	.497	.104	.480
BUNgalow	.668	-.533	.752	.616	1.349	1.367	-1.377	.826
HochHauS	-.420	-.246	-.384	.191	-.689	.424	1.104	.231
ALTbau	.674	.650	.583	-.560	1.046	-1.243	-.742	.698
UNTermiete	-.834	.175	-.850	-.279	-1.525	-.619	1.663	.113
BAUernhaus	1.077	.289	1.030	-.151	1.848	-.335	-1.386	.792
STudentenHeim	-.382	.268	-.414	-.314	-.743	-.697	.377	.387
ReihenHauS	-.340	-.626	-.255	.578	-.458	1.283	-.178	.555
WohnBLock	-.606	-.191	-.576	.113	-1.034	.251	.791	.314
MEAN	0	0	0	0	0	0	0	.501
Std.Deviation	.573	.466	.577	.451	1	1	1	.241
	r_{xy} = .23		$r_{zz'}$ = .03		$r_{z_n z'_n}$ = .03		$r_{BTL/\%}$=-.996	

As for shape, i.e. the general geometric figure formed by the NMDS solution, it may be characterized by DEGERMAN's (1972) term 'S-structure', which means a hyperspheroid (or Guttman 'circumplex') like the well-known color circle (cf. SHEPARD's 1962 re-analysis of EKMAN's 1954 data, reducing Ekman's 5-factor solution to the three-dimensional shape of a double cone). The points of our NMDS structure indeed, with the possible exception of NEU and SOZ, form a roughly circular figure. Such structures do not necessarily lend themselves to a dimensional interpretation.

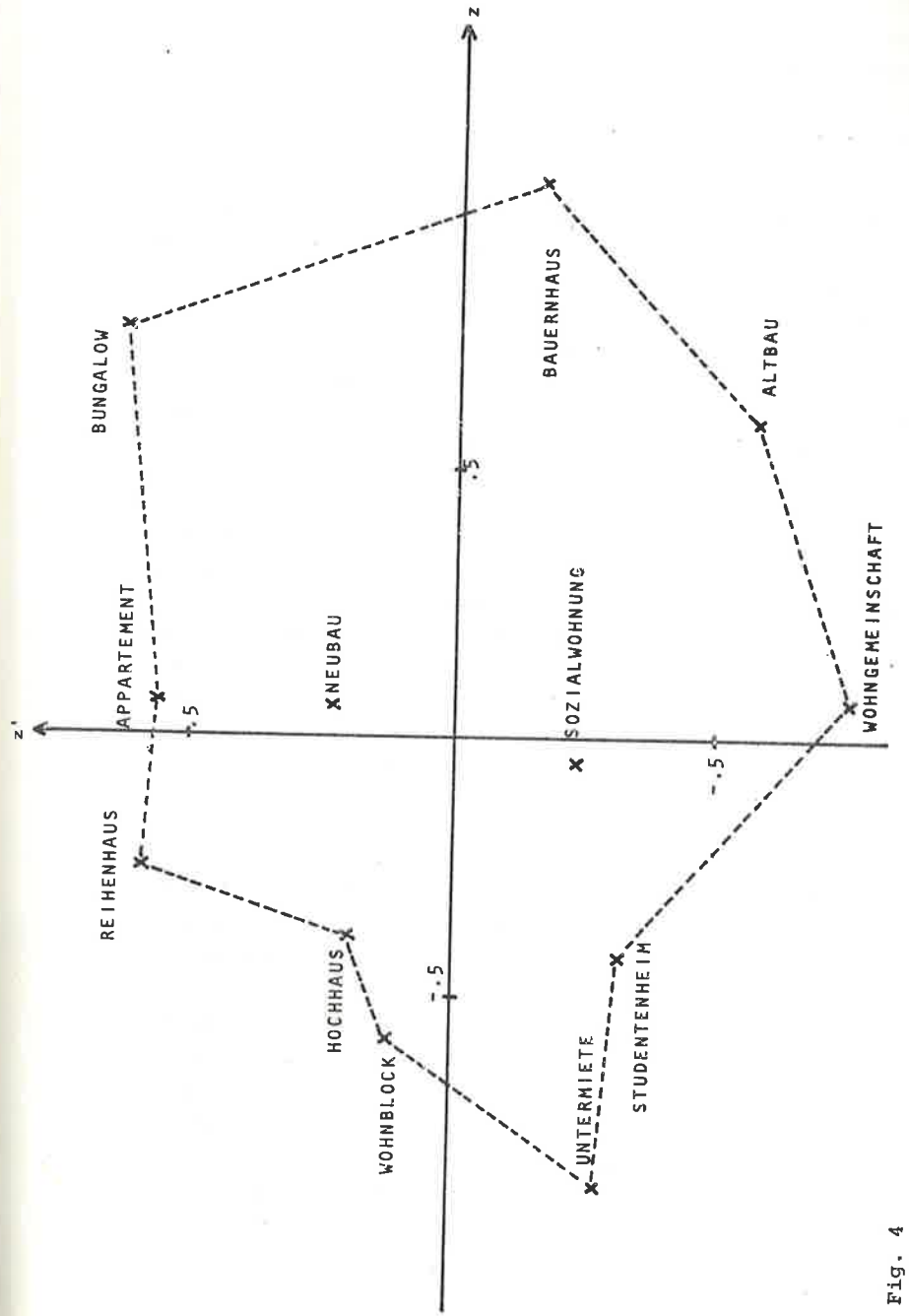


Fig. 4 NMDS structure (orthogonalized and rotated) of the 12 stimulus words

(Cf. the color circle, which is much more convincingly interpreted in terms of polar coordinates, the angle corresponding to the color hue, and the radius to color saturation. The periphery itself may be looked upon as a single, bent dimension.)

On the other hand, there may well be sensible interpretations of Cartesian (or oblique) dimensions. These dimensions being not explicitly qualified, however, (as is with factor loadings of adjectives in a rating task), we need something more than the points' coordinates to answer the question of what farmhouse, cottage (BAU), bungalow (BUN), and "Altbau" (ALT) have in common, as compared to block (WBL), students' dormitory (STH), and subletting (UNT). Two ways there are to get further information, viz. introducing an external variable, and content analysis of the associative responses given.

A variable considered relevant enough to be introduced as an external variable from the start was the subject's preference in housing conditions. The necessary $\binom{12}{2} = 66$ paired comparisons (rating techniques were discarded because of their inaccuracies) were processed according to usual scaling techniques (cf. WOODWORTH & SCHLOSBERG, 1954: 253-256) as well as BTL scaling (BRADLEY & TERRY, 1952; SIXTL, 1967: 209 ff.). The last two columns of Table 3 contain the computed preference values for the 12 stimuli, showing no less than perfect agreement between both scaling methods ($r = -.996$).

Personal preference in housing conditions correlated to a surprisingly high degree with the first of the two POLYCON dimensions (column x of table 3; $r = -.878$). Orthogonalization of the coordinates, as mentioned above, was followed by a rotation designed to produce optimal fit between the first dimension and personal preference. This resulted in a rotation by $-7^{\circ}\ 30'$, obtaining a correlation of $r = -.883$ with the first dimension (and zero correlation with the second one, as well as zero correlation between both the coordinates themselves). Equally, the first dimension, after rotation, contains 78 percent of common variance with personal preference.

We are therefore on firm ground in stating that the first dimension of our obtained structure of meaning is one of subjective evaluation or preference. This result is due to the introduction of an external variable, the methodological value of which thus being illustrated.

Interpreting the second dimension, in the absence of a suitable external variable, is more tedious. No attempt was made to construe a 'dimension of student vs. more established way of life' by contrasting terrace houses (REI), apartment (APP), and bungalow (BUN) on the one hand, and the items low on the second dimension on the other. Common as such a strategy for interpretation is, it seems highly arbitrary and rather distant from the facts. Instead, a content analysis was done.

Content analysis of associative responses, i.e. single words, is much easier to do than an analysis of running text, because one need not worry about problems of context and size of units for analysis. The reflexive process of constructing sensible content categories, however, cannot be dispensed with. (Yet this will not be described here. RITSERT, 1972: 46-55, contains a lucid discussion of this tremendously important issue.)

In the present case, six categories were established that could be described in few words and did not pose difficult problems of how to categorize the responses. These six categories were:

i (individual freedom vs.) dependence
ii restfulness vs. stress and annoyance
iii social life vs. isolation
iv natural beauty (vs. sterile artifact)
v spaciousness & comfort vs. lack of s.&c.
vi expensive vs. cheap

Table 4 shows the relevant categories (arbitrarily defined by an amount of more than 10 % of the responses so classified)

Table 4 Content analysis of free association data (see text)

STIMULUS WORD	CAT. 1 pos.: FREI	CAT. 2 pos.: RUHE	CAT. 3 pos.: GESELL.	CAT. 4 pos.: NATUR	CAT. 5 pos.: GROSS	CAT. 6 pos.: TEUER
APPartment	-	-	13%neg.	-	28% pos. neg.	13%pos.
SOZialwohnung	-	-	-	-	19%neg.	27%neg.
WohnGeMeinschaft	-	20%neg.	33%pos.	-	-	-
NEUbau	-	-	-	32%neg.	23% pos. neg.	-
BUNgalow	-	-	-	27%pos.	13%pos.	21%pos.
HochHauS	-	12%neg.	17%neg.	-	-	-
ALTbau	-	15%pos.	-	-	46% pos. neg.	-
UNTermiete	36%neg.	25%neg.	-	-	-	-
BAUernhaus	-	26%pos.	-	36%pos.	-	-
STudentenHeim	12%neg.	15%neg.	24%pos.	-	15%neg.	-
ReihenHauS	11%neg.	-	-	11% pos. neg.	17%neg.	-
WohnBLock	13%neg.	17%neg.	-	-	12%neg.	-

for each of the stimulus words. Figs. 5, a to c, attempt to enable the reader to visualize the distribution of the content categories among the stimulus words. The first dimension is determined by the categories i, ii, and iv, of which only cat. ii encompasses the whole of the dimension. Yet even cat. ii is relevant to one half of the circular structure only, so it is not identical with the first dimension. An interpretation of dimension 1, however, is still possible by means of the context categories: Negative preference, as we see, is strongly determined by both lack of individual freedom (i-) and by continuous annoyance and stress (ii-), whereas positive preference, apart from restfulness (ii+), is characterized by the prominence of natural beauty (iv+), flowers, gardens, and rural settings. (For more details on specific responses, see STRUBE, 1977: 93-98.)

