QUANTITATIVE LINGUISTICS

Vol. 1

GLOTTOMETRIKA 1

edited

by

G. ALTMANN

Studienverlag Dr. N. Brockmeyer
Bochum 1978

QUANTITATIVE LINGUISTICS

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Glottometrika. – Bochum: Studienverlag Brockmeyer.

1. Ed. by G. Altmann. – 1978.
(Quantitative linguistics; Vol. 1)
ISBN 3-88339-030-5

NE: Altmann, Gabriel [Hrsg.]

ISBN 3 - 88339 - 030 - 5

Querenburger Höhe 281
4630 Bochum

EDITORIAL

In response to the growing intensity of quantification in all domains of linguistics, we propose a new series, called Quantitative Linguistics to be brought out at least 3 times a year at irregular intervals. This form has the advantage over a journal in that both articles and monographs can be published and that the interested reader can buy single volumes without having to subscribe to the total series.

The publication languages will be English, German and French. Authors, however, are asked, to use the first language mentioned, if possible.

The aims of the series are manifold. First, the series is intended to inform the reader about activities in all domains of linguistics. To this end, a rubric, Current Bibliography (of quantitative linguistics), will be established and brought out regularly. The authors publishing in the field of quantitative linguistics are invited to send the Editor bibliographical references about or copies of their most recent work. Special bibliographies concerning a particular linguistic discipline or results in a particular country, etc. are welcome. Further, research reports concerning the developments of a particular method, in a particular discipline, in a particular country, etc. will also be published.

Second, the amount of quantitative data available from languages does not seem at present to be adequate. Therefore, data collections in which uninterpreted results are simply presented with a short commentary will also be accepted. In the articles the raw data should always be presented rather than percentages. In frequency dictionaries, which could be published in separate volumes, the frequencies of all segmentally different forms should be given.

Third, in order to propagate quantitative methodology we plan the publication of textbooks on different mathematical disciplines such as probability theory, stochastic processes, information theory, statistics, special statistical methods (such as nonparametric statistics, multivariate statistics, path analysis, analysis of variance etc.), differential and difference equations, system theory, enumerative graph theory, curve fitting. Such textbooks should address themselves particularly to linguistic applications.

Fourth, current *evaluation of linguistic data*, which will no doubt be contained in the majority of contributions, should proceed with the aid of quantitative methods; therefore neither *pure* algebraic nor *pure* computer linguistics will be accepted. Currently these disciplines have enough publication organs of their own. It is our aim to stimulate the use of quantitative methods, which is best done when a method is included and explained directly in the article. Well known methodologies need not be repeated, but reference to sources where they can be found is essential. Greatest attention will be paid to the development of *new quantitative models* not yet available in other mathematized sciences.

Fifth, since the ultimate goal of all science is the building of theories, we will give preferred status to *theoretical work* concerning the latent mechanisms of language. The time seems right to investigate the deeper layers of language lying below the level of rules. It is our belief that on the one hand, language conceived as a self-regulating system and on the other, the psychological (cognitive) mechanisms of language production and language use can be grasped scientifically in the same manner as is done say in physics. The laws of language can be discovered neither by observing a competent speaker alone nor by analyzing a particular language. They must be derived theoretically. However, preliminary empirical generalizations made on the basis of many languages can suggest the form of a future law. Therefore both attempts at first approximations to laws based on empirical observation,

as well as theoretical derivations and their corroboration on the basis of a particular human language or experimentation in the psychological laboratory will be welcome. Several results in this direction have been published already, particularly in Russian. We intend to *translate* such works provided a satisfactory solution to copyright problems can be achieved.

Sixth, articles on the *methodology* of quantitative linguistics, its epistemological state, its position in the science and its development will be given interested consideration.

Finally, the series is intended to fulfill an *educational function*. Therefore, the authors should keep in mind that they are not only addressing specialists in their own areas of research but also interested "laymen", i.e. colleagues of other branches of linguistics. Hence authors should take pains to present their topic in a manner that is accessible to "newcomers" in their area.

The series will consist of three kinds of publications:

(1) ***Mixed omnibus volumes***, called GLOTTOMETRIKA, containing articles from all disciplines of linguistics. Contributions can be submitted any time.

(2) ***Monothematic omnibus volumes*** on specified subjects. Volumes devoted to the following themes are presently in preparation (here not in chronological order; the addresses of the editors are in brackets):

(a) *Arapov*, M.V., *Brainerd*, B. (Eds.), *Historical Linguistics*. (M.V. Arapov, VINITI AN SSSR, Otdel semiotičeskich problem informatiki, ul. Baltijskaja, 14, Moskva 125219, USSR; B. Brainerd, Dept. of Mathematics, University of Toronto, Toronto 5, Canada.)

(b) *Arapov*, M.V., *Guiter*, H. (Eds.), *Studies on Zipf's Laws*. (Arapov as above; H. Guiter, Université P. Valéry, B.P. 5043, 34032 Montpellier, Cedex, France.)

(c) Frank, H. (Ed.), Interlinguistik als Ausgangspunkt einer Reform des Fremdsprachenunterrichts nach kybernetisch-pädagogischen Gesichtspunkten. (H. Frank, Postfach 1567, 4790 Paderborn, West Germany.)

(d) Matthäus, W. (Ed.), Psycholinguistics. (W. Matthäus, Psychologisches Institut, Ruhr-Universität Bochum, Postfach 102148, 4630 Bochum 1, West-Germany.)

(e) Klein, W. (Ed.), Language Variation. (W. Klein, Max-Planck-Gesellschaft, Berg en Dalseweg 79, Nijmegen, Netherlands.)

(f) Rieger, B. (Ed.), Semantics. (B. Rieger, MESY, Heerlener Str. 30, 5100 Aachen, West Germany or Inst. voor algemeene Literatuurwetenschap, Herengracht 256, Amsterdam 1001, Netherlands.)

(g) Altmann, G., Grotjahn R. (Eds.), Hexameter. (R. Grotjahn, Seminar für Sprachlehrforschung, Ruhr-Universität Bochum, Postfach 102148, 4630 Bochum, West Germany.)

(h) Altmann, G. (Ed.), The law of Menzerath.

(i) Lehfeldt, W. (Ed.), Quantitative Linguistics in Eastern Europe. (W. Lehfeldt, Fachbereich Sprachwissenschaft, Universität, 7750 Konstanz, West Germany.)

(j) Sambor, J. Hérault, D. (Eds.), Morphology. (J. Sambor, ul. Rembielinska 5, m. 126, 03-343 Warszawa, Polandia; D. Hérault, Universitè de Paris VI, 91910 Saint Sulpice de Favieres, Paris, France.)

(k) Goebl, H. (Ed.), Dialectology. (H. Goebl, Romanisches Seminar der Universität Regensburg, Universitätsstr. 31, 8400 Regensburg, West Germany.)

Interested scholars and potential contributors should contact the editors of individual volumes.

(3) Monographs including textbooks and frequency dictionaries. In preparation are:

(a) Alekseev, P.M., Statistical lexicography.

(b) Andreev, N.D., Structural-probabilistic analysis of language.

(c) Andreeva, L.D., Dominantenanalyse in synchroner und diachroner Sprachforschung.

(d) Arapov, M.V., Kherc, M.M., Mathematical models in historical linguistics.

(e) Boy, J., Phonetische Dechiffrierung von Buchstaben.

(f) Lesokhin, M.M., Piotrowski, R.G., Mathematical models in quantitative linguistics.

(g) Moskowich, W., Semantics and statistics.

(h) Moskowich, W., Quantitative linguistics in the Soviet-Union.

(i) Piotrowski, R.G., Texts, machines, men.

(j) Piotrowski, R.G., Bektajev, K., Piotrowskaja, A., Mathematical linguistics.

(k) Skorokhodko, E.E., Semantische Relationen im Lexikon und in Texten.

(l) Valkhov, G., Geometrical methods for scientific modelling.

The series strives for interdisciplinarity and international cooperation. The members of the editorial board are linguists, mathematicians, physicists and psychologists.

Contributions and suggestions can be sent to any member of the editorial board or directly to the Editor G. Altmann, Sprachwissenschaftliches Institut, Ruhr-Universität Bochum, Postfach 102148, 4630 Bochum, West Germany.

The Editorial Board

C O N T E N T S

GENERAL

ALTMANN,G.(Bochum), Towards a theory of language 1

PHONOLOGY

LEHFELDT,W.(Konstanz), Zur Messung der phonetischen Lautdifferenz. Eine Begriffskritische Untersuchung 26

LEXICOLOGY

GEENS,D.(Leuven), On measurement of lexical differences by means of frequency 46

TEXT ANALYSIS

NOWAKOWSKA,M.(Warszawa), Some formal aspects of dynamics of dialogues 73

ALTMANN,G.(Bochum), Zur Anwendung der Quotiente in der Textanalyse 91

MOSKOWICH,W., CAPLAN,R.(Jerusalem), Distributive-statistical text analysis: A new tool for semantic and stylistic research 107

PSYCHOLINGUISTICS

MARX,W., SCHÜRER-NECKER,E.(München), Überlegungen zur Interpretation des Zipfschen Gesetzes am Beispiel der frühen Kindersprache 154

Diskussion: W.Matthäus 168

W.Marx,E.Schürer-Necker 170

SOCIOLINGUISTICS

RODRIGUE-SCHWARZWALD,O.(Ramat Gan), The influence of demographic variables on linguistic performance 173

BIBLIOGRAPHY

KAMINSKA,I.(Wrocław), Bibliography of quantitative linguistics in Poland 198

CURRENT BIBLIOGRAPHY 220

Towards a Theory of Language

G. Altmann, Bochum

1. The development of linguistics since the 19th century shows three major paradigms: the first concentrating on historical problems, the second, in which attention was concentrated on ontological and descriptive problems, and the recent one in which Chomsky aroused interest in simulation problems.

The historians of language observed changes in entities which they had, however, merely vaguely defined. They focused on grasping and describing the evolutionary processes as well as their conditions.

Since de Saussure linguistics has strived for a more precise determination of such vague entities (linguistic units) as well as on the formation of systems of concepts. One might say that the job at hand was to get the ontology "in order": it had to be determined exactly which units were present in language and which part these units played in individual languages. Linguistics in that period aimed at describing and classifying all linguistic units and their mutual rule-governed relations. Almost all textbooks of modern structural linguistics are written in the following way: the contain in the main lists of comprehensively explicated concepts which are only in extremely rare cases connected to one another by means of law-like statements (hypotheses).

Since Chomsky, a new ontology has been created and linguists have tried to simulate the process of language production. It has been said that many entities constructed by structuralists have no relevance for the new ontology. New entities have therefore been defined which fit better into the new descriptive mechanisms than the old ones. Linguists involved in such endeavours tend to maintain that they are constructing a theory. On closer inspection it can be seen that the term "theory" is being used with at least two different meanings.

First, the term is used to designate the mathematical apparatus by means of which the description, the analysis and the synthesis of linguistic entities can be accomplished; this is in other words the so-called theoretical linguistics which maintains the descriptive tools and may be considered a sort of subdiscipline of mathematics or logic.

A second interpretation of the term "theory" might be that it is a system of statements about the production and understanding of utterances. This interpretation is the one underlying the formulation of TARTAGLIA (1972: 16) who says that "the goal of a theory of natural language is the explication of the mechanism by which fluent speakers of a language are able to produce and understand the meaningful utterances of a given language". It is not unfair to call this conception the "official" view of the generativists. It is clear that the "theory of language" refers here actually to a kind of "theory of language behaviour". No matter what our attitude towards such a theory is, it should be evident that it cannot replace a theory of language conceived as a set of law-like statements about the connections and dynamics of the entities of language.

2. When one constructs a theory it is because among other things with the aid of this theory one wants to explain specific phenomena (for other purposes of a theory cf. e.g. SPINNER 1974: 120-123). These explanations, which are the common concern of all sciences, are the answer to a why-question. Reading the relevant literature on the philosophy of science it becomes evident that there are many kinds of such answers (cf. NAGEL, 1961; HEMPEL 1965; STEGMÜLLER 1969 etc.).

As far as language is concerned, one can pursue the explanatory efforts into very remote times. To show only a few examples: in myths one finds numerous answers to the question as to why there are several languages rather than only one; Plato attempts to explain in his Kratylos why words have a specified phonetic form; in recent times historical linguistics has tried to explore the causes of sound change; areal linguistics attempts to explain the evolutionary tendencies in one language with reference to the neighbouring languages; typology tries to grasp the whole mechanism of the construction, functioning and evolution of language and results automatically in the construction of a theory of language. Descriptive linguistics (often equated with linguistics per se) focuses in the main upon answering of how-questions and for this reason it is less successful in formulating explanatory hypotheses.

3. If one wants to construct a theory of language in the above mentioned sense, one must establish in advance some specifications and limitations. The main part of a theory consists of a system of hypotheses. Some of them are empirical (= tenable), i.e. they are corroborated by data; others are theoretical or (deductively) valid, i.e. they are derived from the axioms or theorems of a (not necessarily identical) theory with the aid of permitted operations. A scientific theory is a system in which some valid hypotheses are tenable and (almost) no hypotheses untenable (cf. GALTUNG 1967: 453). Evidently a theory is the better the more of its valid hypotheses are tenable. A theory generating deductively many untenable hypotheses can by no means be considered ideal.

Further the explanatory entities must be determined. In the majority of sciences these entities are laws. In linguistics one operates in addition with rules. We have shown previously (cf. ALTMANN 1977) that rules do not supply us with explanations of language phenomena; they merely represent concepts which enable us to describe certain language phenomena in idealized form. In any future theory of language one should therefore operate exclusively with laws, no matter whether deterministic or stochastic ones. To put it in a nutshell: no laws, no science (BUNGE 1967, I: 318).

The concepts with which one operates in a theory need not be either fruitful or exact at first. Their fruitfullness appears in the process of research and their exactness can be reduced to decidability. Besides, vague concepts are indeed the usual in linguistics (cf. e.g. TLP 2, 1966; ALTMANN 1972b). One must either eliminate them, make them exact, or introduce decision procedures for the determination of concept boundaries (cf. BELLMAN, ZADEH 1970).

The claims laid to a theory, the desiderata, the criteria of its quality (cf. BUNGE 1967, II: 346-361) are actually comparative and valuative criteria which, indeed, should serve to orient the user. At the beginning of research, however, one must deal with them cautiously. At the start one usually has merely a quite nebulous idea of how the properties of a theory will develop in the course of its elaboration. One does not know in advance how simple it will be, whether all its hypotheses will be testable, how many isolated domains it will join, how deep it will be, to what extent it will be able to stimulate further research, etc. Even contradictions can be tolerated at the beginning, a state of

affairs which is for instance tolerated in physics, as well.