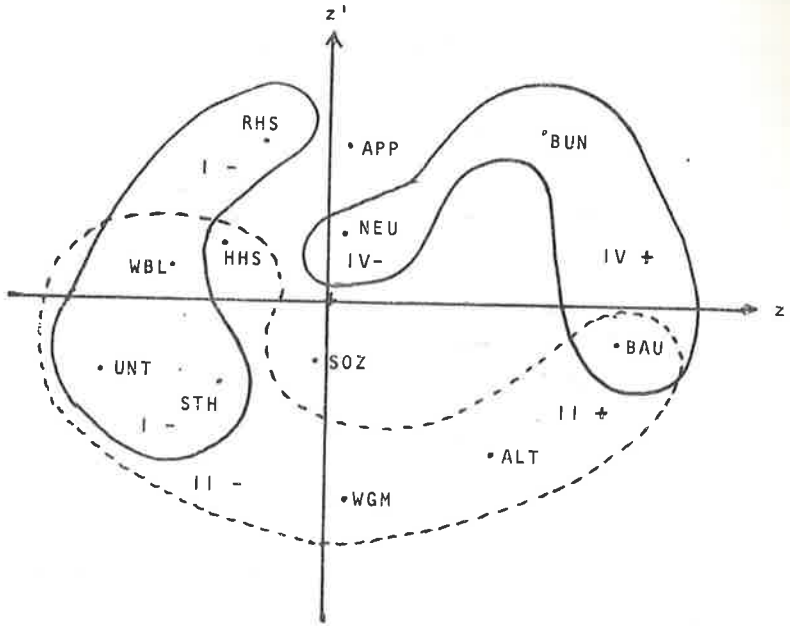


Fig. 5 a

The orthogonal second dimension is ill defined (even cat. iii will not yield to a dimensional interpretation, since to every instance of the category, e.g. APP, there are more words of the same dimensional value, e.g. RHS and BUN, which are not instances of cat. iii). Therefore it seems better to make use of the mainly correlated categories v and vi, by means of which an oblique 2nd dimension may be established, ranging from SOZ and STH to APP and BUN. Though no external variable has been used to validate this dimension, a suitable one now can be

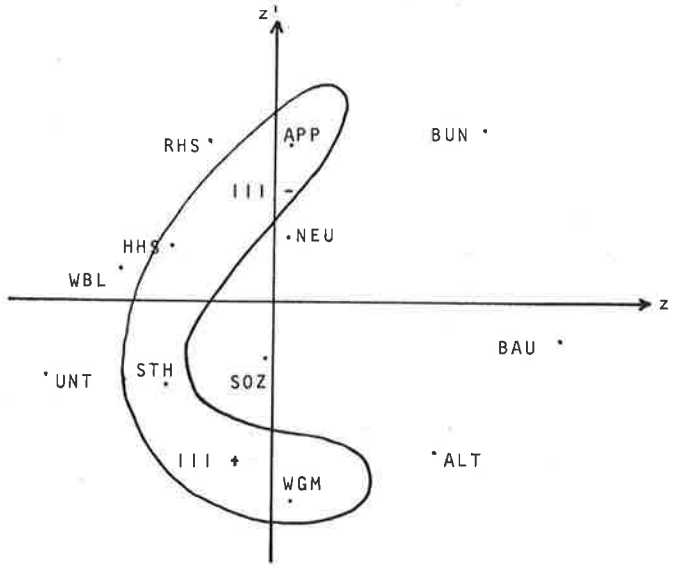


Fig. 5 b

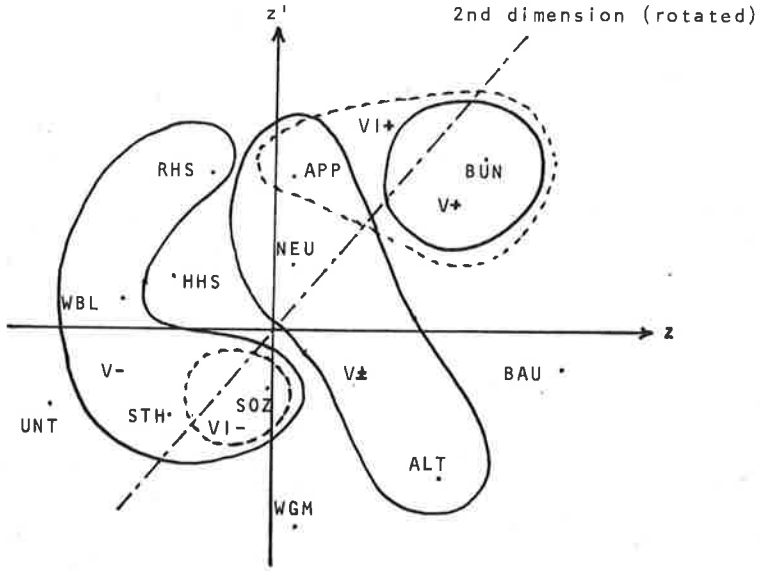


Fig. 5 c

Fig. 5 Range of content categories in NMDS structure (a) Categories i, ii, and iv, (b) Category iii, (c) Categories v and vi

proposed: the price (or rent that has to be paid, or rather the price as estimated by the subjects.

Establishing qualifying categories and/or dimensions, then, is a useful technique to make discussion easier and give a conceptual determination of the associative meaning structure. This is additional, but not at all contrary, to the fact that the visually pregnant shape of 'circular structure' is more adequate to the psychologically relevant aspects of the structure (pertaining, e.g. to memorizing and word association tasks). It may be summarized that the associative meaning structure containing the said 12 stimulus words, with respect to the tested student population, has a roughly circular shape and may further be specified by two meaningful dimensions, oblique to one another, viz. subjective evaluation (or preference), and expensiveness.

Part IV: TWO VALIDATION STUDIES

Does NMDS yield an associative structure proper?

Mediation theories have been discussed in considerable length in part I of this paper, and have been applied to the word association task. So the reader may rightly ask for empirical evidence stressing such a relationship. In other words, the nature of the NMDS structure has to be proven to be an associative one. Now what criteria would be useful to decide this issue?

If the circular structure yielded by the NMDS procedure is an associative one indeed, the distances between the 12 points (standing for the 12 stimulus words) should covary with the probability of direct associations between the words, i.e. with associative strength between stimuli. Unfortunately, however, this criterion is not at all useful, direct associations being rare (remember that Deese went over to coefficients on just these grounds). But while we are perfectly assured that the point distances in our NMDS structure correlate with associative overlap simply for technical reasons, we have rather to face the question concerning the relationship between associative overlap and associative (habit-)strength.

There is clearly need of an indirect criterion measure. Some simple considerations about consequences of association strength for serial learning provide the necessary cue (cf. DEESE 1961). For instance, an experiment by WEINGARTNER (1963) showed that serial learning of lists of words is greatly facilitated when the sequence of words conforms to the associative structure between the words. We may now turn this result upside down and claim that our list of stimulus words, given in the 'right' sequence (i.e., along the circle) in a serial learning task, should be learned more easily than any other, e.g. random or dimensionally ordered, sequence, if the NMDS 'circle' actually represents an associative structure.

The experiment was done with three lists, shown in Table 5 under the headings of 'associative list', 'random list', and 'preference list'. The 'associative list' was constructed by cutting up the circular NMDS structure between the most distant of neighbored points, i.e. between BAU and BUN, proceeding along the 'circle' and leaving out SOZ in order to prevent any cross-cut associations like NEU-SOZ (omission of NEU would have had the same effect).

Table 5 Word lists given in the serial learning task

ASSOCIATIVE LIST	RANDOMIZED LIST	PREFERENCE LIST
1. BUNGALOW	1. APPARTEMENT	1. BAUERNHAUS
2. APPARTEMENT	2. WOHNGEMEINSCHAFT	2. BUNGALOW
3. NEUBAU	3. NEUBAU	3. WOHNGEMEINSCHAFT
4. REIHENHAUS	4. BUNGALOW	4. ALTBAU
5. HOCHHAUS	5. HOCHHAUS	5. APPARTEMENT
6. WOHNBLOCK	6. ALTBAU	6. REIHENHAUS
7. UNTERMIETE	7. UNTERMIETE	7. NEUBAU
8. STUDENTENHEIM	8. BAUERNHAUS	8. STUDENTENHEIM
9. WOHNGEMEINSCHAFT	9. STUDENTENHEIM	9. WOHNBLOCK
10. ALTBAU	10. REIHENHAUS	10. HOCHHAUS
11. BAUERNHAUS	11. WOHNBLOCK	11. UNTERMIETE

Table 6 Results of serial learning

a) Means and std.-deviations of number of trials

Series	mean	std.-dev.
Assoziationsreihe	8.2	3.01
Präferenzreihe	13.9	3.21
Zufallsreihe	14.3	2.75

b) Analysis of variance (level of significance = 1%)

SOURCE OF VARIANCE	SS	df	MS	F
association vs. random & preference series	232.07	1	232.07	25.81 signif.
preference vs. random s.	0.80	1	.80	-
between groups	232.87	2	116.43	12.96 signif.
within groups	242.60	27	8.99	
total	475.47	29		

The lists were presented to ten subjects each (students not having participated in the main experiment) on a memory drum at a speed of 2 sec per word. After one trial, the S had to anticipate the following word. Easiness of attainment was measured by number of trials necessary for complete mastery of the list. The Ss learning the preference list were subsequently tested for their personal preference; the latter usually conformed to the list's sequence, and one S, whose preference grossly differed from the rest, had to be excluded.

Table 6 shows the results. Clearly, the NMDS sequence is significantly easier to learn than either of the other lists, the preference list (reflecting a dimensional ordering of the words)

not differing significantly from a random order. We may therefore conclude that

1. the geometric structure resulting from NMDS processing of overlap coefficients is an associative structure indeed, and
2. dimensional interpretations may lack relationship to the relevant psychological variables even when based on external criteria.

Associative structure and Osgood's semantic space.