4. When constructing a theory of language we proceed on the basic assumption that language is a self-regulating system all of whose entities and properties are brought into line with one another in some way or other. In spite of constant disturbances originating from inner change or outer influences, language must remain in a balance that insures its communicative efficiency. The disturbances may not always disappear but will always be compensated in order to meet the requirements made on language by the speaker. This basic assumption determines partially the path that should be taken in constructing a theory of language. Under different assumptions the theory would probably take a quite different shape.

The assumption of the general coordination of language phenomena with one another specifies the form of the explanative entities and of the desired explanation. The explanative entities should be laws of the form

$$Y = f(X_1, X_2, \ldots, X_n). \tag{1}$$

Here Y is a certain property and $X_1, X_2, \ldots, X_n$ symbolize those properties with which Y is connected in some way; $f(x_1, x_2, \ldots, x_n)$ represents a not yet defined function that expresses the relations of X_i $(i = 1,2,\ldots,n)$ to Y. The properties of language must be measurable, even when cases may arise where the scale contains only two values, 0 and 1 (i.e. a dichotomic scale with "yes-no-decisions"). The question as to why in a given language the examined property Y attains a given magnitude, can be answered by means of the Hempel-Oppenheim scheme with reference to all properties with which Y is connected as well as to the law (1). The explanation takes the (simplified) form

$$\frac{\begin{array}{l} x_{1L}, x_{2L}, \ldots, x_{nL} \\ y = f(x_1, x_2, \ldots, x_n) \end{array}}{y_L = c} \qquad (2)$$

where x_{iL} (i = 1,2,... n) are the antecedence conditions, i.e. the measured values of the property X_1 to X_n in the given language L. When these measured values are substituted in the second line of (2) then the value of the property Y_L, nameley $y_L = c$ results. We say that a phenomenon is explained if we derive it with the aid of laws from the antecedence conditions which are relevant for these laws.

5. The course we pursue will lead us to results through a combined causal, functional and teleological analysis. (For good surveys of these kinds of analysis cf. e.g. NAGEL 1961; HEMPEL 1965; STEGMÜLLER 1969; ESSER, KLENOVITS, ZEHNPFENNIG 1977 etc.)

There is a causal connection between two entities if it can be shown that one of them gives rise to the another, influences it, causes its change etc. The presence of a strong dynamic accent, for example, leads to the weakening of vowels in unstressed syllables and in extreme cases - when vowels disappear completely - to the increase of consonant clusters (cf. KOPPELMANN 1935: 85f.). Here one should speak of determination rather than about what is generally understood as causality (cf. BUNGE 1959).

One speaks of a functional explanation of a phenomenon if one succeeds in showing the function of the phenomenon in language: e.g. one can explain the degree of inflection in language either causally by referring to word length (see below) or by reference

to its function in language. The latter method has the disadvantage that in language it is possible that all functions can be fulfilled by different means. If one wants to hold fast to the functional explanation then one must show with a set of additional statements about the load of all other means of language why precisely the one in question assumed the particular function. These reflections demonstrate by the way certain problems connected with the predicting of changes. Linguistics is nevertheless in this regard in a better position than the social sciences, since the numer of means of expression available in language is limited.

If the overall status to which language ostensibly tends is to be said to account for the status of a particular phenomenon then one is dealing with a teleological explanation. Teleological explanations are at work for example when one explains the constitution of and the changes in a phoneme system by referring to the trend to symmetry, maximal discrimination, "pattern congruity" etc. (cf. MARTINET 1955; MOULTON 1962). Teleological explanations without a good theoretical background raise great difficulties, since the status to which language tends can be determined only vaguely. Success with teleological argument can be attained in the case of self-regulating systems. One can also replace teleological explanations by functional argumentations. We shall discuss these problems in another publication; here we restrict ourselves to the linguistic problems of the causal analysis. We want to demonstrate this problem programmatically in a simplified form and with roughly formulated hypotheses by means of an example and want to show that the main objective of language typology and quantitative linguistics (cf. ALTMANN, LEHFELDT 1973, ALTMANN 1972a) is identical with the building of a theory of language.

Linguistic hypotheses can be obtained for example by means of the following consideration: every language possesses a certain vocabulary in order to express different contents. The expression part of words consists of chains of phonemes, whose combinations are subjected to certain distributional restrictions. The following assumption seems to be plausible: the smaller the number of phonemes in the inventory of a language and the more stringent the distributional restrictions, the longer must be the words in order to avoid excessive homonymy (cf. HOCKETT 1958; SAPORTA 1966: 70, 72 note 15). However, language can protect itself against homonymy by using tones or accents. Therefore one can hypothetically maintain that there is <u>some</u> relation between the number, the distribution of phonemes, tones, accents, word length and homonymy. Further, one can conjecture that word length necessarily affects the degree of synthetism of a language: the longer the word the more morphemes it contains. But according to the "law of Menzerath" (cf. MENZERATH 1954) the longer the word the shorter its syllables. On the other hand the degree of synthetism influences the stringency of word order, the building of subordinate clauses, the compound words (cf. SKALIČKA 1966) and the sentence length which in turn correlates with the word length.

There are nowadays a large number of hypotheses of this kind (cf. USPENSKIJ 1965; VARDUL' 1969; GREENBERG 1966; SKALIČKA 1966). For our example it is sufficient to present the above-mentioned connections in a graph (cf. Fig. 1). It is easy to see that these connections were presented above alternatively in causal or functional or teleological formulations.

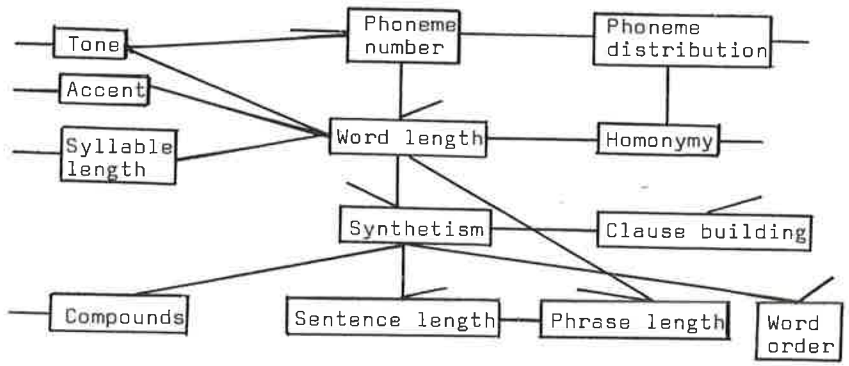

Fig. 1. Some connections between language phenomena

In Fig. 1 the relations of selected language properties are presented as an undirected graph. The direction of the influence has been left open. To each property an arc without the second vertex has been added which symbolizes all unmentioned or still unknown factors influecing it. We see that every property of language is connected with at least one other property; that there is a connection among all levels (e.g. number of phonemes: word order), even if it is indirect; that the arcs or sequences of arcs represent the conjectured laws of language; that these relations are mostly latent and only few of them can be ascertained by means of a single language.However, as soon as the laws have been formulated in form (1) they can be tested on any language.

Linguistics should attempt to discover all such connections in order to be able to build a comprehensive theory of language, i.e. a theory of the structure, of the functioning and of the change of language. The example of the natural sciences shows us how to proceed.

When building a theory of language we encounter several difficulties and problems which will be briefly discussed below: in § 6 problems of concepts, quantification and measurement; in § 7 we show two elementary quantitative models and in § 8 we shall try to point out the necessity of a teamwork. We restrict ourselves, as it were, to the most acute problems and leave most of the methodological problems of theory building aside (for an introduction cf. BUNGE 1967; GALTUNG 1967; ACHINSTEIN 1968; DUBIN 1969; OPP 1967; KÖNIG 1974 etc.).

6. The vertices of the graph in Fig. 1 represent linguistic concepts. In linguistics there are various dictionaries of terms as defined by the various linguistic schools. When building a theory of language we need not restrict ourselves to a particular school; however, it is important that the concepts fit into a given frame, i.e. that they are connected in one way or another with the other concepts of the theory and do not remain isolated.

Nature, society and language do not consist either of qualities or quantities; it is merely the concepts with which we try to grasp reality which are qualitative or quantitative (cf. ESSLER 1971: 65). Of course the vertices of the graph represent qualitative concepts which can, however, be readily quantified. Indeed they must be quantified if they are to be accepted in a theory of this kind.

Quantification is a prescription or definition, according to which one assigns some numbers to a concept. The inflection (Fl) of a language can be quantified e.g. as

$$Fl: = \frac{\text{Number of inflected words}}{\text{Number of all words}} \text{ in a text} \quad (3)$$

This is, however, not the only possible way; they are usually several possibilities of quantification (for several examples cf. ALTMANN, LEHFELDT 1973: 78-91). The decision as to which of the possible quantifications is to be considered as better depends on several circumstances. One must for example ask whether the index has mathematically "good" properties, whether it expresses what we aimed to express, whether it has a good interpretability, whether it is prolific etc. It is impossible to give a general prescription for good quantifications: the problems are of a very manifold nature. A principle that seems to be more or less serviceable in the beginning - when working with several variables - is to choose that index out of those possible which shows the most and highest correlations with others but this, too, is a somewhat circular and laborious procedure.

Quantification itself does not say anything about the presence, beginning, end etc. of the entity that should be counted or measured. Quantification is a rather abstract activity which is not identical with measurement. Measurement is a concrete activity by means of which we identify the entities (such are the usual methods of linguistics), count them and substitute the numbers in the quantifying formula, i.e. ascribe numbers to them. The result of quantification is a definition, a formula; the result of measurement is a number. Repeated measurements for example on different parts of a text or on different texts of a language show that measurement is subject to oscillations so that in practice we are confronted with the situation as presented in Fig. 2.

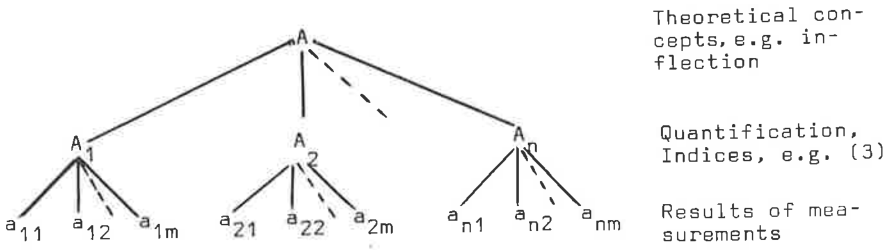

Fig. 2

In this situation one must find a way to make the results of measurement reliable and valid (for a survey cf. SCHEUCH, ZEHNPFENNIG 1974).

7. Notably the greatest difficulty for a linguist is to obtain a formula of form (1) or a whole system of such equalities representing hundreds of connections as suggested in Fig. 1. The difficulties increase as the number of variables in the net of connections increases.

Mathematics supplies us nowadays with sufficient means to enable us to begin constructing such a theory of language. These means include above all differential and difference equations, stochastic processes and statistical techniques working with several variables (path analysis, canonical analysis, multiple regression, factor analysis etc.). In the beginning the simplest way may well be the use of linear systems used frequently in the social sciences (cf. OPP, SCHMIDT 1976; HOLM 1977; HUMMELL, ZIEGLER 1976). They can help us to discover and describe the so-called causal structures.

In § 7.1 we sketch the empirical, "quasiinductive" process - a step-by-step setting up of hypotheses which cannot at first be derived from a set of axioms. They will, however, be immediately

tested and in case of a positive result they will be retained. That means that in this way the theory will at first be built up exclusively of tenable hypotheses. This is a quite customary process in research where the (preliminarily) best hypothesis is usually arrived at by a trial-and-error procedure. The axiomatization and the deducation of hypotheses from the axioms can be accomplished at a later date. When setting up hypotheses one naturally depends upon a background knowledge of immature theories (cf. POPPER 1973: passim). At the beginning only the rational axioms are present implicitly or explicitly, i.e. those which delimit the universe of discourse and determine the assumptions, the requirements and the restrictions. The structural axioms that describe the properties and the dynamics of the universe of discourse can be set up later on.

In § 7.2 we present an example of an elementary (quasi-) deductive procedure beginning with an assumption which will be formulated mathematically and thereafter tested for generality on particular cases. This is the more elegant procedure. Here, too, we depend upon some existing empirical findings which encourage us to venture the hypothesis. The difference to the linear approximation in § 7.1, which can hardly become part of an axiomatized theory, is that the result attained in § 7.2 has a greater chance to be derived later on from higher level hypotheses.

7.1 Here we shall show as an example how different factors act upon the formation of homonymy. First of all explicit hypotheses must be set up; here we must be content with the assumption that the number of phonemes in a phonemic inventory has an effect both upon phoneme associativity (distribution) and word length and that these two properties influence the building of homonyms. The directions of the effect left aside in Fig. 1 will be here indicated by arrows (directed arcs). In this way we obtain Fig. 3.

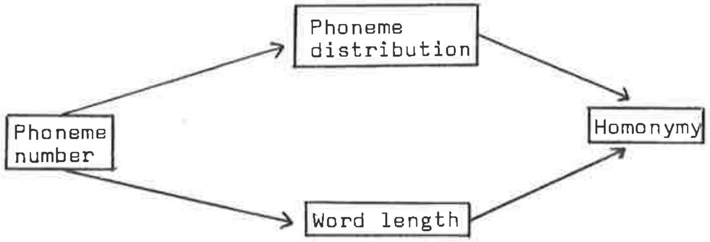

Fig. 3 A model of homonymy

Now we can set up a "causal model" as follows. Let

X_1 = number of phonemes

X_2 = distribution of phonemes

X_3 = word length

X_4 = homonymy

X_a, X_b, X_c = unknown factors

b_{ij} = path coefficients.

(Capital letters symbolize properties or variables, small letters symbolize their values.)

The scheme corresponding to Fig. 3 is shown in Fig. 4. This scheme represents a recursive linear model which can be expressed formally as

$$x_2 = b_{21}x_1 + b_{2a}x_a \tag{4.1}$$

$$x_3 = b_{31}x_1 + b_{3b}x_b \tag{4.2}$$

$$x_4 = b_{42}x_2 + b_{43}x_3 + b_{4c}x_c. \tag{4.3}$$

The model implies several presuppositions which will not be discussed here (cf. the literature above). The process of finding the

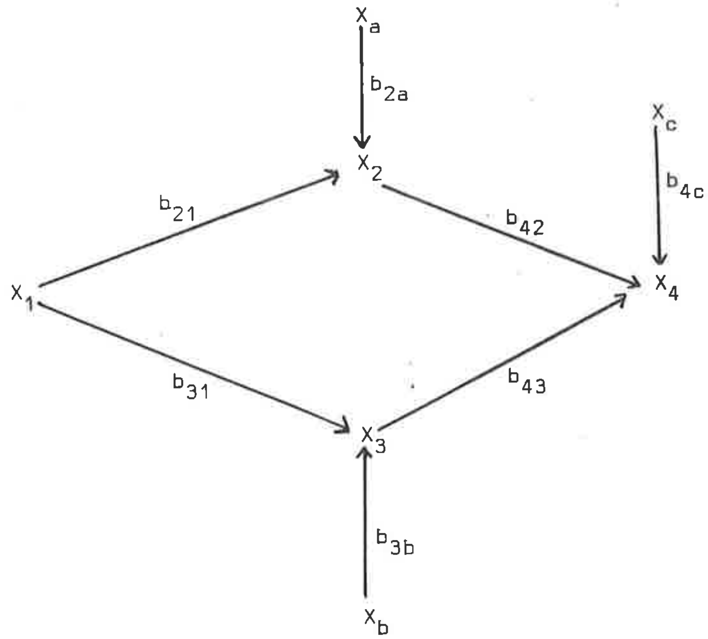

Fig. 4

coefficients b_{ij} is wearisome but not (always) difficult. One should bear in mind that x_i represents not one but many numerical results of measurements from different languages. We standarize them to normal variables by means of the transformation

$$u_i = \frac{x_i - \bar{x}}{s}$$

where $\bar{x}$ is the mean of all measurements and s is the standard deviation of X_i. The path coefficients b_{ij} can be replaced, too, by the standarized path coefficients, β_{ij}, so that instead of

(4.1) we write

$$u_2 = \beta_{21}u_1 + \beta_{2a}x_a \,. \qquad (4.1')$$

We assume that X_a and X_1 are not correlated and multiply (4.1') with the variable U_1, i.e.

$$u_2u_1 = \beta_{21}u_1^2 + \beta_{2a}x_au_1 \,.$$

Summing and dividing by n on both sides yields

$$\frac{1}{n}\sum_{i=1}^{n} u_{2i}u_{1i} = \beta_{21}\cdot\frac{1}{n}\sum_{i=1}^{n} u_{1i}^2 + \beta_{2a}\cdot\frac{1}{n}\sum_{i=1}^{n} x_au_{1i}$$

i.e.

$$r_{21} = \beta_{21}(1) + \beta_{2a}r_{1a} = \beta_{21}$$

since $\frac{\sum_i u_{1i}^2}{n} = 1$ (variance of u_1) and $r_{1a} = 0$ according to our assumption. Here r_{ij} is the correlation coefficient.

In the same way we obtain $\beta_{2a} = r_{2a}$ if we multiply (4.1') with X_a, sum and divide by n. Since r_{2a} cannot be computed directly from the data - X_a is unknown -, we multiply (4.1') with U_2, sum and divide with n and obtain

$$\beta_{2a} = \sqrt{1 - \beta_{21}^2}\,.$$

The path coefficients β_{ij} show the influence, the effect of a variable upon an other one. Here β_{21}^2 represents the explained variance of X_2 and β_{2a}^2 the unexplained one. In the same way one can compute the other β_{ij}s and b_{ij}s and put them into (4.1) till (4.3). These equalities have already the form (1) and may be used as preliminary lawlike statements in the construction of explanations and predictions.

We are aware of the fact that these procedures entail several difficulties: (a) first of all mathematical problems connected

with such systems of equations. They can be found in the literature (cf. e.g. DUNCAN 1975; FISHER 1966; HUMMELL, ZIEGLER 1976). Unfortunately, we shall need even more complicated systems at any higher level of research. (b) Linguistic problems will be apparent as soon as we begin to use these procedures. It could for instance be plausibly assumed that not only the number of phonemes in the inventory but also the size of particular phoneme classes act upon their distribution; that the word length acts upon the distribution, too; that one can expect immense difficulties in the domain of semantics, etc. We shall be forced to set up, to test and to reject a great number of different hypotheses. (c) Methodological difficulties are not to be underestimated. The above method seems to represent a purely inductive, phenomenological approach, as if we were merely setting up empirical generalizations which are not suitable for theory building. There are several objections to this view. First, we are concerned here with the discovery of latent mechanisms of language not with mere characterization or description of a particular surface property such as distribution of word length, grammar etc. Latent mechanisms cannot be discovered by empirical generalizations; they must be conjectured or grasped with hypotheses. The hypotheses do not follow from data but are our tentative creations. They are grounded in our linguistic background knowledge and must be empirically validated. Second, "in the first round" the relations between language properties will be approximated by means of linear additive hypotheses which have a number of weak points. Their rejection stimulates us to build new hypotheses; their acceptance is always only preliminary, so that sooner or later they will be replaced by more complicated hypotheses. A scientific hypothesis must be always correctable, otherwise it becomes a dogma and hinders progress. Third, these procedures help us to search for support in other existing theories, to build hypotheses about connections by analogy and to join in the considerable mathematization which has taken place in other sciences. Fourth, a linguistic "path model" carries a great deal of information which will be continuously complemented by our searching for unknown factors. Such a model has a good and wide testability and goes beyond the directly observable phenomena of language (such as phonetics, grammar). The unknown factors force us to search for new factors; low path coefficients force us to conjecture better hypotheses. This is why these procedures have a great heuristic power (cf. BUNGE 1967, I: 506-515). On the other hand path coefficients have the disadvantage of having a low linguistic interpretability which prevents such linear models from being incorporated into a high level theory. Therefore even in the very beginning one should set up more complicated hypotheses which contain few coefficients. Further, there is the danger that these models often yield only incomplete explanations (cf. HEMPEL 1965: 415-424). So one will be forced to search for further factors and sometimes to decide which part of the variance must be explained by a factor in order to be considered relevant. Fortunately the number of possible factors is not infinite in language; on the contrary, it is very limited so that the situation is not hopeless.

The detection of the interlacing of language phenomena and of their mutual influence affords us relatively deep explanations which, however, can by no means be considered as final. In the first stage only linguistic possiblities will be (exhaustively) examined, i.e. we restrict ourselves to the form of the language.

Later we can consider psychology, sociology, biology, etc. as well. We are interested not only in ascertaining correlations, though even such elementary findings represent progress in linguistics, but also in their explanations, in the detection of (theoretical) laws which bring them about.

7.2 The deductive procedure is no doubt more difficult but it has several advantages. It enables us to set up (theoretically) valid hypotheses which can be incorporated into a theory, but testing them empirically is laborious, since they have always many different interpretations. We shall illustrate the procedure of setting up a non-linear hypothesis by means of a simple example of the "Menzerath's law" which can be generally formulated as follows: "The longer a linguistic unit (constitute) the shorter its components (constituents)". This lawlike statement can be mathematically formulated in several ways. One can assume e.g. that the decrease rate of the constituent (L_1) is constant, i.e. $L_1'/L_1 = -a$ or, in other words, the decrease of the constituent (L_1) with respect to the constitute (L_2) is proportional to the length of the constituent, i.e.

$$\frac{dL_1}{dL_2} = -aL_1 \qquad (5)$$

where a is the coefficient of proportionality. This simple differential equation can be solved by means of separation of variables, i.e.

$$\int \frac{dL_1}{L_1} = -a\int dL_2$$

which gives

$$\ln L_1 = -aL_2 + c$$

and finally,

$$L_1 = e^{-aL_2 + c}$$

or

$$L_1 = be^{-aL_2} \qquad (6)$$

where $b = e^c$. Since linguistic units are almost always discrete, the use of a difference equation would have been perhaps better here. This theoretically derived formula has a number of interpretations according to the determination of L_i as sound length, syllable length, morpheme length, word length, phrase length, clause length, sentence length, etc. A definite confirmation of this universal hypothesis is not possible; one can, however, step by step reach better degrees of confirmation. To this end all possible units in all languages at our disposal must be examined. Our hypothesis is in agreement with the principle of least effort and can be brought in connection with some future theory of balance and/or optimality in language. The coefficients a and b may be determined empirically for a given case, but their theoretical determination is possible, too.

These two examples have we hope shown our conception of a future theory of language. Such a theory does not consist of a set of rules (low level hypotheses) prescribing how to build sounds, words, sentences (the domain of the "competent" speaker) but of a set of laws (high level hypotheses) about the latent dependencies in language. At the same time it seems evident that it is the quantitative part of mathematics that warrants progress along this line. Last but not least, a theory of this kind is the highest aim of typology (cf. ALTMANN, LEHFELDT 1973: 15).

8. Latent mechanisms cannot be detected when analyzing a

single language. The variables X_i in § 7.1 have (in the ideal case) only a single value for a single language from which no correlations between X_i and X_j can be computed. One must examine many languages. The (competent) speaker can neither pass judgements on latent dependencies nor can he detect them intuitively, since in a particular language they appear merely as corroborations of the corresponding stochastic laws. However, once detected, they can be used for predictions as well as explanations in a particular language; what is more, one can even hope that some kinds of rules in individual languages can be roughly derived or predicted from them.

The main obstacle (in this kind of theory building) lies in the circumstances that a single linguist can contribute measurements from few languages only - especially when the number of the variables increases - so that the task of setting up a theory of this kind can only be performed by a <u>team</u> of linguists. The basic prerequisites are already known: the "standard linguists" since de Saussure have defined a great number of concepts, the universalists and historical linguists have set up a number of hypotheses, the "quantitative linguists" have performed some quantifications and measurements, the mathematicians have supplied us with models and the philosophers have elaborated the problems of methodology, so that there is nothing to hinder the carrying out of a synthesis by a team.

9. Though it is possible for linguists to join the methodology of natural sciences (so that it is not necessary to build a separate one), one will be obliged to pay a greater attention to methodological problems, especially those of causality, functional analysis and laws, from the linguistic point of view.

Any such project must proceed in two ways: first, step-by-step empirical generalizations and their statistical systematization will lead to increasing configurations of properties (as in § 7.1), and second, dependencies of small numbers of variables will be derived deductively from hypothetical assumptions (as in § 7.2). Later on, one must strive to choose a small number of axioms and to derive the existing laws from them deductively. One will be obliged to use very general principles such as that of least effort, that of balance between redundancy and effectivity of information transmission by means of language, that of balance between notionality and emotionality of expressive means, further the capacity of man's memory, binary coding etc. For the elaboration and formalization of such a theory the cooperation of psychologists, biologists, physicists, engineers and mathematicians will be necessary.

The explanative and predictive capacity of a theory of this kind would be extremely important both for synchronic and diachronic linguistics. Psycholinguistics would with such a theory be able to enlarge its stock of hypotheses, the areal linguistics would obtain a more secure base, typology and universalism could be unified, even artificial world languages would obtain a more concrete comparative basis, etc. It is not possible to foresee all the applications and consequences of any new theory; nevertheless, we are positive that the final effect of the endeavour here described would surely be the establishment of a very comprehensive theory of language.

References

ACHINSTEIN, P., Concepts of science. A philosophical analysis. Baltimore-London, Hopkins 1968.

ALTMANN, G., Status und Ziele der quantitativen Sprachwissenschaft. In: JÄGER, S. (Ed.), Linguistik und Statistik. Braunschweig, Vieweg 1972a, 1-9.

ALTMANN, G., Zur linguistischen Unbestimmtheit. Linguistische Berichte 22, 1972b, 74-79.

ALTMANN, G., Sprachregeln und Erklärung. Linguistische Berichte 50, 1977, 31-37.

ALTMANN, G., LEHFELDT, W., Allgemeine Sprachtypologie. Prinzipien und Meßverfahren. München, Fink 1973.

BELLMAN, R., ZADEH, L.A., Decision-making in a fuzzy environment. Management Science 17, 1970, 141-164.

BUNGE, M., Causality. Cambridge, Mass., Harvard UP 1959.

BUNGE, M., Scientific research I, II. Berlin, Springer 1967.

DUBIN, R., Theory building. New York, The Free Press 1969.

DUNCAN, O.D., Introduction to structural equation models. New York, Academic Press 1975.

ESSER, H, KLENOVITS, K., ZEHNPFENNIG, H., Wissenschaftstheorie 2. Stuttgart, Teubner 1977.

ESSLER, W.K., Wissenschaftstheorie II. Freiburg-München, Alber 1971.

FISHER, F.M., The identification problem in econometrics. New York, McGraw-Hill 1966.

GALTUNG, J., Theory and methods of social research. Oslo, Universitetsforlaget 1967.

GREENBERG, J.H. (Ed.), Universals of language. Cambridge, Mass., The M.I.T. Press 1966[2].

HEMPEL, C.G., Aspects of scientific explanation. New York, The Free Press 1965.

HOCKETT, Ch. F., A course in modern linguistics. New York, McGraw-Hill 1958.

HOLM, K. (Ed.), Die Befragung 5. München, Francke 1977.

HUMMELL, H.J., ZIEGLER, R. (Eds.), Korrelation und Kausalität I-III. Stuttgart, Enke 1976.

KÖNIG, R. (Ed.), Handbuch der empirischen Sozialforschung I-IV. Stuttgart, Enke 1974[3].

KOPPELMANN, H.L., Ursachen des Lautwandels. Leiden, Sijthoff 1939.

MARTINET, A., Économie des changements phonêtiques. Berne, Francke 1955.

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

MOULTON, W.G., Dialect geography and the concept of phonological space. Word 18, 1962, 23-32.

NAGEL, E., The structure of science. London, Routledge & Kegan Paul 1961.

OPP, K.D., Methodologie der Sozialwissenschaften. Einführung in die Probleme ihrer Theoriebildung. Reinbek, Rowohlt 1976[2].

OPP, K.D., SCHMIDT, P., Einführung in die Mehrvariablenanalyse. Reinbek, Rowohlt 1976.

POPPER, K.R., Objektive Erkenntnis. Hamburg, Hoffman und Campe 1973.

SAPORTA, S., Phoneme distribution and language universals. In: GREENBERG 1966, 61-72.

SCHEUCH, E.K., ZEHNPFENNIG, H., Skalierungsverfahren in der Sozialforschung, In: KÖNIG 1974, 3a: 97-203.

SKALIČKA, V., Ein "typologisches Konstrukt". Travaux Linguistiques des Prague 2, 1966, 157-163.

SPINNER, H., Pluralismus als Erkenntnisfortschritt. Frankfurt, Suhrkamp 1974.

STEGMÜLLER, W., Wissenschaftliche Erklärung und Begründung. Berlin, Springer 1969.

TARTAGLIA, P., Problems in the construction of a theory of natural languages. The Hague, Mouton 1972.