It is an interesting question whether Osgood's semantic differentiation technique results in a semantic structure comparable to an associative structure. Osgood, of course, does not claim to assess all the various aspects of meaning, as the title of his original publication might suggest (OSGOOD et al., 1957: "The measurement of meaning"). He restricts the scope of semantic differentiation to the assessment of connotations, i.e. the affective components of meaning. Established by carefully cross-validated factor analysis (see OSGOOD, 1962, for example), there emerges a three-dimensional (and, because of factor analysis, Euclidean) semantic space comprising E (evaluation), P (potency), and A (activity) as its basic vectors, of which E is usually the most important one.

Since it must be assumed that associative processes reflect denotative as well as connotative aspects of meaning, an intricate rather than simple relationship should be expected between the results of Osgood's technique and our associative one. There should also be some identification possible between EPA-structures and the associative structure, yet in a way that the associative one covers more aspects than the former, notably denotative ones.

The reader is assumed to be familiar with semantic differentiation (a thorough introduction give OSGOOD et al., 1957). Therefore, we may proceed by describing the results of our assessment of the 12 stimulus words, now serving as 'concepts',

with the quite usually constructed semantic differential shown in Table 7.

Table 7 The semantic differential

gut	3	2	1	O	1	2	3	schlecht
häßlich	3	2	1	O	1	2	3	schön
human	3	2	1	O	1	2	3	inhuman
angenehm	3	2	1	O	1	2	3	unangenehm
bitter	3	2	1	O	1	2	3	süß
heiter	3	2	1	O	1	2	3	traurig
schwer	3	2	1	O	1	2	3	leicht
aktiv	3	2	1	O	1	2	3	passiv
schwingend	3	2	1	O	1	2	3	ruhend
monoton	3	2	1	O	1	2	3	abwechslungsreich
weich	3	2	1	O	1	2	3	hart
stark	3	2	1	O	1	2	3	schwach
langsam	3	2	1	O	1	2	3	schnell
aufregend	3	2	1	O	1	2	3	beruhigend
freundlich	3	2	1	O	1	2	3	unfreundlich
einsam	3	2	1	O	1	2	3	gesellig
befreiend	3	2	1	O	1	2	3	bedrückend

The 17 items included 6 referring to E, 3 to P, and 4 to A, as well as another 4 items of unknown EPA reference supposed to cover important aspects of the concepts assessed. Complete judgments were obtained from all 79 subjects participating in the main experiment.

The usual kind of factor analysis (modified Hotelling procedure and Varimax rotation, program FACTOR of the SPSS series) produced, oddly enough, a series of factor eigenvalues of 8.21, 1.09, and 0.42. This is very nearly the kind of distorted solution Osgood himself has often described, e.g.: "The greater the emotional or attitudinal loading of the set of concepts to be judged, the greater the tendency of the semantic framework to collapse into a single combined dimension." (OSGOOD, 1962: 25).

Indeed the first factor, explaining 48 percent of the total variance, clearly represents Evaluation, while the much lesser second factor is rather one of Activity (though without containing "active-passive" itself). The P items partly merged with E, having rather low communality even in an attempt at a three-factor solution. Tables 8 and 9 show, respectively, factor loadings of the 17 items, and factor scores for the 12 concepts in the two-factor solution. (For a more detailed discussion, see STRUBE, 1977: ch. IV).

Table 8 Factor loadings of semantic differential scales

SCALE	f.1	f.2	h^2
1. gut-schlecht	.789	-.037	.623
2. häßlich-schön	-.799	.041	.641
3. human-inhuman	.807	-.051	.653
4. angenehm-unangenehm	.849	-.074	.727
5. bitter-süß	-.802	.066	.648
6. heiter-traurig	.877	.117	.783
7. schwer-leicht	-.618	-.086	.389
8. aktiv-passiv	.729	.321	.643
9. schwingend-ruhend	.350	.557	.433
10. monoton-abwechslungsreich	-.755	-.224	.621
11. weich-hart	.637	-.069	.410
12. stark-schwach	.471	.098	.231
13. langsam-schnell	-.054	-.589	.350
14. aufregend-beruhigend	-.247	.496	.307
15. freundlich-unfreundlich	.843	.014	.710
16. einsam-gesellig	-.655	-.174	.460
17. befreiend-bedrückend	.819	.079	.677

Table 9 Factor scores of the 12 stimulus words on semantic differential factors

STIMULUS	f. 1	f. 2
Neubau	.544	-.203
Untermiete	.729	.054
Bauernhaus	-1.216	.527
Appartement	.101	-.071
Sozialwohnung	.489	.208
Bungalow	-.813	.195
Wohngemeinschaft	-.915	-.921
Hochhaus	1.151	-.107
Wohnblock	.721	.142
Altbau	-.684	.740
Reihenhaus	-.066	.348
Studentenheim	-.040	-.912

The important first factor, Evaluation, correlates nearly perfectly with the preference scale (r = -.92 uncorrected!), and reasonably high with the first NMDS dimension (r = -.77). We are therefore entitled to assume that the result of Osgood's technique indeed represents the connotative aspect contained in the associative structure. Regarding their respective theoretical claims, it may be concluded that both these techniques, both based on mediation theory, have been found to produce converging results. Notably the method of analyzing associative responses, which has been proposed in this paper, has resulted in greater differentiation of structure as well as encompassing both denotative and connotative aspects of meaning.

SUMMARY

Theoretical views on meaning as conzeptualized in association theory and earlier experimental work in this field are reviewed. A method for the analysis of associatively related word fields is described, using modified overlap coefficients computed from word association data as input for a NMDS (nonmetric multi-dimensional scaling) program. An experimental example is given, and problems of interpretation are discussed. Finally, two validation studies on the method are provided.

REFERENCES

ANDERSON, J.R., BOWER, G.H., Human associative memory. Washington, Winston 1973.

ATTNEAVE, F., Dimensions of similarity. In: Am.J.Psychol., 63, 1950, 516-556.

BOUSFIELD, W.A., The Problem of meaning in verbal learning. In: COFER 1961, 81-91.

BOUSFIELD, W.A., COHEN, B.H., WHITMARSH, G.A., Associative clustering in the recall of words of different taxonomic frequencies of occurence. In: Psychol. Reports, 4, 1958, 39-44.

BOYD, J.P., Information distance for discrete structures. In: SHEPARD et al. 1972, 213-223.

BRADLEY, R.A., TERRY, M.E., The rank analysis of incomplete block designs. I.: The method of paired comparisons. In: Biometrika 39, 1952, 324-345.

CLARK, H.H., Word association and linguistic theory. In: LYONS 1970, 271-286.

COFER, C.N. (Ed.), Verbal learning and verbal behavior. New York, McGraw-Hill 1961.

COFER, C.N., Verbal learning. In: MARX 1969, 299-378.

COOMBS, C.H., A theory of data. New York, Wiley 1964.

DEESE, J., Associative structure and the serial reproduction experiment. In: J.Abn.Soc.Psychol. 63, 1961, 95-100.

DEESE, J., On the structure of associative meaning. In: Psychol. Rev. 69, 1962, 161-175.

DEGERMANN, R.L., The geometric representation of some simple structures. In: SHEPARD et al. 1972, 193-211.

EBBINGHAUS, H., Über das Gedächtnis. Untersuchungen zur experimentellen Psychologie. (1885). Reprint: Darmstadt, Wiss. Buch-Ges. 1971.

EKMAN, G., Dimensions of color vision. In: J.Psychol. 38, 1954, 467-474.

GALTON, F., Psychometric experiments. In: BRAIN 2, 1880, 149-162.

GALTON, F., Inquires into human faculty and its development. London, Macmillan 1883.

GIGERENZER, G.: Nonmetrische multidimensionale Skalierung als Modell des Urteilsverhaltens. Zur Integration von dimensionsanalytischer Methodik und psychologischer Theoriebildung. Diss. Univ. München 1977.

GIGERENZER, G., STRUBE, G., Zur Revision der üblichen Anwendung dimensionsynalytischer Verfahren. In: Zeitschrift f. Entwicklungspsychologie u. Pädagog. Psychologie 10, 1978, 75-86.

GUTTMAN, L., A general nonmetric technique for finding the smallest coordinate space for a configuration of points. In: Psychometrika 33, 1968, 469-506.

HERMANN, T., DEUTSCH, W., Psychologie der Objektbenennung. Bern, Stuttgart, Wien, Huber 1977.

HULL, C.L., LUGOFF, L.S., Complex signs in diagnostic free association. In: J.Exp.Psychol. 4, 1921, 111-136.

HYMAN, R., WELL, A., Judgements of similarity and spatial models. In: Percept. Psychophysics 2, 1967, 238-248.

HYMAN, R., Perceptual separability and spatial models. In: Percept. Psychophysics 3, 1968, 161-165.

JENKINS, J.J., MINK, W.D., RUSSELL, W.A., Associative clustering as a function of verbal association strength. In: Psychol. Reports 4, 1958, 127-136.

JENKINS, J.J., RUSSELL, W.A., Associative clustering during recall. In: J. Abn. Soc. Psychol. 47, 1952, 818-821.

JENKINS, P.M., COFER, C.N., An exploratory study of discrete free associations to compound verbal stimuli. In: Psychol. Reports 3, 1957, 599-602.

JUNG, C.G., RIKLIN, F., Diagnostische Assoziationsstudien. In: Z. Psychiatrie Neurol. 3, 1904, 193-215.

KENT, H.G., ROSANOFF, A.J., A study of association in insanity. In: Am. J. Insanity 67, 1910, 37-96 und 317-390.

KRUSKAL, J.B., Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. In: Psychometrika 29, 1964, 1-28 (a).

KRUSKAL, J.B., Multidimensional scaling: A numerical method. In: Psychometrika 29, 1964, 115-129 (b).

LAUCHT, M., HERRMANN, T., Zur Direktheit von Direktiva. Arbeiten der Forschungsgruppe Sprache und Kognition am Lehrstuhl Psychologie III, Bericht Nr. 1. Universität Mannheim, 1978.

LEUNINGER, H., MILLER, M.H., MÜLLER, F., Psycholinguistik. Ein Forschungsbericht. Frankfurt, Fischer-Athenäum 1972.

LINGOES, J.C., A general survey of the Guttman-Linoes nonmetric program series. In: SHEPARD et al. 1972. 49-68.

MARX, W., Die Messung der assoziativen Bedeutungsähnlichkeit. In: Z. Exp. Angew. Psychol. 23, 1976, 62-76 (a).