TLP: Travaux Linguistique de Prague.

USPENSKIJ, B.A., Strukturnaja tipologija jazykov. Moskva, Nauka 1965.

VARDUL', I.P. (Ed.), Jazykovye universalii i lingvističeskaja tipologija. Moskva, Nauka 1969.

Zur Messung der phonetischen Lautdifferenz
Eine begriffskritische Untersuchung

Werner Lehfeldt, Konstanz

Eine wissenschaftliche Disziplin, die bestrebt ist, in ihren Gegenstandsbereich immer tiefer einzudringen, muß fortlaufend die von ihr gebrauchten Mittel, d.h. ihre Methoden, Modelle, Instrumente usw., überprüfen und, falls nötig, verändern. Wenn wir von einer "Vertiefung" der Einsichten über einen Objektbereich sprechen, so soll mit diesem Ausdruck zweierlei bezeichnet werden:

1. eine immer vollständigere und präzisere Erfassung der untersuchten Gegenstände;
2. die Verbesserung der Aussagen über regelhafte Zusammenhänge zwischen Eigenschaften der untersuchten Gegenstände und die Entdeckung neuer Zusammenhänge dieser Art, die es gestatten sollen, bekannte Erscheinungen besser als bisher zu erklären und neue Erscheinungen mit größerer Zuverlässigkeit vorherzusagen.

Zu den Mitteln einer Wissenschaft, die ständig zu prüfen sind, gehören auch die jeweils gebrauchten Systeme von Begriffen, mit denen Aussagen formuliert werden. Die Antwort auf die Frage, wie genau und wie vollständig die Beschreibung der Gegenstände ist, und welchen Grad der Zuverlässigkeit Erklärungen und Voraussagen aufweisen, hängt zu einem guten Teil davon ab, welche Art von Begriffen eine Wissenschaft verwendet und wie sie ihre Begriffe verwendet. Dies zeigt die Geschichte der Wissenschaften mit hinreichender Deutlichkeit. So sind die Erfolge der Physik unmittelbar mit der Einführung einer quantitativen Begriffs-

sprache verknüpft gewesen.

Die wichtige Rolle, die unter den Mitteln einer Wissenschaft die Begriffe spielen, macht einsichtig, daß die Überprüfung dieser Mittel die Begriffskritik einschließen muß. In der vorliegenden Arbeit wollen wir eine solche Begriffskritik an sprachwissenschaftlichen Beispielen veranschaulichen. Bevor wir uns diesen Beispielen zuwenden, erscheint es zweckmäßig, etwas präziser darzulegen, welche Aufgaben die Begriffskritik erfüllen muß. Wir wollen drei Punkte näher betrachten:

1. Die erste Aufgabe besteht darin, festzustellen, welcher Art die von einer Disziplin jeweils gebrauchten Begriffe sind: Handelt es sich um klassifikatorische (qualitative), um komparative (topologische) oder um quantitative Begriffe? (Vgl. STEGMÜLLER 1970, 19-109). Der Versuch, auf diese Frage eine Antwort zu finden, führt nicht selten zu der Einsicht, daß den Wissenschaftlern nicht klar ist, mit was für einer Art von Begriffen sie umgehen. In derartigen Fällen ist es notwendig, eine Begriffsrekonstruktion zu versuchen, d.h. herauszufinden, um was für eine Begriffsart es sich möglicherweise handeln könnte (für Begriffe der Sprachtypologie vgl. LEHFELDT, ALTMANN 1975). Finden wir beispielsweise in einer typologischen Untersuchung die Aussage, eine Sprache <u>x</u> sei agglutinierender als eine Sprache <u>y</u>, so könnte man vermuten, daß dieser Aussage die Anwendung eines komparativen Begriffes zugrundeliegt. Es handelt sich zunächst wirklich um nicht mehr als eine Vermutung, da der Gebrauch eines grammatischen Komparativs nicht mit der Verwendung eines komparativen Begriffes verwechselt werden darf.

2. Wenn die erste Frage beantwortet bzw. wenn eine entsprechende Vermutung ausgesprochen ist, dann ist zu fragen, ob die Begriffe immer korrekt angewendet worden sind, das soll heißen, in einer

Weise, die sich mit den jeweiligen Adäquatheitsbedingungen vereinbaren läßt. Verdeutlichen wir dies an dem einfachsten denkbaren Fall: Wenn in einer sprachtypologischen Abhandlung alle betrachteten Sprachen als entweder flektierend oder agglutinierend oder amorph eingestuft werden, dann haben wir es offensichtlich mit klassifikatorischen Begriffen zu tun, und wir müssen daher fragen, ob die Adäquatheitsbedingungen für klassifikatorische Begriffe erfüllt sind. Wir haben zu prüfen, ob diese Begriffe eine erschöpfende und disjunkte Klassifizierung der bis zu dem Zeitpunkt der Arbeit bekannten natürlichen Sprachen gestatten. Die in dem letzten Satz enthaltene zeitliche Einschränkung ist nicht unwesentlich, da, wie alle anderen Begriffe, so auch die klassifikatorischen Begriffe, sofern sie unabhängig voneinander definiert sind, auf empirischen Hypothesen beruhen, die möglicherweise in der Zukunft widerlegt werden. Für das von uns herangezogene Beispiel läßt sich ohne Schwierigkeiten zeigen, daß zumindest die zweite Adäquatheitsbedingung, nach der die Klasseneinteilung disjunkt sein muß, nicht erfüllt ist, da es zahlreiche Sprachen gibt, die sowohl Flexion wie Agglutination kennen (vgl. SAPIR 1921, 116).

Wie wir sehen, erfüllt die Begriffskritik die Aufgabe, die Wissenschaftler zur korrekten Konstruktion und zum korrekten Gebrauch der Begriffe zu veranlassen. Wie wichtig diese Aufgabe ist, kann man sich am besten klarmachen, wenn man sich vor Augen hält, warum im einzelnen auf einen korrekten Begriffsgebrauch so großer Wert gelegt werden muß. Erläutern wir diesen Punkt zunächst am Beispiel der quantitativen Begriffe.

Die Einführung quantitativer Begriffe in einen Gegenstandsbereich ist gleichbedeutend der Metrisierung dieses Bereiches.

Wenn eine Metrisierung geglückt ist, dann kann man Messungen vornehmen, d.h. man kann ein empirisches Relationensystem strukturgetreu in ein numerisches Relationensystem abbilden. Die Messungen ergeben Zahlen, auf die bestimmte mathematische Operationen angewendet werden können, deren Ergebnisse wiederum in Aussagen über das interessierende empirische Relationensystem rückübersetzt werden sollen. Natürlich sind die mathematischen Operationen nur dann empirisch belangvoll, können nur dann zu zutreffenden Erklärungen und/oder Voraussagen führen, wenn die Messungen korrekt waren. Die Korrektheit der Messungen hängt ihrerseits in starkem Maße von der Korrektheit der Metrisierung ab. Leider wird diesem Umstand oft nicht die ihm gebührende Aufmerksamkeit geschenkt. Man muß sich ganz klarmachen, daß mathematische Modelle, mögen sie auch noch so kompliziert und elegant aussehen, und mögen sie dem "nichteingeweihten" Leser auch noch so imponieren, empirisch vollkommen uninteressant sind, falls sie auf inkorrekten Messungen beruhen.

Was hier über die quantitativen Begriffe gesagt wurde, gilt zu einem guten Teil auch von den komparativen Begriffen, nicht zuletzt deshalb, weil man bei der Einführung vieler wichtiger quantitativer Begriffe an bereits vorhandene komparative Begriffe anknüpft. Auch komparative Begriffe erlauben es, empirische Relationensysteme in numerische abzubilden. Über die Struktur dieser Abbildung muß man präzise Vorstellungen haben, um in der Lage zu sein, die erlaubten mathematischen Operationen auszuwählen. Wir kommen auf die in diesem Abschnitt genannten Punkte noch im einzelnen zurück.

3. Schließlich hat die Begriffskritik auch noch die Frage zu untersuchen, ob die wissenschaftlichen Begriffe fruchtbar sind,

d.h. ob sie es gestatten, Gesetzmäßigkeiten zu erkennen. Auch dieser Punkt wird oft nicht hinreichend erkannt. So werden beispielsweise oftmals Klassifizierungen vorgenommen, ohne daß man sich fragt, wozu solche Einteilungen gut sein sollen. In letzter Zeit läßt sich aber doch beobachten, daß bei der Konstruktion linguistischer Klassifikationen Fruchtbarkeitserwägungen stärker als bisher eine Rolle spielen (vgl. CARVELL, SVARTVIK 1969, 34f.). Solche Erwägungen können u.U. auch dazu anregen, Begriffe primitiver Struktur durch komplizierte zu ersetzen.

Nach diesen allgemeinen Erörterungen wollen wir uns der Kritik einiger in der Linguistik gebrauchter Begriffe zuwenden und dabei insbesondere die unter den Punkten 1 und 2 genannten Aufgaben der Begriffskritik im Auge behalten. Wir werden im einzelnen zwei Vorschläge betrachten, die artikulatorische Differenz von je zwei Lautsegmenten zu bestimmen und in einer Zahl auszudrücken. Diese Vorschläge stammen von PETERSON und HARARY (1961) sowie GERŠIĆ (1971). Besonders der Vorschlag von PETERSON und HARARY ist in der Linguistik recht bekannt geworden und soll daher am ausführlichsten analysiert werden.

Bei der Bestimmung der artikulatorischen Differenz von zwei Lautsegmenten handelt es sich um eine abgeleitete oder sekundäre Bestimmung (der Ausdruck "Messung" wird hier absichtlich vermieden), sie ist eine Funktion mehrerer zugrundeliegender fundamentaler oder primärer Begriffe, die sich auf diejenigen einzelnen artikulatorischen Merkmale beziehen, die bei der Differenzbestimmung nach dem Willen des Sprachwissenschaftlers berücksichtigt werden sollen. Je größer die Zahl der Merkmale ist, die wir an einem Segment untersuchen und bezüglich derer wir zwei Segmente vergleichen, ein um so besseres Differenzmaß werden wir konstruieren können. Ein praktisch recht befriedigendes

Maß würde sich beispielsweise aufstellen lassen, wenn es gelänge, alle die von PIKE (1943) angegebenen und beschriebenen Lautparameter zu erfassen.

Die artikulatorische Differenz von Lautsegmenten kann selbstverständlich nur dann korrekt bestimmt werden, wenn die zugrundegelegten fundamentalen Begriffe korrekt konstruiert sind und korrekt angewandt werden. Eben die Struktur dieser Begriffe ist es, die wir in den von uns ausgewählten Beispielsfällen im einzelnen prüfen wollen.

Bis heute ist es nicht gelungen, artikulatorische Lautmerkmale in hinreichender Zahl und in hinreichendem Maße zu metrisieren, d.h. zu quantifizieren, es sind ja auch "nicht einmal in Ansätzen Meßmethoden entwickelt worden, die eine Bestimmung der für eine adäquate Beschreibung der Artikulationsbasis erforderlichen Elemente im physiologischen Bereich ermöglichen könnten" (KELZ 1974, 217). Damit entfällt eine wichtige Voraussetzung für die Messung der artikulatorischen Differenz von Lautsegmenten.

Man kann aber bereits auf der Grundlage einer Ordinalskala zu einer approximativen Bestimmung der artikulatorischen Differenz gelangen, und wir wollen sehen, ob das Differenzmaß von PETERSON und HARARY und das von GERŠIĆ möglicherweise auf solchen Skalen beruhen. Zu diesem Zweck sei zunächst der Vorschlag von PETERSON und HARARY ausführlicher beschrieben.

PETERSON und HARARY (1961) haben ein Maß der phonetischen, genauer der artikulatorischen Differenz zwischen zwei Lautsegmenten entwickelt. Jedes Segment wird mit Rücksicht auf fünf Merkmale beschrieben, die hier Parameterklassen heißen. Drei Parameterklassen, "secondary articulation", "laryngeal action" und "air direction", sind Vokalen und Konsonanten gemeinsam, zwei Parameterklassen, "tongue hump relative to pharynx" und "tongue height", dienen der Charakterisierung nur der Vokale, zwei andere, "place of articulation" und "manner of articulation", der Charakterisierung nur der Konsonanten. Daraus wird ersichtlich, daß mit dem Maß von PETERSON und HARARY die phonetische Differenz zwischen einem Vokal und einem Konsonanten nicht erfaßt werden kann.

Die Parameterklassen beschreiben jedes Segment im Hinblick auf die physiologischen Mechanismen und Prozesse, die an der Erzeugung von Lauten beteiligt sind. Sie sind hierarchisch gestuft, da bestimmten Parameterklassen eine größere Bedeutung für die Bildung eines Vokals oder eines Konsonanten beigemessen wird als anderen. Wie diese Hierarchiebildung begründet wird, soll uns noch beschäftigen.

Jede Parameterklasse ist unterteilt in verschiedene Ränge, die hier Parameter heißen. Die Zahl der Parameter kann von Klasse zu Klasse schwanken. Sie beträgt 10 in der Klasse "place of articulation" und 2 beispielsweise in der Klasse "air direction". Jedem Parameter wird eine Zahl zugeschrieben, wobei eine Rangordnung angestrebt wird, wie die Autoren selbst einige Male andeuten. Auch auf die Begründung für die Rangordnungen innerhalb der einzelnen Parameterklassen werden wir unten ausführlicher eingehen.

Die Parameterklassen und die einzelnen Parameter werden von PETERSON und HARARY wie folgt spezifiziert:

V o k a l e

1. Secondary articulation
 .0 Spread
 .1 Rounded
 .2 Rilled (curled)
 .3 Lateralized
 .4 Retroflexed
 .5 Velarized
 .6 Nasalized
 .7 Faucalized

2. Laryngeal action
 .0 Whispered
 .1 Breathy
 .2 Clear
 .3 Laryngealized

3. Air direction
 .0 Egressive
 .1 Ingressive

4. Tongue hump relative to pharynx
 .0 Front
 .1 Front central
 .2 Central
 .3 Back central
 .4 Back

5. Tongue height
 .0 Highest
 .1 High
 .2 High-mid
 .3 Mid
 .4 Low-mid
 .5 Low
 .6 Lowest

K o n s o n a n t e n

1. Secondary articulation
 .0 Spread
 .1 Labialized
 .2 Palatalized
 .3 Rilled (curled)
 .4 Lateralized
 .5 Retroflexed
 .6 Velarized
 .7 Nasalized
 .8 Faucalized

2. Laryngeal action
 .0 Voiceless
 .1 Voiced

3. Air direction
 .0 Egressive
 .1 Ingressive

4. Place of articulation
 .0 Bilabial
 .1 Labio-dental
 .2 Dental
 .3 Alveolar
 .4 Alveolo-palatal
 .5 Palatal
 .6 Velar
 .7 Uvular
 .8 Pharyngeal
 .9 Glottal

5. Manner of articulation
 .0 Nasal
 .1 Liquid
 .2 Fricative (sibilant)
 .3 Trill (flap)
 .4 Plosive

Jedes Segment wird aufgefaßt als Menge von Vokal- bzw. Konsonantenparametern. Jeder Parameter wird durch eine Dezimalzahl dargestellt, wobei die erste Zahl die betreffende Klasse, die zweite die betreffende Ausprägung bezeichnet.