MARX, W., Die statistische Sicherung des Überlappungskoeffizienten. In: Z. Exp. Angew. Psychol. 23, 1976, 267-270 (b).

NOBLE, C.E., An analysis of meaning. In: Psychol. Rev. 59, 1952, 421-430.

OSGOOD, C.E., Comments on Prof. Bousfield's paper. In: COFER 1961, 91-106.

OSGOOD, C.E., Studies on the generality of affective meaning systems. In: Am. Psychologist 17, 1962, 10-28.

OSGOOD, C.E., On the whys and wherefores of E.P. and A. In: J. Pers. Soc. Psychol. 12, 1969, 194-199.

OSGOOD, C.E., SUCI, G.J., TANNENBAUM, P.H., The measurement of meaning. Urbana, Chicago, London, Univ. Illinois Press 1957.

PAWLOW, I.P., Auseinandersetzung mit der Psychologie. (ed. by G. BAADER & U. SCHNAPPER). München, Kindler 1973.

RAPOPORT, A., FILLENBAUM. S., An experimental study of semantic structures. In: ROMNEY et al. 1972.

RITSERT, J., Inhaltsanalyse und Ideologiekritik. Ein Versuch über kritische Sozialforschung. Frankfurt, Fischer-Athenäum 1972.

ROMNEY, A.K., SHEPARD, R.N., NERLOVE, S.B. (Eds.), Multidimensional scaling. Vol. 2: Applications. New York, London, Seminar Press, 1972.

ROSCH, E.H., On the internal structure of perceptual and semantic categories. In: T. MOORE (Ed.), Cognitive development and the acquisition of language. New York, Academic Press, 1973.

RUSSELL, W.A., JENKINS, J.J., The complete Minnesota norms for response to 100 words from the Kent-Rosanoff Association Test. In: Technical Report n. 11, Univ. of Minnesota 1954.

SHEPARD, R.N., The analysis of proximities: Multidimensional scaling with an unknown distance function. In: Psychometrika 27, 1962, 125-140 and 219-246.

SHEPARD, R.N., ROMNEY, A.K., NERLOVE, S.B. (Eds.), Multidimensional scaling. vol. 1: Theory. New York, London, Seminar Press, 1972.

SIXTL. F., Meßmethoden der Psychologie. Weinheim, Beltz, 1967.

SKINNER, B.F., Verbal behavior. New York, Appleton-Century-Crofts, 1957.

STRUBE, G., Wortbedeutung als psychlogisches Problem. Zur Konkurrenz von Forschungsprogrammen in der Sprachpsychologie. Diss. Universität München, 1977 (a).

STRUBE, G., Bemerkungen zum Schicksal des psychometrischen Forschungsprogramms. In: G. STRUBE (Hg.), Binet und die Folgen(Die Psychologie des 20. Jahrhunderts, Bd. V). Zürich, Kindler, 1977 (b).

SZALAY, L.B., DEESE, J., Subjective meaning and culture. An assessment through word associations. Hillsdale, N.J., Lawrence Erlbaum, 1978.

TORGERSON, W.S., Theory and methods of scaling. New York, Wiley 1958.

WEINGARTNER,H., Associative structure and serial learning. In: J. Verb. Learn. Verb. Behav. 2, 1963, 476-479.

WENDER, K., Die psychologische Interpretation nichteuklidischer Metriken in der multidimensionalen Skalierung. Diss. Techn. Univ. Darmstadt 1969.

WOODWORTH, R.S., SCHLOSBERG, H., Experimental psychology. Rev. ed. New York, Holt, Rinehart & Winston 1954.

YOUNG, F.W., A model for polynomial conjoint analysis algorithms. In: SHEPARD et al. 1972, 69-105.

CURRENT BIBLIOGRAPHY

ABBREVIATIONS

NTI Naučno-texničeskaja informacija. Serija 2: Informacionnye processy i sistemy

PSML Prague Studies in Mathematical Linguistics

GENERAL

1. ANSHEN, F.: Statistics for Linguists. Rowley, Mass.: Newbury House Publishers, Inc., 1978, IX + 73 pp.

2. ARAPOV, M.V., ŠREJDER, Ju.A.: Zakon Cipfa i princip dissimmetrii sistemy [Zipf's law and the principle of dissymmetry of system]. Semiotika i informatika 10, 1978, 74-95.

3. BAJRAMOVA, L,K.: Matematiko-statističeskoe izučenie sootvetstvija perevodov originalu [A mathematical-statistical study of the relationship of translations to originals]. Očerki grammatiki i leksikologii russkogo jazyka. Kazan' 1977, 39-51.

4. ČIKOIDZE, G.B., CKITIŠVILI, Ž.D.: Ocenka informativnosti otdel'nyx urovnej jazykovogo processora [Estimation of the information content of levels of a linguistic processor]. Trudy instituta sistem upravlenija AN GSSR 16, 1977, 3, 47-59.

5. GERD, A.S. (Red.): Strukturnaja i prikladnaja lingvistika. Mežvuzovskij sbornik. Vypusk 1. Leningrad: Izdatel'stvo Leningradskogo universiteta, 1978, 208 pp.

6. GORODECKIJ, B.Ju.: Leksiko-statističeskaja inventarizacija kompleksa poď jazykov [A lexical-statistical inventarisation of a complex of sublanguages]. Problemy teoretičeskoj i èksperimental'noj lingvistiki. Moskva: Izdatel'stvo Moskovskogo universiteta, 1977, 21-42.

7. GRABOVSKAJA, S.V.: O statističeskoj ocenke kartoteki dlja

lingvističeskix issledovanij [On the statistical estimation of a card-index for linguistic investigations]. Strukturnaja i matematičeskaja lingvistika 5, 1977, 26-32.

8. HŘEBÍČEK, L.: The Turkish language reform and contemporary grammar (The difference between the spoken (S) and written (T) texts on the level of grammatical morphemes). Archiv orientální 46, 1978, 334-337.

9. JAKUBAJTIS, T.A.: Ispol'zovanie ÈVM v lingvističeskix issledovanijax [The use of computers for linguistic research]. Voprosy jazykoznanija 1979/3, 127-131.

10. JONES, A., CHURCHHOUSE, R.F. (eds.): The Computer in Literary and Linguistic Studies. Cardiff: The University of Wales Press, 1976, VIII + 362 pp.

11. KEMP, K.W.: Personal observations on the use of statistical methods in quantitative linguistics. In [10], 59-77.

12. KÖNIGOVÁ, M.: Statistical analysis of citations from quantitative linguistics. PSML 6, 1978, 177-195.

13. KOPEJKIN, S.V., OSTAPENKO, V.E.: Zakon Cipfa i sopostavitel'nyj analiz častotnyx struktur anglijskogo, francuzskogo, rumynskogo i russkogo jazykov na baze matematičeskix modelej [Zipf's law and the contrastive-comparative analysis of frequency structures in English, French, Roumanian and Russian on the basis of mathematical models]. Naučnye trudy Kujbyševskogo pedagogičeskogo instituta 193, 1977, 91-94.

14. KRÁLIK, J.: On the dispersion and its computation. PSML 6, 1978, 149-158.

15. KUZEMSKAJA, N.A., SKOROXOD'KO, È.F.: Ob odnom podxode k razrabotke kriteriev količestvennoj ocenki kačestva perevoda [On one approach to the elaboration of criteria for a quantitative estimation of the quality of a translation]. Strukturnaja i matematičeskaja lingvistika 5, 1977, 69-75.

16. LEVIČ, A.P.: Informacija kak struktura sistem [Information as the structure of systems]. Semiotika i informatika 10,

1978, 116-132.

17. MARCUS, S.: Mathematical and computational linguistics and poetics. Rosetti, A., Golopentia-Eretescu, S. (eds.): Current Trends in Romanian Linguistics. Bucureşti: Editura Academiei Republicii Socialiste România, 1978, 559-588.

18. MARTYNENKO, G.Ja.: Nekotorye zakonomernosti koncentracii i rassejanija ėlementov v lingvističeskix i drugix složnyx sistemax [Some regularities of the concentration and dispersion of elements in linguistic and other complex systems]. In [5], 63-79.

19. MUXAMEDOV, S.: Linguostatističeskie issledovanija v tjurkskix jazykax [Linguostatistical investigations in the Turkish languages]. Issledovanija po literaturovedeniju i jazykoznaniju. Taškent 1977, 195-197.

20. ORLOV, Ju.K.: Nekotorye aspekty organizacii informacii čelovekom (statističeskaja struktura proizvedenij iskusstva) [Some aspects of the organization of information by man (the statistical structure of works of art)]. Trudy Instituta Kibernetiki AN GSSR 1, 1977, 267-275.

21. PIOTROVSKIJ, R.-G.: Inženernaja lingvistika i lingvističeskie avtomaty [Engineering linguistics and linguistic automata]. Vestnik Akademii nauk SSSR 1978/9, 112-119.

22. SIGURD, B.: Arbitrariness, frequency and abbreviation. Studia Linguistica 32, 1978, 169-173.

23. SPIVAK, D.L.: Jazyk kak nečetnaja sistema [Language as a fuzzy system] (Review of Piotrovskij, R.G., Bektaev, K. B., Piotrovskaja, A.A.: Matematičeskaja lingvistika. Moskva: Vysšaja škola, 1977). NTI 1978/10, 12, 29-30.

24. SUXOTIN, B.V.: Optimizacionnye metody issledovanija jazyka [Optimizational Methods of Language Research]. Moskva: Nauka, 1976, 170 pp.

25. TĚŠITELOVÁ, M.: Kvantitativní lingvistika [Quantitative Linguistics]. Praha: Státní pedagogické nakladatelství, 1977, 177 pp.

26. TULDAVA, Ju.: Statističeskaja proverka odnorodnosti s pomošč'ju metoda serii [Statistical examination of homogeneity by means of the serial-method]. Methodica VII. Acta et commentationes Universitatis Tartuensis. Tartu 1978, 143-151.

27. ZUBOV, A.V., ČAPLJA, A.I., ČAPLJA, S.G.: Avtomatičeskoe vydelenie ključevyx slov [Automatic selection of key-words]. In [5], 198-204.