Beispiel: Dem Lautsegment [p] entspricht folgende Parametermenge: [p] = {secondary articulation: spread; laryngeal action: voiceless; air direction: egressive; place of articulation: bilabial; manner of articulation: plosive}

Dem entspricht folgender Vektor:

[p] = [1.0; 2.0; 3.0; 4.0; 5.4]

Auf der Grundlage der skizzierten Voraussetzungen entwickeln PETERSON und HARARY ein Maß der phonetischen Differenz, in dem die folgenden Faktoren Berücksichtigung finden:

a. die Rangordnung der Parameterklassen;

b. die Zahl der Parameter, hinsichtlich derer sich zwei Segmente in einer Parameterklasse unterscheiden;

c. die Rangunterschiede zwischen je zwei Segmenten innerhalb einer Parameterklasse.

Das Maß der phonetischen Differenz zwischen zwei Segmenten i und j bezüglich einer Parameterklasse k ist definiert als

$$D_k(i,j) = k\varphi_k(i,j) + \frac{n[S_k(i) \oplus S_k(j)]}{10} + \delta_k(i,j)\,\frac{|S_k(i) \oplus S_k(j)|}{10\,T_k(i,j)}$$

Erläuterungen:

K = 1, 2, 3, 4, 5 symbolisiert die Rangzahlen der einzelnen Parameterklassen

$S_k(i)$ = Menge der Parameter des Segments i in der k-ten Parameterklasse

Beispiel:

$S_1(k')$ = {palatalized}

$S_2(t)$ = {voiceless}

$S_3(m)$ = {egressive}

$S_5(s)$ = {fricative}

$S_4(t)$ = {alveolar}

$S_5(k)$ = {plosive}.

$$\varphi_k(i,j) = \begin{cases} 0 & \text{wenn } S_k(i) = S_k(j) \\ 1 & \text{wenn } S_k(i) \neq S_k(j) \end{cases}$$

Beispiel:

$\varphi_5(p,t) = 0$,

weil $S_5(p) = S_5(t)$ = {plosive}.

$\varphi_4(p,t) = 1$,

weil $S_4(p)$ = {bilabial} $\neq S_4(t)$ = {alveolar}

$\varphi_1(k',g) = 1$,

weil $S_1(k')$= {palatalized} $\neq S_1(g)$ = {velarized}.

$n[S_k(i) \oplus S_k(j)]$ = Anzahl der Parameter in der symmetrischen Differenz von $S_k(i)$ und $S_k(j)$, d.h. Zahl der Parameter, die für i und j in der k-ten Klasse unterschiedlich sind.

Beispiel:

$n[S_4(p) \oplus S_4(t)] = 2$

denn es ist

$S_4(p)$ = {bilabial}

$S_4(t)$ = {alveolar}

$S_4(p) \oplus S_4(t)$ = {bilabial, alveolar}.

$$\delta_k(i,j) = \begin{cases} 0 & \text{wenn } S_k(i) = \emptyset \text{ oder } S_k(j) = \emptyset \\ 1 & \text{wenn } S_k(i) \neq \emptyset \text{ oder } S_k(j) \neq \emptyset \,. \end{cases}$$

Beispiel:

$\delta_4(p,t) = 1,$

weil $S_4(p) = \{\text{bilabial}\} \neq \emptyset$ und $S_4(t) = \{\text{alveolar}\} \neq \emptyset$.

$|S_k(i) \oplus S_k(j)|$ = absoluter Wert des Unterschiedes zwischen allen Paaren der Dezimalindizes der Parameter von i und j in der k-ten Parameterklasse; d.h., dieser Ausdruck symbolisiert den Rangunterschied zwischen i und j in der k-ten Klasse.

Beispiel:

$|S_4(b) \oplus S_4(k)| = |4.0 - 4.6| = 0.6.$

$T_k(i,j) = n[S_k(i) - S_k(j)]n[S_k(j) - S_k(i)]$

$T_k(i,j)$ symbolisiert die Zahl der Glieder in $|S_k(i) \oplus S_k(j)|$.

Beispiel:

$T_4(b,k) = 1$

denn es ist

$S_4(b) - S_4(k) = \{\text{bilabial}\} \rightarrow n[S_4(b) - S_4(k)] = 1$

$S_4(k) - S_4(b) = \{\text{velar}\} \rightarrow n[S_4(k) - S_4(b)] = 1.$

Die totale phonetische Differenz D(i,j) ist gleich der Summe der phonetischen Differenzen zwischen i und j in allen Parameterklassen:

$$D(i,j) = \sum_{k=1}^{5} D_k(i,j)$$

Beispiel: Zu berechnen ist die totale phonetische Differenz zwischen den Segmenten [t] und [m].

$$
\begin{array}{llr}
D_1(t,m) = 1x0 + 0 + 0 & & = 0 \\
D_2(t,m) = 2x1 + \frac{2}{10} + 1\frac{0.1}{10x1} & = 2 + 0.2 + 0.01 & = 2.21 \\
D_3(t,m) = 3x0 + 0 + 1\frac{0}{10x1} & & = 0 \\
D_4(t,m) = 4x1 + \frac{2}{10} + 1\frac{0.3}{10x1} & = 4 + 0.2 + 0.03 & = 4.23 \\
D_5(t,m) = 5x1 + \frac{2}{10} + 1\frac{0.4}{10x1} & = 5 + 0.2 + 0.04 & = 5.24 \\
\hline
D(t,m) & & = \underline{\underline{11.68}}
\end{array}
$$

Wir wollen nun die Voraussetzungen des Differenzmaßes von PETERSON und HARARY eingehender betrachten.

Die phonetische Differenz wird mit Rücksicht auf Merkmale ("Parameterklassen") berechnet, die augenscheinlich auf einer Ordinalskala gemessen werden sollen (vgl. S. 144). Wenn wir uns z.B. die vierte Parameterklasse für Konsonanten, "place of articulation", anschauen, so können wir jeden einzelnen Parameter wie "bilabial", "labio-dental" usw. als Klassennamen aller der Segmente auffassen, die hinsichtlich der Artikulationsstelle als gleich betrachtet werden. Hierbei ist zu beachten, daß der Ausdruck "gleich" nicht Identität bedeutet (vgl. S. 144). "Man gibt sich mit einer Approximation zufrieden, das heißt es hängt von der Feinheit der Differenzierung ab, 'wie gleich' sozusagen Objekte in einer Klasse sind" (SIXTL 1967, 5).

Die Zahl der Klassen hängt u.a. von dem Grad der Genauigkeit des Meßverfahrens ab und davon, welche Unterschiede man überhaupt erfassen will. Wenn man wie PIKE (1943, 124) nur solche Unterschiede registrieren will, die oberhalb der Wahrnehmungsschwelle liegen, dann wird die Klassenzahl voraussichtlich kleiner sein, als wenn man verschiedenartige Meßgeräte einsetzt, mit denen sich feinere Schattierungen festhalten lassen.

Die einzelnen Klassen sind nicht nur voneinander verschieden, sondern in bestimmter Weise geordnet, wie dies durch die Rangzahlen zum Ausdruck gebracht wird.

Es ist nun darauf zu achten, daß jede Ordinalskala auf einem komparativen oder topologischen Begriff basiert. Wenngleich dieser Begriff nicht immer explizit angegeben zu sein braucht, so muß es doch jedesmal möglich sein, ihn zu rekonstruieren. So können wir das Merkmal "place of articulation" als komparativen Begriff auffassen, der durch die Relationen "weiter vorne artikuliert als" und "an der gleichen Stelle artikuliert wie" konstituiert wird (vgl. auch LINDNER 1969, 175). Die Meßskala und die durch sie erzeugte Rangordnung werden durch diesen Begriff und nur durch ihn vollständig festgelegt. Bezogen auf unser Beispiel: Ein dentales Lautsegment erhält deshalb und nur deshalb einen niedrigeren Rang als beispielsweise ein velares, weil es "weiter vorne als dieses" artikuliert wird. Die Bezeichnungen wie "bilabial", "labio-dental" usw. sind also nur insofern von Bedeutung, als sie je verschiedene Ränge bezeichnen. Diese sind logisch allein durch den genannten Begriff festgelegt. Selbstverständlich muß empirisch gewährleistet und mit Hilfe einer empirischen Regel überprüfbar sein, daß die einzelnen Klassen, auf die die Namen hindeuten, und ihre Anordnung dem zugrundeliegenden komparativen Begriff entsprechen. Bei der Parameterklasse "place of articulation" dürfte diese Bedingung erfüllt sein.

Die Rangordnung der Parameter wird, wenn überhaupt, so nur in recht allgemeinen Ausdrücken begründet, z.B.: "... there is a contiguity between many of the adjacent parameters" (S. 144),

oder: "A vowel or consonant parameter class is a maximal set of related vowel or consonant parameters, respectively" (S. 144). Etwas näher wird lediglich die Parameterklasse "secondary articulation" begründet, wie aus folgendem Zitat ersichtlich: "The order of the individual categories under the parameter class of secondary articulation is according to position along the vocal tract" (S. 144). Wir sind daher darauf angewiesen, die komparativen Begriffe zu rekonstruieren.

Betrachten wir die fünfte Parameterklasse für Konsonanten:

5. Manner of articulation
   .0 Nasal
   .1 Liquid
   .2 Fricative (sibilant)
   .3 Trill (flap)
   .4 Plosive

Die Autoren versuchen, den hier implizit vorausgesetzten Merkmalsbegriff durch Andeutungen über die kontinuierliche Modifikation eines Explosivlautes zu einem Vibranten usw. wenigstens anzudeuten: "A plosive may be continuously modified until it becomes a flap, which in turn may be doubled or tripled to form a trill; a flap or trill may be continously modified to form a fricative; the degree of constriction for a voiced fricative may be continuously relaxed until a liquid is formed; etc." (S. 147). Bezeichnenderweise jedoch sagen sie nichts darüber, in welcher Weise man eine nasale Artikulation als Ergebnis der Modifikation der Artikulation eines Liquiden auffassen könnte, inwiefern ein Nasal einem Liquiden bezüglich der Artikulationsart "vorausgeht".

Auch die folgende Beobachtung ist geeignet, Zweifel an der Korrektheit der Ordinalskala aufkommen zu lassen: Während bei

PETERSON und HARARY die Parameter "nasal" und "plosive" in der Rangordnung die beiden extremen entgegengesetzten Ränge einnehmen, 0 und 4, folgen sie in dem noch zu besprechenden Merkmalsystem von GERŠIĆ (1971, 102) unmittelbar aufeinander: 0 plosive, 1 nasale.

Zweifel können beispielsweise auch im Hinblick auf das Merkmal "secondary articulation" erhoben werden. Die einzelnen Parameter dieser Klasse sollen angeordnet sein "according to position along the vocal tract" (S. 144). Aber es bleibt unklar, wieso beispielsweise "spread" weiter vorne im Artikulationstrakt liegen soll als "labialized".

Ebenso wie jede einzelne Parameterklasse, die auf einer Ordinalskala gemessen wird, einen komparativen Begriff voraussetzt, so muß auch die rangmäßige Ordnung der Merkmale selbst, wie sie bei PETERSON und HARARY zu finden ist, auf einem solchen Begriff beruhen. Die Autoren unterscheiden für Vokale und Konsonanten je fünf Parameterklassen, die, wie unsere Aufstellung zeigt, hierarchisch geordnet sind.

Allgemein wird diese Hierarchie mit dem Hinweis begründet, daß Merkmale, die mit anderen korreliert seien, einen niedrigeren Rang als diese zugeordnet bekämen. So heißt es z.B.: "Degree of lip rounding is normally closely correlated with tongue height; ..., and rounding is shown as a secondary articulation ..." (vgl. dazu GERŠIĆ 1971, 100f.).

Eine ausführlichere Begründung fehlt jedoch. So sagen die Autoren auf S. 146: "A description of the consonants according to their place of articulation and manner of articulation is basic". Damit soll begründet werden, warum die Parameterklassen "place of articulation" und "manner of articulation" höhere Ränge als die übrigen Klassen einnehmen. Diese Begründung kann

erstens in sich angezweifelt werden, da z.B. die Ausprägungen des Merkmals "laryngeal action", also "voiceless" und "voiced", mit den genannten höher eingestuften Parameterklassen nicht korreliert sind. Außerdem bleibt auch unklar, warum "manner of articulation" höher rangiert als "place of articulation".

Unser Versuch, die von PETERSON und HARARY vorausgesetzten komparativen Begriffe zu rekonstruieren, hat also, aufs Ganze gesehen, zu einem negativen Ergebnis geführt. Deshalb ist es uns auch nicht möglich, mit dem PETERSON-HARARYschen Maß der phonetischen Differenz zu operieren.

Wenden wir uns nun noch dem Vorschlag von GERŠIĆ (1971) zu, die phonetische Differenz zwischen zwei Lauten zu berechen (vgl. auch GERŠIĆ 1972; DROMMEL et al. 1973). Hierbei werden wir uns auf einige Punkte beschränken, die uns wichtig erscheinen.

Ähnlich wie PETERSON und HARARY stellt GERŠIĆ für Vokale und Konsonanten gesonderte Parameterklassen auf. Diese Entscheidung wird durch reine Zweckmäßigkeitserwägungen (die Vektoren würden sonst zu kompliziert u.ä.), also nicht phonetisch begründet, so wie das PETERSON und HARARY immerhin versucht haben.

Seine Parameterklassen betrachtet GERŠIĆ als gleichwertig, "um das Element subjektiver Entscheidungen zu eliminieren, das bei der Bildung von Hierarchien bei PETERSON und HARARY zu beobachten ist" (S. 101). Damit ist er der Notwendigkeit enthoben, eine Rangfolge der Parameterklassen zu begründen.

Jede Parameterklasse soll eine Ordinalskala darstellen. Dazu heißt es bei GERŠIĆ: "Diese Parameterklassen wurden in Stufen unterteilt und den Stufen wurden in der üblichen Weise ganze positive Zahlen zugeordnet" (S. 102).