PHONOLOGY

28. DOBROGOWSKA, K.: Badanie zależności między akcentem a intonacją w mowie nieemocjonalniej [A study of the interdependence between accent and intonation in unemotional speech]. Lingua Posnaniensis 21, 1978, 65-76.

29. FAJZIEV, A.A.: Statističeskoe izučenie čeredovanija glasnyx i soglasnyx bukv v uzbekskom literaturnom tekste [A statistical study of alternations of vowels and consonants in Usbekian literary texts]. Verojatnostnye processy i matematičeskaja statistika. Taškent 1978, 172-178.

30. HŘEBÍČEK, L.: The phonological adaptations of borrowings: methodological problems. Rapports, Co-Rapports, Communications tchécoslovaques pour le IV Congrès de l'Association Internationale d'Edtudes du Sud-Est Européen, Prague 1979, 371-378.

31. KARLGREN, M.: On the arbitrariness of shorthand signs. Studia Linguistica 32, 1978, 119-136.

32. KRÁMSKÝ, J.: Quantitative analysis of near-identical phonological systems. PSML 6, 1978, 9-38.

33. KUZINA, V.: Statistika bukv v tekstax raznyx tipov sovremennogo latyšskogo jazyka [Statistics of letters in modern Latvian texts of various types]. Statistika un valodas funkcionālie stili. Rīga: Zinātne, 1977, 97-106.

34. LINDBLOM, B.: Phonetic aspects of linguistic explanation. Studia Linguistica 32, 1978, 137-153.

35. MIYAZIMA, T.: The new style Chinese characters and stroke numbers [In Japanese]. Mathematical Linguistics 11, 1978, 301-306.

36. NOMURA, M., ITO, K.: To what degree are present-day Chinese characters in Japanese phonetic? [In Japanese]. Mathematical Linguistics 11, 1978, 306-311.

37. PADLUŽNY, A.I.: Narys akustyčnaj fanetyki belaruskaj movy [Essay on Acoustical Phonetics of Byelorussian]. Minsk: Navuka i texnika, 1977, 166 pp.

38. PLUCIŃSKI, A.: Interdependencies between the fundamental frequency and the intensity levels in Polish colloquial speech. Lingua Posnaniensis 21, 1978, 77-89.

39. REMMEL, M.: Traces of phonetical parameters in phonostatistical distributions. Estonian Papers in Phonetics. Tallinn 1977, 81-88.

40. RIETVELD, A.C.M.: Über die artikulatorische Ähnlichkeit deutscher Vokale. Phonetica 36, 1979, 115-129.

41. STEIN, D.: Phonotactic and semantic constraints on token occurrence frequencies of an affixed verb form. Papiere zur Linguistik 19, 1978, 58-77.

42. ŽILINSKIENĖ, V.: Lietuvių kalbos raidžių dažnumas publicistikos tekstuose [Letter frequency in Lithuanian publicistic texts]. Kalbotyra 29/1, 1978, 83-95.

MORPHOLOGY

43. AXMEDOVA, U.K.: Nekotorye statističeskie xarakteristiki obrazovanija slov iz slogov v uzbekskom literaturnom jazyke [Some quantitative characteristics of the formation of words from syllables in the Uzbekian written language]. Verojatnostnye processy i matematičeskaja statistika. Taškent 1978, 11-18.

44. GERD, A.S., KAPORULINA, L.V., KOLESOV, V.V.: Lingvostatističeskoe opisanie russkogo imennogo sklonenija [A statistical linguistic description of the Russian nominal declension]. In [5], 110-138.

45. JAKUBAJTIS, T.: Statističeskaja xarakteristika častej reči v latyšskom jazyke [Statistical characterization of the parts of speech in Latvian]. Statistika un valodas funkcionālie stili. Rīga: Zinātne, 1977, 7-28.

46. LEHFELDT, W.: Formenbildung des russischen Verbs. Versuch einer analytisch-synthetisch-funktionellen Beschreibung der Präsens- und der Präteritumflexion. München: Sagner, 1978, 114 pp.

47. LUDVÍKOVÁ, M.: On the occurrence of syllables in different word positions. PSML 6, 1978, 39-45.

48. MARTIN, W.: On the evolution of word-length in Dutch. In [10], 271-284.

49. MARX, W.: Statistische Information und assoziative Bedeutung verschiedener Wortarten. Zeitschrift für experimentelle und angewandte Psychologie 25, 1978, 431-440.

50. MIZUTANI, S.: A critical analysis of the notion keiyo dosi (adjective verb) [In Japanese]. Mathematical Linguistics 11, 1978, 283-301.

51. MÜLLER, B.S.: Einige statistische Angaben über zusammengesetzte Substantive im Deutschen. Germanistische Linguistik 1977/1-2, 171-198.

52. NIKONOV, V.A.: Dlina slova [Word-length]. Voprosy jazykoznanija 1978/6, 104-111.

53. RIZAEV, S.A., JUSUPOVA, D.Ju.: O distributivno-statističeskom issledovanii leksiko-morfologičeskoj struktury slova sovremennogo uzbekskogo jazyka [On the distributional-statistical investigation of lexico-morphological word-structure in contemporary Uzbek]. Issledovanija po literaturovedeniju i jazykoznaniju. Taškent 1977, 198-203.

54. SLIPČENKO, L.D.: Častota fonem i dlina slova (Na materiale anglijskogo jazyka) [Phoneme frequency and word-length (based on material from the English language)]. Strukturnaja i matematičeskaja lingvistika 5, 1977, 104-109.

55. ŠAJKEVIČ, A.Ja.: Vydelenie klassov slov i paradigm posredst-

vom distributivno-statističeskogo metoda (na materiale komedij Šekspira) [The creation of word-classes and paradigms with the help of the distributional-statistical method (based on Shakespeare's comedies)]. Prikladnaja lingvistika 18, 1976, 96-134.

56. TULDAVA, Ju.: Kvantitativnoe issledovanie struktury odnosložnogo slova v èstonskom jazyke [A quantitative investigation of the structure of Estonian monosyllabic words]. Linguistica 10, 1978, 115-135.

57. ZAPLATKINA, N.: Obščee i različnoe v sistemax odnosložnyx slov slavjanskix jazykov [Common and difference in the systems of monosyllabic words of the Slavonic languages]. Strukturnaja i matematičeskaja lingvistika 5, 1977, 40-46.

58. ZILINSKENE, V.: Imja suščestvitel'noe litovskogo jazyka i ego kategorii v statističeskom aspekte [The Lithuanian noun and its categories seen from the statistical point of view]. Statistika un valodas funkcionālie stili. Rīga: Zinātne, 1977, 29-52.

Syntax

59. BYSTROVA, L.V.: Vyvčennja syntagmatyčnyx zv'jazkiv sliv za dopomogoju statystyčnyx metodiv [The study of syntagmatic connections of words by statistical methods]. Movoznavstvo 1978/4, 44-48.

60. DUŠKOVÁ, L.: The simple sentence in Czech and in English. PSML, 1978, 83-92.

61. GRABOVSKAJA, S.V.: O primenenii statistiki v grammatičeskix issledovanijax [On the application of statistics in grammatical research]. Strukturnaja i matematičeskaja lingvistika 6, 1978, 19-24.

62. KADOMCEV, V.I.: O formal'nyx priznakax kommunikativnyx tipov predloženij [On the formal characteristics of communicative sentence types]. Sootnošenie soderžatel'noj i formal'noj storon jazykovyx edinic. Kemerovo 1977, 31-33.

63. MEN'ŠIKOV, I.I.: O nekotoryx zakonomernostjax vzaimoraspoloženija komponentov frazy v sovremennom russkom jazyke (količestvennyj analiz) [On some regularities in the interarrangement of phrase components in modern Russian (quantitative analysis)]. In [5], 30-38.

64. NEBESKÁ, I.: On the order of clauses in a group of complex sentences expressing causal relations. PSML 6, 1978, 73-81.

65. PETUNIN, Ju.I., NADTOČIJ, E.D.: O korrektnosti statističeskix issledovanij sintaksisa pri raspredelenijax otličajuščixsja ot normal'nogo [On the correctness of statistical syntactical studies with the help of distributions other than normal]. Strukturnaja i matematičeskaja lingvistika 7, 1979, 74-82.

66. ŠULJAK, N.T.: K voprosu matematičeskogo opisanija èlementarnyx sintaksičeskix javlenij [On the problem of the mathematical description of elementary syntactical facts]. Strukturnaja i matematičeskaja lingvistika 6, 1978, 109-117.

67. ŠULJAK, N.T., MAKARENKO, V.G.: K voprosu o statističeskom opisanii sintaksičeskix konstrukcij [On the problem of the statistical description of syntactical constructions]. Strukturnaja i matematičeskaja lingvistika 7, 1979, 108-115.

68. UHLIŘOVÁ, L.: On the statistical distribution of adverbials. PSML 6, 1978, 59-72.

Semantics

69. BAŽENOVA, N.S.: Nekotorye svojstva semantičeskix setej terminologičeskix sistem [Some properties of the semantic nets of terminological systems]. Strukturnaja i matematičeskaja lingvistika 6, 1978, 9-13.

70. HIROSI, N.: On image of onomatopoeia by semantic differential method [In Japanese]. Mathematical Linguistics 11, 1979, 362-370.

71. JOSHITAKE, Ô.: An analysis of word meanings by the check list method - an approach from a computative semantics [In Japanese]. Mathematical Linguistics 11, 1979, 341-352.

72. NAKANO, H.: On image of onomatopoeia by semantic differential method [In Japanese]. Mathematical Linguistics 11, 1978, 312-317.

73. PLOTNIKOV, B.A.: Distributivno - statističeskij analiz leksičeskix značenij [Distributional Statistical Analysis of Lexical Meanings]. Minsk: Vyšėjšaja škola, 1979, 136 pp.

LEXICOLOGY

74. ALEKSEEV, P.M., KAŠIRINA, M.E., TARASOVA, E.M.: Častotnyj slovar' anglijskix fizičeskix terminov [Frequency dictionary of English physical terms]. In [5], 80-88.

75. ANDRJUŠČENKO, V.M.: K voprosu ob ispol'zovanii koėfficienta stabil'nosti v kačestve mery upotrebitel'nosti [On the application of the coefficient of stability as a measure of usage]. Andrjuščenko, V.M. (ed.): Issledovanija v oblasti vyčislitel'noj lingvistiki i lingvostatistiki. Moskva: Izdatel'stvo Moskovskogo universiteta, 1978, 3-40.