Bei einigen Parameterklassen ist es nicht weiter schwer, den vorausgesetzten komparativen Begriff zu rekonstruieren, z.B. bei der Klasse "Dauer" von Vokalen

0 kurz
1 halblang
2 lang
3 überlang

Bei anderen ist es schwieriger. Die Parameterklasse "sekundäre Artikulation" für Konsonanten ist folgendermaßen unterteilt:

0 ohne sekundäre Artikulation
1 palatalisiert
2 leicht aspiriert
3 stark aspiriert

Es müßte begründet werden, in welchem Sinne die Palatalisierung der Aspirierung "vorausgeht".

Völlig unmöglich ist es, herauszufinden, welcher komparative Begriff der Stufung der Parameterklasse "Art der Artikulation" für Konsonanten zugrundeliegen soll. Dies wird besonders deutlich, wenn man die Stufung von GERŠIĆ und die von PETERSON und HARARY nebeneinanderhält:

GERŠIĆ		PETERSON und HARARY	
0	plosive	0	Nasal
1	nasale	1	Liquid
2	laterale	2	Fricative (sibilant)
3	vibrante	3	Trill (flap)
4	frikative	4	Plosive
5	Konsonantenverlust		

Wie die Verbindungslinien in diesem Schema zeigen, stimmen die beiden Stufungen nur in einem Rang überein. Die größte Differenz liegt bei den Plosiven vor: Bei GERŠIĆ nehmen sie den niedrigsten, bei PETERSON und HARARY den höchsten Rang ein. Dieser Unterschied und all die anderen sind nicht ohne weiteres zu verstehen. Dies gilt um so mehr, als GERŠIĆ die von PETERSON und HARARY gewählte Rangordnung an keiner Stelle seines Buches kritisiert, obwohl er sich doch mit der Arbeit dieser Autoren mehrfach auseinandersetzt, ihren Ansatz weiterentwickeln will.

So können wir also auch das von GERŠIĆ (1971, 114) vorgeschlagene Maß der phonetischen Differenz zwischen zwei Lauten kaum akzeptieren.

Überlegen wir uns zum Schluß, welche Einsichten und Folgerungen aus unserer Untersuchung gezogen werden können. Wir haben gesehen: Die Bestimmung der artikulatorischen Differenz ist eine abgeleitete Messung, die erst dann in zuverlässiger Weise durchgeführt werden kann, wenn die zugrundezulegenden direkten Messungen korrekt vorgenommen worden sind. Damit diese Bedingung erfüllt werden kann, ist es zunächst nötig, all die artikulatorischen Parameter, die den Bezugsrahmen abgeben sollen, zu metrisieren. Unsere Untersuchung hat sich kritisch mit zwei Vorschlägen beschäftigt, die in die Richtung einer solchen Metrisierung weisen. Die Kritik, die an diesen Vorschlägen geübt wurde, entwickelte sich aus dem Versuch, ihre begrifflichen Grundlagen rekonstruierend zu klären. Wir müssen uns aber darüber klar sein, daß wir es nicht mit einem rein wissenschaftstheoretischen Problem zu tun haben. Es müssen darüber hinaus Meßapparaturen entwickelt werden, um quantitative Angaben über die ausgewählten Parameter zu gewinnen. Gerade die Lösung dieser Frage stellt die Phonetik vor nicht unbeträchtliche Schwierigkeiten (vgl. z.B. HARDCASTLE 1974). Diese Schwierigkeiten sind erst zu einem Teil gelöst (vgl. z.B. SLIS 1970), ein Bezugssystem quantifizierter artikulatorischer Parameter steht noch nicht zur Verfügung.

Literatur

CARVELL, H.T., SVARTVIK, J., Computational Experiments in Grammatical Classification. The Hague-Paris 1969.

DROMMEL, R., GERŠIĆ, S., HINTZENBERG, D., Eine Methode zur numerischen Erfassung der Suprasegmentalia. Phonetica 27, 1973, 1-20.

GERŠIĆ, S., Mathematisch-statistische Untersuchungen zur phonetischen Variabilität, am Beispiel von Mundartaufnahmen aus der Batschka. Göppingen 1971.

GERŠIĆ, S., Phonetische Variabilität des Dialekts, dargelegt am Beispiel des Donauschwäbischen. S. JÄGER (ed.): Linguistik und Statistik, Braunschweig 1972, 33-50.

HARDCASTLE, W.J., Instrumental investigations of lingual activity during speech: a survey. Phonetica 29, 1974, 129-157.

KELZ, H., Artikulationsbasis und phonetische Beschreibungsparameter. Kommunikationsforschung und Phonetik, Hamburg 1974, 217-238.

LEHFELDT, W., ALTMANN, G., Begriffskritische Untersuchungen zur Sprachtypologie. Linguistics 144, 1975, 49-78.

LINDNER, G., Einführung in die experimentelle Phonetik. München 1969.

PETERSON, G.E., HARARY, F., Foundations of phonemic theory. In: JAKOBSON, R. (ed.), Structure of Language and its Mathematical Aspects. Providence, Rhode Island, 1961, 139-165.

PIKE, K.L., Phonetics. Ann Arbor 1943.

SAPIR, E., Die Sprache. München 1961 (zuerst engl. 1921).

SIXTL, F., Meßmethoden der Psychologie. Theoretische Grundlagen und Probleme. Weinheim 1967.

SLIS, I.H., Articulatory measurement on voiced, voiceless and nasal consonants. Phonetica 21, 1970, 193-210.

STEGMÜLLER, W., Theorie und Erfahrung. Heidelberg-New York 1970.

On Measurement of Lexical Differences by Means of Frequency

Dirk Geens, Leuven

Within the framework of a research project on the syntactic properties of English dialogues, a corpus of theatrical English was compiled at the Institute for Applied Linguistics of the Catholic University at Leuven. This sample contains the texts of 62 present-day (= 1966-1972) British plays*, which were selected in such a way that the sample is homogeneous for syntactic research and representative of the literary genre that provides the material (GEENS 1973, GEENS 1975).

Although the aim of this project is a syntactic study, it is obvious that some interesting by-products, such as studies on the lexical level and stylistic investigations, can be obtained from this corpus as well. The intention of our present study is to propose some methods for quantitative analysis of language material, particularly as regards the <u>content</u> of texts rather than the <u>form</u>. To this end we selected the problem of comparing the vocabulary of the theatre-corpus with a standard corpus of written English: our immediate goal was to describe a number of properties that <u>characterize the genre</u> with regard to the vocabulary that is used. Central throughout our study were the frequencies of the words; our attention was mainly focussed on

1) the individual frequency-differences of lexical entries,
2) and the frequency-differences that occur with word-categories.

---

*The texts were selected mainly on the basis of the "naturalness" of the dialogues they contained.

IT MAKES A DIFFERENCE

We did not work with the frequency-distributions of the vocabulary, as these seem to be less suitable for a description of content, due to the degree of abstraction from the texts. During the set-up of this study we took the line that the frequency-differences between individual lexical entries had to be examined. Consequently what mattered first of all was to define a way to measure these differences in such a way that attributive descriptions in the fashion of 'much', 'little', 'considerable', 'minimal' and the like could be avoided.

The argumentation developed in this section will be illustrated with examples taken from the theatre-corpus. To this purpose the corpus was divided into two parts, T_j and T_k, with respectively n_j and n_k words such that $n_k = n_j / 2$ and, $n_j + n_k = n_{corpus}$. In the examples given below f denotes the absolute frequency, $\hat{p}$ the relative frequency, D the difference, and the indexes j and k identify the two samples.

Absolute Difference

Expressing the difference in terms of absolute numbers - as can be done with the formula

$$D = f_j - f_k$$

- has the evident disadvantage that incongruities in sample size and coverage cannot be taken into account. In the following example we compare the words <u>coffee</u> (120 occurrences in the theatre corpus) and <u>house</u> (650 occurrences) by means of their absolute frequencies.

	f_j	f_k	D
coffee (120)	110 (.015%)	10 (.002%)	100
house (650)	375 (.053%)	275 (.078%)	100

In both the cases the absolute difference is 100, though in the case of coffee the relative difference is .013%, whereas it is -0.025% for the word house.

Relative Difference

To avoid such falsified results one might decide to work with relative frequencies; a possible formula is

$$D = \hat{p}_j - \hat{p}_k$$

When we examine the example in the preceding section we see that difference in sample size has been ruled out now, although differences in coverage have not been accounted for yet. Consider the following example, where the words the (39,405 occurrences in the theatre-corpus) and those (656 occurrences) are compared.

	f_j	f_k	D
the (39,405)	26,372 (3.767%)	13,033 (3.723%)	.044%
those (656)	540 (0.077%)	116 (0.033%)	.044%

The relative difference is .044% in both cases. Still we can not proclaim that a difference of .044% is the same for an item that covers some 3.8% of the corpus as for one that only covers .06%. Therefore we reject relative difference as a measurement method as well.

A Balanced Difference Coefficient

In fact what has to be measured is not only the value of the difference, but also its intensity. To achieve this purpose we have used the formula

$$DC = \frac{\hat{p}_j - \hat{p}_k}{\hat{p}_j + \hat{p}_k}$$

which expresses the proportion relative difference/coverage.

This ratio accounts for sample size as well as for coverage. It produces a real number with values in between -1 and +1, whose absolute values express the intensity of the relative differences; the sign expresses the direction of those differences. For $\hat{p}_j = 0$ the value calculated is -1; for $\hat{p}_k = 0$ the value is +1; and for $\hat{p}_j = \hat{p}_k$ it is 0.

For the same relative differences between lexical items different coefficients are yielded according to the total coverage of the lexical item in the corpora (DE KOCK, GEENS 1976).

Significance of a Difference

Since the difference coefficients do not correspond to a distribution they cannot have a predictive value; in other words we are able to measure the intensity and the direction of a difference, but we have no information concerning the significance of a difference. This implies that the coefficient can be used to make a categorization of lexical items, but also that we shall have to consider that fact that important differences can be insignificant because they are based on too few observations. Therefore we believe that a study which aims at describing the properties of texts by comparison of the lexical items must be complemented by a calculus of the significances of differences observed. For however considerable a difference may be, it can be based on too small a number of occurrences. As a rule the reliability of extrapolations decreases with the amount of observations on which they are based.

To measure the significance of a difference we use the standardized variable

$$z = \frac{\hat{p}_j - \hat{p}_k}{\sqrt{pq\left(\frac{1}{n_j} + \frac{1}{n_k}\right)}}$$

which is approximately normally distributed with paramters (0,1). Here

$$\hat{p}_j = \frac{f_j}{n_j}$$

$$\hat{p}_k = \frac{f_k}{n_k}$$

n_j, n_k = sample sizes

$$p = \frac{f_j + f_k}{n_j + n_k}$$

$$q = 1 - p$$

The probability of z can easily be found in the tables of the normal distribution.

These techniques have been applied during our study of the characteristics of the theatre corpus in comparison with a standard corpus of English, for which we have used the well-known

corpus of present-day American English by H. KUCERA and W.N. FRANCIS, referred to as Brown Corpus (KUCERA, FRANCIS 1967).

THE THEATRE CORPUS COMPARED WITH THE BROWN CORPUS

The Brown corpus contains 1,014,232 tokens, dispersed over 50,406 types, whereas the theatre corpus contains 1,035,472 tokens and 33,521 types.

As far as tokens are concerned they are both approximately the same size, the difference being some 2%. However, if types are compared the Brown corpus is substantially bigger, the difference there being about 33%. The type-token ratios are for the Brown and the theatre corpus: 0.050 and 0.032 respectively, indeed a difference that points to a more diversified vocabulary in the Brown corpus. We have the impression that among the most important reasons for that difference three are self-evident:

1) there is a structural reason in that the Brown corpus contains 15 genres and consists of 500 samples; in fact, as more different authors have contributed to the vocabulary we may assume that the sum of their vocabularies taken individually is greater than that of the 62 contributing authors in the theatre corpus;
2) thematically there is a greater variation in the texts form which the Brown corpus was sampled; it covers such genres as fiction, technical writing, scientific writing, newspaper articles, and so on. As a matter of fact the theatre corpus consists of only one genre (one which was not included in the Brown corpus);

3) for stylistic reasons the authors of the plays may very well be expected to use deliberately a language register which is traditionally considered less varied in terms of vocabulary; it is a widespread assumption (though never proved) that in discourse the lexicon is less extended.

Together the Brown corpus and the theatre corpus contain 64,123 types, in 2,049,704 tokens. Of these 30,602 types occur only in the Brown corpus, 13,171 only in the theatre corpus, and 19,804 overlap. It is natural that our attention will be devoted primarily to the types that overlap, for the outsiders are considered less interesting for the following reasons:

1) they contain different spelling-forms between American and British English;
2) the bulk of proper nouns occurs in those groups;
3) the content words in these categories are so strongly topic-bound that their frequencies do not convey any relevant information;
4) there is no real basis for a comparison, except for the fact that they are outsiders in one of the corpora.

For the words that overlap we computed intensity-values, direction and significance of the differences, as described above. In order to distinguish the characteristics of both the corpora a multi-dimensional categorization was considered desirable: on the one hand categorization should be based on our statistical definitions of difference, whereas on the other hand syntactic and semantic categorizations must be performed.

The syntactic categorization was done automatically by means of a slightly adapted version of our automatic syntactic analysis program (GEENS 1974). Insofar as words were grouped into semantic

classes this had to be done manually. Exactly because of the amount of manual work to be done we have up till now limited ourselves to the function words (which are most suitable for general observations concerning the structure of the corpus) and the content words at the extremities of the differential categorization.

Function Words

The selection of function words we examined covers 456,051 tokens of the Brown corpus (44%) and 541,428 tokens of the theatre corpus (52%). It comprises:

- the subject and object forms of personal pronouns,
- the possessive pronouns,
- the reflexive pronouns,
- the demonstratives,
- the articles,
- the indefinite pronouns,
- the conjunctions,
- the prepositions,
- the exclamations,
- the auxiliary verbs,
- the interrogative pronouns.

For obvious reasons we were unable to distinguish between the syntactically ambiguous forms; neither of the corpora contains syntactic information of the kind required for disambiguation of those forms. As a result, for the following forms a theoretical frequency was calculated on the basis of their relative frequencies in the classes they belong to:

- you and it
- her
- his and its
- for and since.

Some other forms, like that, can, will, may and might as well as the interrogatives have retained their ambiguity, as it was impossible in those cases to calculate a theoretical frequency (mainly because they belong to three classes or belong to open classes). In the tables the syntactically ambiguous forms are italicized; those for which no theoretical frequency was calculated are given in brackets.