76. AUBERT, R., TOMBEUR, P.: Concilium Vaticanum I. Concordance, Index, Listes de fréquence, Tables comparatives. Louvain-la-Neuve: Presses du CETEDOC, 1977, XVII + 275 pp.

77. BOLCHAZY, L.J., GICHAN, G., THEOBALD, F.: Concordance to Thomas More's Utopia. Hildesheim: Olms, 1978, 388 pp.

78. DENISOV, P.N., MORKOVKIN, V.V., SAF'JAN, Ju.A.: Kompleksnyj častotnyj slovar' russkoj naučnoj i texničeskoj leksiki. 3047 slov [A Complex Frequency Dictionary of Russian Scientific and Technical Lexis. 3047 Words]. Moskva: Russkij jazyk, 1978, 408 pp.

79. DIETZE, J.: Zur rechnergestützten Erarbeitung von Thesauren. Wissenschaftliche Zeitschrift der Universität Halle, 28, 1979 G, 107-120.

80. DUBROCARD, M.: Juvénal, Satires. Index verborum, relevés statistiques. Hildesheim: Olms, 1976.

81. GODET, J.-F., MAILLEUX, G.: Corpus des Sources Franciscaines, III: Legenda trium sociorum. Anonymus Perusinus, Fr. Juliani de Spira Vita s. Francisci. Sacrum commercium s. Francisci cum domina paupertate. Concordance, Index verborum, Listes de fréquence, Tables comparatives. Louvain-la Neuve: Presses du CETEDOC, 1976, X + 395 pp.

82. GODET, J.-F., MAILLEUX, G.: Corpus des Sources Franciscaines, V: Opuscula s. Francisci, Scripta s. Clarae. Concordance, Index verborum, Listes de fréquence, Tables comparatives. Louvain-la Neuve: Presses du CETEDOC, 1976, VIII + 425 pp.

83. GUERET, M., ROBINET, A., TOMBEUR, P.: Spinoza. Ethica. Concordances, Index, Listes de fréquences, Tables comparatives. Louvain-la-Neuve: Presses du CETEDOC, 1977, XXI + 538 pp.

84. JABLONSKAJA, N.N.: Častotnyj slovar' nemeckogo poď jazyka xirurgii [Frequency dictionary of the German surgery sub-language]. In [5], 94-103.

85. KURCZ, I., LEWICKI, A., SAMBOR, J., WORONCZAK, J.: Słownictwo współczesnego języka polskiego: Listy frekwencyjne. Tom V: Dramat artystyczny [The Vocabulary of Contemporary Polish: Frequency-lists. Vol. 5: The Artistic Drama]. Warszawa: Polska Akademia Nauk. Instytut Języka Polskiego, 1977, 631 pp.

86. LEBEDEV, B.M.: Častotnyj slovar' nemeckogo poď jazyka televidenija [Frequency dictionary of the German TV sub-language]. In [5], 102-110.

87. MAŽĖJKA, N.S., SUPRUN, A.Ja.: Častotnyj sloŭnik belaruskaj movy: Publicystyka [Frequency Dictionary of the Byelorussian Language: Press]. Minsk: Vydavectva BDU im. Ŭ.I. Lenina, 1979, 216 pp.

88. ORLOV, Ju.K.: Model' častotnoj struktury leksiki [A model of the frequency structure of the lexis]. Andrjuščenko, V.M. (ed.): Issledovanija v oblasti vyčislitel'noj ling-

vistiki i lingvostatistiki. Moskva: Izdatel'stvo Moskovskogo universiteta, 1978, 59-118.

89. ROŽKOV, V.V.: O statističeskom opredelenii granic mnogoznačnoj leksiki [On statistical definition of the boundaries of ambiguous words]. [5], 88-93.

90. RUSKOVA, M.P.: Statističeskoe raspredelenie leksiki v bolgarskoj pis'mennosti XVIII v. (Kotlenskij damaskin) [Statistical distribution of vocabulary in the Bulgarian written language of the 18th century (Kotlensky Damaskin)]. In [5], 138-145.

91. SIMONOV, M.D.: Količestvennoe vyraženie rodstvennosti jazykov. (Na leksičeskom materiale) [Quantitative expression of the genetic relationships between languages (based on lexical material)]. Issledovanija po jazykam narodov Sibiri. Novosibirsk 1977, 101-107.

92. VERETA, V.I.: O častotnom slovare naučno-texničeskoj leksiki sovremennogo ukrainskogo jazyka [On a frequency dictionary of the scientific-technical vocabulary of modern Ukrainian]. Strukturnaja i matematičeskaja lingvistika 5, 1977, 9-17.

Text Analysis

93. ARAPOV, M.V.: Izmerenie leksičeskogo bogatstva tekstov [The measurement of the lexical richness of texts]. Tekst-Język-Poetyka. Wrocław: Ossolineum, 1978, 185-210.

94. ARAPOV, M.V., TER-GASPARJAN, L.I., XERC, M.M.: Sravnenie častotnyx slovarej [Comparison of frequency dictionaries]. NTI 1978/4, 20-29.

95. BEČKA, J.V.: Application of quantitative methods in contrastive stylistics. PSML 6, 1978, 129-147.

96. BEKTAEV, K.B.: Statistiko-informacionnaja tipologija tjurkskogo teksta [Statistical-Informational Typology of Turkish Texts]. Alma-Ata: Nauka, 1978, 183 pp.

97. BORODA, M.G.: Častotnye struktury muzykal'nyx tekstov [Frequency structures of musical texts]. Sbornik statej posvjaščennyx 60-letiju Velikoj Oktjabr'skoj socialističeskoj revoljucii. Tbilisi: Mecniereba, 1977, 178-202.

98. CHISHOLM, D.: A survey of computer-assisted research in modern German. Computers and Humanities 11, 1978, 279-287.

99. DINU, M.: How to estimate the weight of stage relations? Poetics 6, 1977, 209-227.

100. DROMMEL, R.H.: Schätztests im Rahmen einer experimentellen Textlinguistik: Ein Werkstattbericht. Folia Linguistica 11, 1977, 237-258.

101. FRAUTSCHI, R.L.: Some parameters of public taste in Enlightment French fiction. In [10], 156-176.

102. GEFFROY, A., GUILHAUMOU, J., HARTLEY, A., SALEM, A.: Factor analysis and lexicometrics: Shifters in some texts of the French revolution (1793-1794). In [10], 177-193.

103. HELBICH, J.: Some results of an experiment in statistical selection of keywords. PSML 6, 1978, 159-175.

104. IVANOV, V.I.: Razmery načal'nyx i zaključitel'nyx predloženij v čuvašskix i nemeckix rasskazax [On the length of initial and final sentences in Chuvashian and German stories]. Sopostavitel'naja lingvistika i obučenie inostrannym jazykam v uslovijax dvujažycija 3, Čeboksary 1978, 36-47.

105. IVANOV, V.I., IVANOVA, T.S.: O razmerax predloženija v sovremennoj čuvašskoj i nemeckoj xudožestvennoj proze [On sentence-length in modern Chuvashian and German literary prose]. Sopostavitel'naja lingvistika i obučenie inostrannym jazykam v uslovijax dvujazycija 3, Čeboksary 1978, 48-74.

106. JAKUBAJTIS, T.A.: O stabil'nosti častot i odnorodnosti tekstovyx sovokupnostej [On the stability of frequencies and the homogeneity of text-entities]. Izvestija Akademii nauk Latvijskoj SSR 1978/6, 69-83.

107. KJETSAA, G.: Sound and meaning according to Lomonosov.

In [10], 230-239.

108. LEIGHTON, J.: Automatic analysis of simple rhetorical devices in 17th century German sonnets. In [10], 246-254.

109. LIÉNARD, E.: Répertoire prosodiques et métriques. Lucrèce, De rerum naturae, L. III, Valerius Flaccus, Argonautica, L. VII, Germanicus, Aratea. Bruxelles: Éditions de l'Université, 1978, 204 pp.

110. MAEGAARD, B., SPANG-HANSSEN, E.: La segmentation automatique du français écrit (Coll. Documents de linguistique quantitative). Paris 1978, 122 pp.

111. MALONEY, G., VILLENEUVE, X.F., DIAZ, I.: La longeur des mots comme mesure d'homogénéité chez Hippocrate. Phoenix 3, 1977, 95-122.

112. MARTINDALE, C.: A quantitative analysis of diachronic patterns in some narratives of Poe. Semiotica 22, 1978, 287-308.

113. MARTINDALE, C.: Sit with statisticians and commit a social science: interdisciplinary aspects of poetics. Poetics 7, 1978, 273-282.

114. MATEEVA, A.C.: Nekotorye količestvennye xarakteristiki strukturnoj složnosti celogo teksta (na materiale bolgarskogo jazyka) [Some Quantitative Characteristics of the Structural Complexity of a Whole Text (Based on Material from Bulgarian)]. Avtoreferat kandidatskoj dissertacii. Moskva 1978, 26 pp.

115. MICHAELSON, S., MORTON, A.Q.: Things ain't what they used to be. In [10], 78-84.

116. MIZUTANI, Sh.: Distributions of word-occurrences in Japanese popular songs 1929-37 [In Japanese]. Mathematical Linguistics (Tokyo) 11, 1978, 191-197.

117. MIZUTANI, Sh.: Discrimination between Waka poems of plum- and cherry-blossoms by their word-uses [In Japanese]. Mathematical Linguistics 12, 1979, 1-13.

118. NADAREJŠVILI, I.S.: Sravnitel'nyj statističeskij analiz leksiki kak metod izučenija tvorčestva pisatelja (Na materiale prozy K. Gamsaxurdija) [The comparative stati-

stical analysis of the vocabulary as a method of investigating the works of a writer (based on material from the prose of K. Gamsaxurdija)]. Strukturnaja i matematičeskaja lingvistika 6, 1978, 45-52.

119. NADAREJŠVILI, G.Š., NADAREJŠVILI, I.Š., ORLOV, Ju.K.: Nestacionarnye javlenija v processe porožděnija teksta [Shifting phenomena in the process of text-generating]. Trudy Instituta Kibernetiki AN GSSR 1, 1977, 276-285.

120. PIIROJA, E.: Über die Messung der Schwierigkeit deutschsprachiger Texte (Anhand des Lehrbuches für das 2. Studienjahr deutscher Philologen). Ütü töid germaani filoloogia alalt, Tartu, 1978/1, 42-55.

121. RADDAY, Y.T.: A Hebrew computer Bible (In Hebrew). Hebrew Computational Linguistics 13, 1978, 92-99.

122. ROCŁAWSKI, B.: Funkcjonowanie parametrów fonostatystycznych w tekstach języka artystycznego [Fonctionnement des paramètre phonostatistiques dans des textes en langue artistique]. Bukak, J., Wilkón, A. (ed.): Z zagadnień języka artystycznego. Kraków 1977, 339-354.

123. SEVBO, I.P., PETUNIN, Ju.I., GALJUTA, E.D.: Èksperiment po raspoznavaniju avtora osnovannyj na predvaritel'nom statističeskom issledovanii sintaksičeskix struktur [Experiment of identifying the author, based on a preliminary statistical investigation of syntactic structures]. Strukturnaja i matematičeskaja lingvistika 5, 1977, 96-103.

124. ŠULJAK, N.T.: Primenenie ÈVM dlja statističeskogo analiza sintaksisa teksta [Application of electronic calculating-machines to the statistical analysis of text-syntax]. Strukturnaja i matematičeskaja lingvistika 5, 1977, 123-128.

125. SUPRUN, A.E.: K količestvennoj ocenke leksičeskogo bogatstva teksta [On the quantitative evaluation of the lexical richness of a text]. Filologičeskie nauki 1979/1, 44-48.

126. TALLENTIRE, D.R.: Confirming intuitions about style using

concordances. In [10], 309-328.

127. TĚŠITELOVÁ, M.: On some questions of spoken scientific discourses of men and women. (From the point of view of quantitative analysis of their vocabulary). PSML 6, 1978, 47-58.

128. TULDAVA, Ju.: Ob izmerenii trudnosti teksta [On the measurement of text-complexity]. Methodica IV. Acta et commentationes Universitatis Tartuensis. Tartu 1975, 102-120.

129. TULDAVA, Ju.: Quantitative relations between the size of text and the size of vocabulary. Strukturnaja i matematičeskaja lingvistika 4, 1977, 28-35.

130. WAITE, S.V.F.: Word position in Plautus: Interplay of verse ictus and word stress. In [10], 92-105.

131. XOVANOV, G.M.: Nekotorye voprosy količestvennogo analiza povtorenija slova v tekste [Some problems of the quantitative analysis of word repetition in text]. Andrjuščenko, V.M. (ed.): Issledovanija v oblasti vyčislitel'noj lingvistiki i lingvostatistiki. Moskva: Izdatel'stvo Moskovskogo universiteta, 1978, 41-58.

PSYCHOLINGUISTICS

132. AKSENENKO, O.N.: Ob odnoj metodike ocenki "ponimanija" soobščenija [On a method of estimating the "understanding" of a message]. Strukturnaja i matematičeskaja lingvistika 6, 1978, 3-9.

133. ASHCRAFT, M.H.: Property norms for typical and atypical items from 17 categories: A description and discussion. Memory and Cognition 6, 1978, 227-232.

134. BISANZ, G.L., LAPORTE, R.E., VESONDER, G.T., VOSS, J.V.: On the representation of prose: New dimensions. Journal of Verbal Learning and Verbal Behavior 17, 1978, 337-357.

135. DAVELAAR, E., COLTHEART, M., JONASSON, J.T.: Phonological recoding and lexical access. Memory and Cognition 6,

1978, 391-402.

136. FUCHS, A.: Fakten zum Cliff'schen Gesetz. Zeitschrift für experimentelle und angewandte Psychologie 25, 1978, 35-45.

137. GRAESSER, A.C.: Tests of a holistic chunking model of sentence memory through analyses of noun intrusions. Memory and Cognition 6, 1978, 527-536.

138. HABERLANDT, K., BINGHAM, G.: Verbs contribute to the coherence of brief narratives: Reading related and unrelated sentence triples. Journal of Verbal Learning and Verbal Behavior 17, 1978, 419-425.

139. HEALY, A.F., LEVITT, A.G.: The relative accessibility of semantic and deep-structure syntactic concepts. Memory and Cognition 6, 1978, 518-526.

140. HOFFMANN, J., KLIX, F.: Zur Prozeßcharakteristik der Bedeutungserkennung von sprachlichen Reizen. Zeitschrift für Psychologie 185, 1977, 315-374.

141. HOLYOAK, K.J., GLASS, A.L.: Recognition confusions among quantifiers. Journal of Verbal Learning and Verbal Behavior 17, 1978, 249-264.

142. HOUSTON, S.H.: Toward a differentiation of descriptive and psycholinguistic language models: Perceptual and orthographic/phonological analysis of spelling strategies. Journal of Psycholinguistic Research 7, 1978, 353-399.

143. JÄGER, S., FISCHER, V., MÜLLER, W., SCHMIDT, E., WOLF, M.: Warum weint die Giraffe? Ergebnisse des Forschungsprojektes "Schichtenspezifischer Sprachgebrauch von Schülern". Kronberg/Ts.: Scriptor 1978 (Monographien, Linguistik und Kommunikationswissenschaft Bd. 37, 669 pp.).

144. JOHNSON-LAIRD, P.N., GIBBS, G., MOWBRAY, J. de: Meaning, amount of processing, and memory for words. Memory and Cognition 6, 1978, 372-375.

145. JÖRG, S., HÖRMANN, H.: The influence of general and specific verbal labels on the recognition of labeled and unlabeled parts of pictures. Journal of Verbal Learning and Verbal Behavior 17, 1978, 445-454.

146. KLIX, F., HOFFMANN, J.: Process and structure in sentence verification tasks. Zeitschrift für Psychologie 186, 1978, 48-57.

147. LORCH, R.F., Jr.: The role of two types of semantic information in the processing of false sentences. Journal of Verbal Learning and Verbal Behavior 17, 1978, 523-537.

148. McCLOSKEY, M.E., GLUCKSBERG, S.: Natural categories: Well-defined or fuzzy sets? Memory and Cognition 6, 1978, 462-472.

149. MacLEOD, C.M., HUNT, E.B., MATHEWS, N.N.: Individual differences in the verification of sentence-picture relationships. Journal of Verbal Learning and Verbal Behavior 17, 1978, 493-508.

150. MacWHINNEY, B., BATES, E.: Sentential devices for conveying givenness and newness: A cross-cultural developmental study. Journal of Verbal Learning and Verbal Behavior 17, 1978, 539-558.

151. MARX, W.: Statistische Information und assoziative Bedeutung verschiedener Wortarten. Zeitschrift für experimentelle und angewandte Psychologie 25, 1978, 431-440.

152. MISHLER, E.G.: Studies in dialogue and discourse III. Utterance structure and utterance function in interrogative sequences. Journal of Psycholinguistic Research 7, 1978, 279-305.

153. MOESER, S.D.: Effects of questions on prose unitization. Journal of Experimental Psychology, Human Learning and Memory 1978/4, 290-303.

154. MUSTAFINA, S.Š.: Opyt formirovanija dvux vidov rečevoj dejatel'nosti - slušanija i govorenija na innostrannom jazyke [An experiment on the formation of two kinds of speech activity - speaking and hearing - in a foreign language]. Voprosy psixologii 1978/3, 118-122.

155. NEWMAN, J.E., DELL, G.S.: The phonological nature of phoneme monitoring: A critique of some ambiguity studies. Journal of Verbal Learning and Verbal Behavior 17, 1978, 359-374.

156. ORTONY, A., SCHALLERT, D.L., REYNOLDS, R.E., AUTOS, S.J.: Interpreting metaphors and idioms: Some effects of context on coprehension. Journal of Verbal Learning and Verbal Behavior 17, 1978, 465-477.

157. RIPS, L.J., SMITH, E.E., SHOBEN, E.J.: Semantic composition in sentence verification. Journal of Verbal Learning and Verbal Behavior 17, 1978, 375-401.

158. SHATZ, M.: On the development of communicative understandings: An early strategy for interpreting and responding to messages. Cognitive Psychology 10, 1978, 271-301.

159. SHOBEN, E.J., WESCOURT, K.T., SMITH, E.E.: Sentence verification, sentence recognition, and the semantic - episodic distinction. Journal of Experimental Psychology, Human Learning and Memory 1978/4, 304-317.

160. THEIOS, J., MUISE, J.G.: The word identification process in reading. Castellan, N.J., Jr., Pisoni, D.B., Potts, G.R. (eds.): Cognitive Theory, Vol. 2, 289-327. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1977.

161. TOWNSEND, D.J., BEVER, T.G.: Interclause relations and clausal processing. Journal of Verbal Learning and Verbal Behavior 17, 1978, 509-521.

162. WHALEY, C.P.: Word - nonword classification time. Journal of Verbal Learning and Verbal Behavior 17, 1978, 143-154.

163. YEKOVICH, F.R., WALKER, C.H.: Identifying and using referents in sentence comprehension. Journal of Verbal Learning and Verbal Behavior 17, 1978, 265-277.

HISTORICAL LINGUISTICS

164. BARON, N.S.: Language Acquisition and Historical Change. Amsterdam: North-Holland, 1977, 320 pp.

165. BERSCHIN, H., FELIXBERGER, J., GOEBL, M.: Französische Sprachgeschichte. München: Hueber, 1978, 377 pp.

166. JONES, W.J.: A quantitative view of Franco-German loan currency (1575-1648). Zeitschrift für Dialektologie und Linguistik, 45, 1978, 149-160.

167. KLIMEŠ, L.: On some quantitative aspects of the Czech sentence in the 18th century. PSML 6, 1978, 93-116.

168. KREJN, I.M.: Puti proniknovenija francuszkix zaimstvovanij v anglijskij literaturnyj jazyk XIX v. (Opyt statističeskogo issledovanija) [Ways of infiltration of French loan-words into the English written language of the nineteenth century (Attempt of a statistical investigation)]. Strukturnaja i matematičeskaja lingvistika 5, 1977, 63-69.

LANGUAGE VARIATION

169. BURDENJUK, G.M., Grigorevskij, V.M.: Jazykovaja interferencija i metody ee vyjavlenija [Linguistic Interference and Methods of its Study]. Kišinev: Štiinca, 1978, 127 pp.