A particular problem was posed by the auxiliaries: some of the abbreviated forms of auxiliaries coincide ('d = had and would, 'll = shall and will, 's = is and has). Moreover, in the Brown corpus the abbreviated forms are recorded as one word-form, together with their subject, whereas in the theatre corpus they are as a rule recorded apart from their subject. For these cases again a theoretical frequency was calculated.

As the rule applied to measure the difference-coefficients was

$$DC = \frac{\hat{p}_{\text{Brown Corpus}} - \hat{p}_{\text{theatre corpus}}}{\hat{p}_{\text{Brown Corpus}} + \hat{p}_{\text{theatre corpus}}}$$

it is clear that negative values will point to a greater ascendancy in the theatre corpus, positive values will then point to preponderance in the Brown corpus.

Tables 1 to 5 show the function words arranged according to the difference-coefficients obtained by this computation. With respect to the significance-coefficients, we decided to set the z value to a minimum of 2.58, which corresponds to a 99%

threshold value. For all the word-forms examined this value was exceeded except for those forms marked by an asterisk in the tables.

The difference-coefficients were also calculated for the syntactic categories as a set of word-forms; in the tables the area in which the coefficients of these groups can be located will be shaded.

The function words as a group can be regarded as a stable body; indeed, by calculating the difference-coefficient on the basis of all the words we examined no significant difference could be observed (DC = 0.0).

The personal pronouns were marked by a high difference-coefficient, -0.5; indeed, this considerable difference confirms our expectations, namely that the usage of personal pronouns (in particular the 1st and 2nd persons, whose coefficients were respectively -0.7 and -0.8) is much more intense in dialogues - even written discourses - than in prose. The possessive pronouns received a coefficient of 0.0 (that is, no noticeable difference) for attributive usage, though it was -0.4 for pronominal usage. Again this is an important difference. With regard to the reflexive pronouns, the difference observed was -0.1. Some general observations can be attached to the results of the comparison of these categories, which all imply references within the same stratum:

1) in every class the order in which the persons appear is the same, 2nd person - 1st person - 3rd person;
2) the areas in which the different persons can be located within the table are very different in size:
   - the 2nd person forms occur in a very narrow area, ranging

from -0.6 to -0.8,

- the 1st person covers a less narrow area, which ranges from +0.7 to -0.4;

3) within the group of 3rd person forms there is some similarity in the sequence in which masculine, feminine, neutral and plural forms appear: this order is

- feminine, masculine, plural, neutral for possessive and reflexive pronouns,

- neutral, feminine, plural and masculine for personal pronouns subject forms,

- feminine, neutral, masculine and plural for personal pronouns object forms;

4) for the 1st person the difference observed was always greater for the singular than for the plural forms;

5) the group difference-coefficient can always be located within the area occupied by the first person.

The difference-coefficients for indefinite pronouns range between +0.5 and -0.5, the group difference is -0.1, which seems to point to a slight preponderance in the theatre corpus. No difference was made between pronominal and attributive usage for the pronouns each, one, any, some and all.

The interrogative pronouns range from +0.5 to -0.6, whereas their group coefficient is -0.2. Again those forms seem to be more frequent in the theatre corpus. It should be noted that the word-forms in this group can actually function as interrogative pronouns, relative pronouns and subordinators. Since the only other relative pronoun is that (which is syntactically ambiguous on its turn) we were unable to compute a theoretical frequency here.

The coordinators, subordinators and prepositions are characterized by a greater usage in the Brown corpus, (their difference-coefficients are respectively +0.1, 0.0 and +0.1). The 0.0 coefficient for subordinators may be explained by the very high frequency of if (2199 and 3937). In the group of prepositions toward and towards can be found at the extremes of the table; this seems to point to regional differences in usage.

The articles and the demonstratives are preponderant in the Brown corpus, with difference-coefficients of respectively +0.2 and +0.1. For the demonstratives two group coefficients have been calculated: +0.1 for the group without that (which is highly ambiguous), and 0.0 for the group including that.

The group coefficient for the auxiliary verbs is -0.2 and thus the difference here is in favour of the theatre corpus. The coefficients for the forms in this group range in between +0.4 and -0.8. For the individual present and past forms of the verbs be, do and have the difference coefficients in the table are listed with the word-form. In order to compare the usage of the tenses as well, the difference of the forms in sub-groups containing the past and present tenses of those verbs was computed. These coefficients are represented under the heading 'present/past tenses of ...'. The total number of forms of those verbs is represented under the heading 'forms of ...'.

As for the exclamations it could be expected that they are far more frequent in the theatre corpus. Their difference-coefficient is -0.8; the values assigned to the forms in this class range from +0.5 to -0.9.

These calcualtions lead to the following observations:

1) in discourse the first and second person of the personal

pronouns are preponderant to the third person;

2) the usage of anaphoric forms exceeds the usage of deictic forms in the theatre corpus;

3) the usage of conjoining (coordinators, subordinators and prepositions) forms in the Brown corpus goes far beyond the limited usage made of those forms in the theatre corpus;

4) as the latter forms help in constructing more complex sentences, it can be extrapolated that syntactically the dialogues are characterized by a less complex sentence structure than the prose Brown corpus;

5) the function words constitute a grammatical superclass that is stable throughout the two corpora, considering the fact that as a group they do not yield a significant difference factor;

6) the theatre corpus shows more variety in the usage of auxiliary verbs, which as a rule indicates a more diversified picture of modality and aspect;

7) the exclamations evidently dominate in the theatre corpus; besides the significantly higher number of occurrences of the overlappings there are an important number of outsiders in this corpus.

Content Words

Clearly we had to examine the content words in a different way: they belong to open classes and it was impossible so far to make a complete semantic categorization manually. This leads us to the assumption that an analysis of their representatives in the difference categories at the extremes should provide us with some

clues for further interpretation. Therefore we selected the content words that have difference-coefficients between +0.9 and +0.85, or -0.9 and -0.85. A categorization in meaning classes was provided manually. In any case the significance test pointed to highly significant differences.

In the theatre corpus the majority of the nouns points to ***animate human beings***, often referring to a human being by means of synonyms, titles, indications of relationships. Some examples:

(1) titles and vocatives: duchess, mum, mister, madam

(2) synonyms: bloke, lads, chaps

(3) relationships: daddy, auntie, dad.

Sometims reference is made to a special characteristic or an occupation. Examples of these classes are:

(4) characteristic: idiot, bastard, neighbours

(5) occupation: gunner, whore, constable, archbishop, gardner

Other nouns display an ***indirect relationship*** to human beings by referring to names of food, clothes, parts of the body, utensils, buildings, parts of houses, furniture and the like. Examples are:

(6) food: ale, tea, biscuit, marmelade

(7) clothes: jeans, stocking, jumper, trousers, waistcoat, tunic

(8) parts of the body: moustache, snout

(9) utensils: briefcase, transistor, wig, handbag, torch

(10) buildings: dispensary, greenhouse, bungalow

(11) parts of houses: livingroom, doorbell, exit, curtain

(12) furniture: tabernacle, cupboard, armchair, sofa

In the Brown corpus fewer nouns point to ***human beings***, and if so, rather to professions and roles in political activity than to proper identification names. Examples:

(1) professions: attorney, reporters, dealers, architect

(2) political activity: candidates, mayor, voters, sheriff

The majority of the nouns that are far more frequent in the Brown corpus however, refer to ***inanimate objects***. They generally deal with business, administration, science and technology:

(3) business: company's, funds, trends, ownership, industry

(4) administration: commission, allotment, conference

(5) science: hydrogen, systems, frequency, techniques

An exhaustive list of the nouns interpreted is in table 6.

It should be noted that the words with a difference in favour of the Brown corpus tend to be derived from Latin roots, whereas those favoured in the theatre corpus have more Anglo-Saxon roots. This feature makes the vocabulary of the Brown corpus sound learned and educated, whereas the theatre corpus vocabulary sounds "ordinary" and rather resembles the vocabulary of every-day language.

In the theatre corpus most verbs in these classes refer to actions that require an ***animate subject***; as a rule they indicate actions that involve some physical activity, sometimes they are actions that involve mental activity:

(1) physical activity: stares, laughs, kneels, lifts, shouts

(2) mental activity: examines, dreamt, imagining, boring

The verbs frequent in the Brown corpus are of a more abstract nature and breathe an atmosphere of ***business***, ***science and technology***:

increase, manufacturing, intensified, registered, anticipated.

As with the nouns, the verbs in the theatre corpus relate to people, and have Anglo-Saxon roots, whereas those of the Brown corpus relate to objects and derive from Latin roots.

An exhaustive list of the verbs examined is given in table 7.

The adjectives of the theatre picture characteristics of human beings or things related to human beings:

tidy, naughty, horrid, unfaithful, dearest, jealous, silly;

the Brown corpus adjectives again lie within the sphere of business and science:

fundamental, urban, experimental, optical, bronchial, nuclear.

Similarly, the same interpretation holds true for the adjectives. A complete list ist given in table 8.

As to the adverbs, no clues of any importance were adduced by the differential categorization to enable them to be mapped out distinctly in semantic strata as could be done for the nouns, verbs and adjectives. In fact only a few adverbs yielded a difference coefficient of |0.6| or more, so that it may be expected that they constitute a relatively stable category within the two corpora.

To recapitulate, we observed the following differences between the Brown corpus and the theatre corpus:

(1) the Brown corpus contains more different words, and consequently makes a more diversified impression than the theatre corpus; the 'richness' of the Brown corpus has its origin in the structure of the corpus, the themes that were employed and the style of the language it was sampled from;

(2) although the function words seem to constitute a stable body in the comparison of the two corpora, the anaphoric function words prevail in the theatre corpus, whereas the connectors prevail in the Brown corpus; for this reason the Brown corpus can be assumed to contain more complex syntactic structures than the theatre corpus;

(3) semantically the content words related to business, administration, science and technology dominate in the Brown corpus, whereas the theatre corpus counts more words related to human beings;

(4) at the same time this accounts for the fact that among the words examined those from the Brown corpus derive mainly from Latin roots and sound learned or educated, whereas those from the theatre corpus tend to derive from Anglo-Saxon roots and sound "ordinary", as if reflecting everyday life.

To summarize, our study of a technique to compare two totally independent vocabularies led to the following conclusions:

(1) the method we used confirmed our intuitive hypotheses and proved to be reliable;

(2) our study contains enough evidence for the fact that frequency can be used to analyze vocabularies, and even the content of samples;

(3) there are significant differences between the Brown and the theatre corpus, but to proclaim that these are due to the opposition British/American English or to the opposition discourse/writing would be very rash; probably both these oppositions, together with other - structural - reasons, play their part;

(4) if two large text collections display such divergencies, it should be clear that frequency-lists have to be handled with care when used as resourcer; no absolute pronouncements about a language can safely be made when based on frequency only.

References

DE KOCK, J., GEENS, D., Por una estilìstica lìnguìstica y cuantitativa con la ayuda del ordenador. Madrìd, Revìsta de la Universidad Complutense, Vol. XXV, Nr. 102, 1976, 113-129.

DE KOCK, J., GEENS, D., Pour une stylistique lìnguistique et quantitative avec l'aide de l'ordinateur. Paris, Linguistique Automatique et Langues Romanes (ed. J. DE KOCK, C. BOITET), 1977, 198-215.

GEENS, D., A Frequency-List of Sentence Structures: Preliminary Considerations. Leuven, Review of the ITL, Nr. 21, 1973, 39-55.

GEENS, D., A Venture in Computational Analysis of English Syntax. Leuven, Preprints of the Department of Lingustics, Nr. 19, 1974.

GEENS, D., Analysis of Present-Day English Theatrical Language. Leuven, Progress in Applied Lìnguistics, Nr. 6 (Progress Report) 1975.

KUCERA, H., FRANCIS, W.N., Computational Analysis of Present-Day American English. Providence, R.I., Brown U.P., 1967.

Table 1. Difference Coefficients for personal pronouns, possessive pronouns and reflexive pronouns.

DC	Personal Pronouns: Subject	Personal Pronouns: Object	Possessive Pronouns: Attributive	Possessive Pronouns: Pronominal	Reflexive Pronouns
+0.9					
+0.8					
+0.7			ITS		
+0.6					
+0.5				ITS	ITSELF
+0.4			THEIR		
+0.3					THEMSELVES
+0.2			HIS	THEIRS*	
+0.1					HIMSELF
0.0	HE		HER //////		
			OUR //////		
-0.1	THEY*	THEM		HIS	HERSELF* //////////
					OURSELVES* //////////
-0.2	SHE	HIM			
-0.3	IT	IT		HERS*	
				OURS*	
-0.4	WE	HER		///////	
-0.5	//////////	//////////			
-0.6			MY	MINE	MYSELF
				THINE	YOURSELVES
-0.7	I	ME	YOUR	YOURS	YOURSELF
-0.8	YOU	YOU			
-0.9					

* An asterisk in the tables indicates the words that yielded a z value below the threshold level (2.58).

* Shaded areas in the tables mark the difference coefficient level for the class of words as a group.

Table 2. Difference Coefficients for indefinite pronouns and interrogative pronouns.

DC	Indefinite Pronouns	Interrogative Pronouns
+0.9		
+0.8		
+0.7		
+0.6		
+0.5	SOMEWHAT	WHICH
+0.4	SUCH	WHOSE
+0.3	EACH	
+0.2	SELF	WHOM
	OTHER	
	BOTH	
+0.1	OTHERS	
0.0	ANOTHER	WHEN
	ONE	
	ONES	
	ONE'S	
	EVERY	
	ANY	
	SOME	
-0.1	NONE* //////////////////	
-0.2		WHERE //////////////////
		WHENCE //////////////////
-0.3	ALL	
	EVERYBODY	
	SOMEBODY	
	ANYONE	
	EVERYONE	
-0.4	NOBODY	HOW
	ANYBODY	
	NOTHING	
	EVERYTHING	
	SOMETHING	
-0.5	ANYTHING	
	SOMEONE	
-0.6		WHAT
		WHY
-0.7		
-0.8		
-0.9		

Table 3. Difference Coefficients for Coordinators, Subordinators, and Prepositions.

DC	Coordinators	Subordinators	Preposistions
+0.9			TOWARD
+0.8			
+0.7		ALTHOUGH	CONCERNING PER
+0.6			DURING WITHIN VIA
+0.5			BETWEEN AMONG CONSIDERING
+0.4		WHEREAS	BY BEYOND UPON SINCE
+0.3	OR UNLIKE THAN	LEST*	OF ABOVE UNTIL UNDER
+0.2	NEITHER	WHILE WHETHER SINCE	THROUGH FROM BENEATH IN ONTO AGAINST AROUND
+0.1	EITHER AND NOR		ACROSS FOR
0.0	FOR BUT	UNLESS THOUGH BECAUSE	EXCEPT TO INTO OVER AT OUTSIDE AFTER WITH WITHOUT ON INSIDE
-0.1			BEHIND* BESIDE*
-0.2		IF	ABOUT OUT
-0.3			UP DOWN
-0.4			
-0.5			
-0.6			TOWARDS
-0.7			
-0.8			
-0.9			

Table 4. Difference Coefficients for articles, demonstratives and auxiliary verbs.