170. KRAUS, J.: Stil' i obščestvennaja interakcija [Style and social interaction]. PSML 6, 1978, 117-128.

171. PIVOVAROVA, E.A.: K probleme vydelenija funkcional'nyx stilej russkogo literaturnogo jazyka [On the problem of distinguishing functional styles of the Russian standard language]. Strukturnaja i matematičeskaja lingvistika 6, 1978, 68-73.

DIALECTOLOGY

172. ALTMANN, G.: Zur Ähnlichkeitsmessung in der Dialektologie. Germanistische Linguistik 1977/3-4, 305-310.

173. ANCITIS, A.: Govor Akniste: Statistika i dinamika govora [The Subdialect of Akniste: Statistics and Dynamics of a Subdialect]. Riga 1977.

174. BERSCHIN, H.: Was leistet die elektronische Datenverarbeitung in der Dialektologie? Germanistische Linguistik 1977/3-4, 41-54.

175. FOSSAT, J.-L.: Vers un traitement automatique des données dialectologiques en dialectométrie. Germanistische Linguistik 1977/3-4, 311-334.

176. GOEBL, H.: Zu Methoden und Problemen einiger dialektometrischer Meßverfahren. Germanistische Linguistik 1977/3-4, 333-365.

177. GOEBL, H.: Analyse dialectométrique de quelques points de l'AIS. Lingue e dialetti nell' arco alpino occidentale. Torino: Centro Studi Piemontesi, 1978, 282-295.

178. NAUMANN, C.L.: Klassifikation in der automatischen Sprachkartographie. Germanistische Linguistik 1977/3-4, 181-210.

179. PŠENIČNOVA, N.N.: O primenenii statistiki pri izučenii govorov [On the application of statistics in the investigation of subdialects]. Govory territorij pozdnego zaselenija. Saratov 1977, 49-57.

180. SHUY, R.W., McGREEDY, L., FIRSCHING, J.: Toward the description of areal norms of syntax: some methodological suggestions. Germanistische Linguistik 1977/3-4, 367-395.

181. THOMAS, A.R.: A cumulative matching technique for computer determination of speech-areas. Germanistische Linguistik 1977/3-4, 275-288.

LANGUAGE ACQUISITION

182. AGER, D.E.: The importance of the word in the analysis of register. In [10], 194-207.

183. ALEKSEEV, P.M., GERMAN-PROZOROVA, L.P., GRIGOR'EV, V.P. et al.: Verojatnostno-statističeskij podxod k obučeniju inostrannym jazykam (osnovnye položenija) [The Probabilistic Approach to the Teaching of Foreign Languages (Basic Principles)]. Inženernaja lingvistika i optimiza-

cija prepodovanija inostrannyx jazykov. Leningrad 1976.

184. LISKER, L.: On learning a new contrast. Jazayery, M.A., Polomé, E., Winter, W. (eds.): Linguistic and Literary Studies. In Honor of A.A. Hill. The Hague: Mouton, 1978, Vol. I, 201-216.

All contributions to the CURRENT BIBLIOGRAPHY should be sent to Prof.Dr. Werner Lehfeldt, Universität Konstanz, Fachbereich Sprachwissenschaft, P.O. Box 5560, D.7750 Konstanz.

List of Contributors

Andrejev, Nikolaj, Institut jazykoznanija AN SSSR, Universitetskaja nab., dom 5 B-164, Leningrad, USSR

Altmann, Gabriel, Ruhr-Universität Bochum, Sprachwissenschaftliches Institut, Postfach 10 21 48, 4630 Bochum 1, West Germany

Grotjahn, Rüdiger, Ruhr-Universität Bochum, Seminar für Sprachlehrforschung, Postfach 10 21 48, 4630 Bochum 1, West Germany

Halstead, M.H., Purdue University, Department of Computer Sciences, West Lafayette, Indiana 47907, USA

Hantrais, Linda, University of Aston, Department of Modern Languages, Gosta Green, Birmingham B4 7ET, England

Harmatiuk, Sandra J., University of Notre Dame, Frechman Year of Studies, Notre Dame, IN 46556, USA

Itahashi, Shuichi, Electrotechnical Laboratory, Nagato-Cho, Chioda-Ku, Tokyo, Japan

Lehfeldt, Werner, Universität Konstanz, Fachbereich Sprachwissenschaft, Postfach 7733, D-7750 Konstanz, West Germany

McKinnon, Alastair, McGill University, Department of Philosophy, 1001 Sherbrooke Street West, Montreal, PQ, Canada H3A 1G5

Ratkowsky, David, CSIRO, Tasmanian Regional Laboratory, Stowell Avenue, Hobart, TAS. 7000, Australia

Sterner, Heinz, Querenburger Höhe 97, 4630 Bochum 1, West Germany

Strube, Gerhard, Universität München, Institut für Psychologie, Friedrichstraße 22, 8000 München 40, West Germany

Yokoyama, Shoichi, Electrotechnical Laboratory, Nagato-Cho, Chioda-Ku, Tokyo, Japan

INSTRUCTIONS TO AUTHORS

1. Contributions should be in English, German or French, but in English if at all possible.
2. MSS should be sent in duplicate to G.Altmann, Sprachwissenschaftliches Institut der Ruhr-Universität Bochum, Postfach 102148, D-4600 Bochum 1, West Germany, or to a member of the editorial board.
3. An abstract in English should be included at the beginning of the contribution.
4. Brief information about the author(s) (e.g. institution, position, main research interests) should be included on a separate page.
5. Authors are responsible for obtaining permission to publish material (figures, tables, etc.) for which they do not hold the copyright.
6. Each contributor to Glottometrika will receive one copy of the volume free of charge. Offprints may be purchased from the publisher if ordered at the proof stage.
7. Preparation of the manuscript. In order to facilitate the editing of the series the MSS should be submitted in final form, ready for photo-offset, observing the conventions delineated below. If this is not possible the MSS will be retyped at the expense of the editors.

Conventions:

Title: IBM Selectric typeface Orator if possible; 4 cm from top, centered.
Name and institutional affiliation (+ place if necessary): IBM Courier Italic (if possible), 2 cm under title.
Abstract: single-spaced, maximum of 20 lines, IBM Courier 10, 2 cm below name and institution.
Text: Courier 10, 1 1/2-spaced, 2 cm below the abstract should come the first line.
Paragraphs: indent 3 spaces.
Margins: 2 cm left as well as right, 3 cm above and below.
Chapter or paragraph titles, subtitles: any contrasting typeface.
Emphasized words within text as well as italics (i.e. for foreign words): underlined or in italics.
Notes and Bibliography should be listed separately at the end of the article.
The Bibliography should list the literature used in alphabetical order by author's last name, in the following form:

Brainerd, B., Graphs, topology and text. Poetics, 1977, 6, 1-14

Charpentier, C.M., La statistique linguistique pratiquée dans les pays anglo-saxons. Eléments de bibliographie (1970-1976). In: David, J. & Martin, R. (Eds.), Etudes de statistique linguistique. Paris: Klincksieck, 1977, 89-119

Titles of journals should not be abbreviated.
References both in the text and in the notes should take the form of the following examples: (cf. Brainerd, 1977:10); (David & Martin, 1977:12).
Figures and graphs should be ready for offset printing and be large enough to remain legible after a reduction in size of 50%. Furthermore they should be numbered consecutively, identified at the bottom with legends and, if possible, placed at an appropriate spot in the running text.
Tables should be numbered consecutively, provided with titles at the top and, if necessary, with legends (also at the top) and should be typed single-spaced. If at all possible, tables should be integrated into the running text, just as with figures and graphs.

Q U A N T I T A T I V E L I N G U I S T I C S

Appeared

Vol. I. Altmann, G. (Ed.), Glottometrika 1
Vol. II. Grotjahn, R., Linguistische und statistische Methoden in Metrik und Textwissenschaft
Vol. III. Grotjahn, R. (Ed.), Glottometrika 2

In Preparation

Alekseev, P., Statistische Lexikographie
Altmann, G., Statistik für Linguisten 1
Altmann, G., Lehfeldt, W., Einführung in die quantitative Phonologie 1
Andreev, N.D., Structural probabilistic analysis of language
Andreeva, L.D., Dominantenanalyse in synchroner und diachroner Sprachforschung
Arapov, M.V., Brainerd, B. (Eds.), Historical linguistics
Arapov, M.V., Guiter, H. (Eds.), Studies on Zipf's laws
Arapov, M.V., Cherc, M.M., Mathematische Modelle in der historischen Linguistik
Boy, J., Phonetische Dechiffrierung von Buchstaben
Frank, H. (Hrsg.), Interlinguistik
Goebl, H. (Ed.), Dialectology
Grotjahn, R. (Ed.), Hexameter
Grotjahn, R., Hopkins, E. (Eds.) Empirical Research on Language Teaching and Language Acquisition
Klein, W. (Ed.), Language variation
Lehfeldt, W. (Ed.), Quantitative Linguistics in Eastern Europe
Lesochin, M.M., Piotrowski, R.G., Mathematical models in quantitative linguistics
Matthäus, W. (Ed.), Psycholinguistics
Matthäus, W. (Ed.), Glottometrika 3
Moskowich, W., Semantics and statistics
Moskowich, W., Quantitative Linguistics in the Soviet-Union
Piotrowski, R.G., Text, Maschine, Mensch
Piotrowski, R.G., Bektaev, K.B., Piotrowskaja, A.A. Mathematische Linguistik
Rieger, B. (Ed.), Semantics. 2 Bd.
Sambor, J. (Ed.), Morphology
Skorochod'ko, E.F., Semantische Relationen im Lexikon und in Texten
Strauß, U., Struktur und Leistungsgrößen der Vokalsysteme
Valchov, G., Geometrical methods for scientific modelling

The series publishes
- Mixed volumes
- Monothematic volumes
- Textbooks
- Monographs
- Frequency dictionaries

Contributions can be sent to G. Altmann, Sprachwissenschaftliches Institut der RUB, Postfach 102148, 4630 Bochum, West Germany
Orders for individual volumes or the entire series should be directed to Studienverlag Dr. N. Brockmeyer, Querenburger Höhe 281, 4630 Bochum, West Germany.