DC	Articles	Demonstratives	Auxiliary Verbs
+0.9			
+0.8			
+0.7			
+0.6			
+0.5			
+0.4			WERE DARED (MAY)
+0.3		THESE	WAS past of BE USED
+0.2	AN THE		
+0.1		THOSE	WOULD
0.0	A	THIS (THAT)	MUST COULD (MIGHT) forms of BE forms of HAVE
-0.1			ARE HAS SHOULD
-0.2			IS (WILL) (CAN) DID present of BE present of HAVE
-0.3			HAVE HAD LET
-0.4			OUGHT DOES
-0.5			DARE forms of DO
-0.6			SHALL DO present of DO
-0.7			
-0.8			AM
-0.9			

Table 5. Difference Coefficients for exclamations.

DC	Exclamations
+0.9	
+0.8	
+0.7	
+0.6	
+0.5	GODDAMN*
+0.4	
+0.3	GODDAM*
+0.2	
+0.1	
0.0	WHOA*
	ALAS*
	MM*
-0.1	
-0.2	GOSH*
-0.3	AHEM*
	NO-O*
-0.4	
-0.5	AW*
	OOOO*
-0.6	HI*
	OOPS*
-0.7	BAH*
	MMMM*
	HUH
-0.8	SHH*
	YEAH
	MMM
	UGH
-0.9	HMM
	AH
	OH
	YA
	AYE
	HA

Table 6. Difference Coefficients for nouns.

+0.9 - +0.85	attorney, candidates, mayor, reporters, allies, dealers, sitter, president's, dealer, giants, manufacturers, architect, reader, listeners, merchant, sheriff, voters, salesmen
	associations, formula, barn, fabrics, magnitude, proposals, solutions, stem, judgement, limitations, publication, company's, prevention, aids, bridges, concerts, frames, tangent, conviction, rates, clay, cents, composition, coverage, exhibit, exploration, stadium, materials, diffusion, narrative, utopia, construction, funds, sections, decade, equivalent, losses, craft, crystal, refrigerator, interference, growth, sources, capabilities, chart, complement, conversion, merger, mode, replacement, trends, billion, anniversary, commission, allotment, nations, emphasis, temperature, conference, region, onset, completion, dictionary, corporate, span, frequency, supervision, timber, gulf, meanings, obligations, ownership, planets, sphere, successes, teams, areas, hydrogen, budget, transportation, stress, industry, aspects, goal, distribution, certainty, alliance, jazz, maximum, procedure, tax, earnings, entries, participation, removal, shore, addition, features, negotiations, curves, data, congres, products, firms, techniques, dollars, radiation, league, organizations, systems, railroad, wagon, measurements, factors, characteristics, states, institute, motors, assembly, acceptance, structures, missile, trend, avenue, diameter, medium, developments, investment, route, marketing, oxygen, technology, downtown, limits, highway, axis, regions, treasury, districts, skills, faculty, core, objective, columns, industries, ratio, accuracy, unity, cooperation, roles, measurement, countries, campus, pathology, performances, stems, minimum, markets, administration, effectiveness, era,

Table 6. (continuation)

	emission, plastics, awareness, organization, location, considerations, dates, density, dimensions, flux, instances, occurrence
-0.9 - -0.85	mum, duchess, bloke, daddy, mister, madam, holiness, gunners, majesty, lads, chaps, lunatic, lordship, dad, nuns, sod, neighbours, sir, darling, divider, feller, ma'am, buggers, maitre, idiot, whore, queen's, blokes, lad, highness, chap, constable, constables, majesty's, blacks, auntie, mug, nun, undertaker, vice-chancellor, bastard, earl, archbishop, eminence, gardner, abbot, colonel's pigeons, roach, goldfish, rats, parrot wig, jumper, stocking, trousers, tunic, drapes, jeans, handbag, waistcoat, sword, moustache, snout, biscuit, marmelade, hash, ale, tea, cask, pub, shillings, livingroom, pause, dispensary, honour, sofa, shit, tabernacle, heaven's, favour, briefcase, humour, cupboard, disgust, defence, transistor, greenhouse, fortnight, parcel, behaviour, torch, judgement, clap, doorbell, packet, armchair, dune, madness, anthem, exit, rubbish, kettle, backside, char, bungalow, dustbin, fancies, pram, wills, ale, blackout, vase, bugle, curtain, vocation, banner

Table 7. Difference Coefficients for verbs.

+0.9 - +0.85	operating, located, increased, varying, estimate, corresponding, resulted, coating, determining, promote, sponsored, grinned, measuring, furnish, representing, obtained, expanding, paused, provides, operated, determine, desired, viewed, increase, insure, increases, manufacturing, slid, subjected, intensified, foams, registered, anticipated, listed, sampling, sighed, indicated, stated, varied, hesitated, regarding, enforced, emphasize, gains, expanded, included, glued
-0.9 - -0.85	stares, sits, nods, laughs, picks, pours, kneels, crosses, piss, leans, watches, exits, sob, snuts, lifts, sighs, jumps, kisses, shrugs, giggling, taps, smashing, throws, giggle, nicked, flicks, screams, shouts, cheers, listens, shakes, clears, dials, fuck, climbs, catches, collapses, knitting, sneer, straightens, undressed, wanders, grabs, pushes, rushes, gazes, smiles, opens, tosses, pulls, knows, closes, claps, turns, chuckles, dial, frown, groan, mumbling, seizing, stumbling, tidying, goes, moves, waits, walks, fetch, rises, bagged, creeps, bends, glances, sleeps, looks, wears, hesitates, examines, dream, fainted, disgusting, imagining, reacts, please, apologize, boring, fade, stops, puts, pauses, rings, cheating, oblige, stinking, shun, starve, finishes, disappears, shone

Table 8. Difference Coefficients for adjectives.

+0.9 - +0.85	characteristic, dominant, additional, evident, largest, fundamental, external, urban, visual, long-range, experimental, variable, pacific, widespread, republican, bronchial, oral, democratic, economic, minimal, current, flexible, available, technical, nuclear, civic, notable, governmental, residential, effective, significant, intermediate, industrial, newer, primary, organic, optical
-0.9 - -0.85	bloody, subaltern, tidy, rotten, naughty, horrid, unfaithful, dearest, backwards, cosy, dear, jolly, jealous, creepy, crafty, disgraceful, sarcastic, silly

Some Formal Aspects of Dynamics of Dialogues

Maria Nowakowska, Warsaw

1. Introduction

This paper presents an application of the general theory of actions (see NOWAKOWSKA 1973, 1975) to a construction of theory of dialogues. The approach suggested here is different than that given in an earlier paper on the same subject (see NOWAKOWSKA 1976). Roughly, the central idea may be described as follows: instead of building the semantics of dialogues in a traditional way (i.e. by designing more and more elaborate formal schemes for expressing the meanings of sentences in their contexts, situations, etc.; for the most refined system of this type see PETÖFI 1976), one could try to approach the subject by explicating the formal structure of various effects of semantic constraints operating in a dialogue. This approach is, in some sense, similar to that used in propositional calculus, where one pretends that the problem of the truth of simple propositions has already been solved, and proceeds to explicate the conditions for truth of more complex propositions.

2. The basic formal scheme

Let L stand for the natural language under consideration; thus, L is a class of strings of words of a given vocabulary. The elements of L, denoted generally by u,v,... with or without subscripts, will be referred to as speech units, such as sentences, questions, exclamations, etc. Further, let

$$L^{(n)} = \{ \alpha : \alpha = u_1u_2...u_n, u_i \in L, i = 1,...,n \}$$

be the class of all strings, each consisting of n speech units from L. Finally, let

$$L^{\infty} = L \cup L^{(2)} \cup L^{(3)} \cup \ldots$$

be the set of all finite strings of speech units from L.

If $\alpha \in L^{\infty}$, then $|\alpha|$ will denote the number of speech units in α.

Let now $K = \{1,\ldots,k\}$ denote the set of persons participating in the dialogue.

Let D^* be the set of all pairs (α,f), where $\alpha \in L^{\infty}$ and f: $\{1,\ldots, |\alpha|\} \to K$ is a function which assigns elements of K to integers between 1 and $|\alpha|$.

The intended interpretation here is such that α is the string of consecutive speech units uttered, whil f is a function which contains information as to who uttered which speech unit. If $\alpha = u_1u_2\ldots u_n$ (so that $|\alpha| = K$), and $f(i) = j$, then u_i is uttered by j-th member of the set K.

Naturally, not every pair (α,f) can be treated as an admissible dialogue. Thus (similarly as in formal linguistics), a set $D \subset D^*$ will be distinguished as a primitive of the system, and its elements (α,f) interpreted as dialogues.

It seems natural to make the following two assumptions.

Assumption 1. $k > 1$, i.e. the set K of participants in a dialogue consists of more than one element.

Assumption 2. If $(\alpha,f) \in D$, then

$$(\forall j \in K)(\exists i \in \{1,\ldots,|\alpha|\}): f(i) = j.$$

The meaning of the second assumption is such that if (α,f) is a dialogue between persons in K, then every person utters at least one speech unit. Thus, assumption 2 rules out persons who

remain silent throughout the dialogue.

As an immediate consequcne of assumption 2, let us record the following

Proposition. If $(\alpha,f) \in D$, then $|\alpha| \geq k$.

3. Various types of moments in a dialogue

To express further notions, let us introduce the following notation. For n = 1,2,... let

$$D^n = \{(\alpha,f) \in D: |\alpha| > n\}$$

be the class of all dialogues which contain more than n speech units.

Let now $(\alpha,f) \in D^n$ and $\alpha = \beta\gamma$ with $|\beta| = n$. It follows that the string γ is not empty (since the dialogue does not end at n-th speech unit).

Definition 1. n is a possible termination epoch of the dialogue (α,f), if $(\beta,f') \in D$, where f' is the function f restricted to the set of arguments $\{1,\ldots,n\}$.

Definition 2. n is a regeneration epoch of the dialogue (α,f) if $(\gamma,f'') \in D$, where $f''(j) = f(j+n)$, $j = 1,2,\ldots$.

Intuitively, n is a possible termination epoch if the dialogue (α,f) truncated to the first n speech units is also a dialogue. Similarly, n is a regeneration epoch if the subsequent part of the dialogue (following n-th speech unit) is also a dialogue.

It follows from the above definitions that the intended interpretation is such that a dialogue must constitute a self-contained whole, so that not every fragment of a dialogue is a continuation of the dialogue.

The most common example of r-connectedness is 1-connectedness, where u_{n+1} is simply a reply to a direct question u_n (so that the string β' is empty). For instance, consider the speech units:

u_n = "Have you ever been to Rome?"

u_{n+1} = "No, I haven't".

If u_n were replaced by a unit such as u'_n = "Do you think it will rain tomorrow?" (provided it is admissible in view of the preceding part of the dialogue), then u_{n+1} would not be admissible as an answer to u'_n.

Naturally, an utterance in the dialogue may be r-connected for various values of r, i.e. it may depend on different preceding speech units. This suggests introducing the following two numbers which would characterize to some extent the internal ties in a dialogue.

Let $\alpha = u_1u_2 \ldots u_nu_{n+1} \ldots$ and let $(\alpha, f) \in D$.

Put $A(u_n|\alpha, f)$ = number of indieces r such that u_n is r-connected

and

$$B(u_n|\alpha, f) = \max \{r: u_n \text{ is r-connected}\}.$$

Then $A(u_n|\alpha, f)$ shows how strongly the speech unit u_n is related to the preceding speech units in the dialogue, while $B(u_n|\alpha, f)$ shows how far back into the conversation reaches the connection of u_n.

The functions A and B may be related to the previous concepts, and they may also be used for formulating some general hypotheses about the dialogues.

Firstly, one can conjecture that if m is a regeneration epoch of the dialogue (α, f), then for all $n > m$ one has $B(u_n|\alpha, f) > n-m$. This means that if m is regeneration epoch, then all subsequent speech units are not connected with speech units occurring before epoch m.

An important feature of some dialogues is that they concern some principal topic, and - except for digressions - the dialogue "returns" to the speech unit (or units) in which this topic was posed (formulated, stated, etc.).

One may say that m is an anchor point of the dialogue if there is a set of epochs $n_1, n_2, \ldots$ with $m < n_1 < n_2 < \ldots$ such that u_{n_i} is (n_i-m)-connected. In other words, the speech units $u_{n_1}, u_{n_2}, \ldots$ are all connected with the speech unit u_m. Intuitively, the discussion returns to the theme stated in the speech unit u_m.

This property, despite its strong intuitive appeal, is difficult to express in formal definition, since one would have to specify how long the sequence $n_1, n_2 \ldots$ should be in order for m to be an anchoring epoch.

Finally, the function A shows the strength of the relationships in the dialogue. The extreme case would be a dialogue with $A(u_n|\alpha, f) = 1$ for all n (since if $A(u_n|\alpha, f) = 0$ for all n, α consists of a string of disconnected speech units, and hence is not a dialogue). If $A(u_n|\alpha, f) = 1$ and $B(u_n|\alpha, f) = 1$ for all n, then the dialogue α is such that each speech unit depends only on the directly preceding speech unit, and not on any other units. Such dialogues are simply more or less random chattering. On the other extreme, one may expect that in scientific dialogues, both $A(u_n|\alpha, f)$ and $B(u_n|\alpha, f)$ will be large for all (or most) of the speech units u_n.

It would perhaps be of some interest to make a statistical study of various types of dialogues (e.g. plays) in order to see whether the properties of processes $\{A(u_n|\alpha,f)\}$ and $\{B(u_n|\alpha,f)\}$ $(n = 1,2,\ldots)$ are significantly different for various authors, types of scenes, etc.

5. Eristic dialogues

In this section, the considerations will concern a special category of dialogues, namely eristic dialogues. They are characterized by the fact that the partners (assume for simplicity that there are two of them) try to convince one another about some issue, regarding which they initially differ.

The considerations will necessitate enriching the initial set of basic concepts. To introduce convenient formalisms, let $K = \{1,2\}$, i.e. let there be only two speakers, and let $(\alpha,f) \in D$. Let $v_1, v_2, \ldots$ denote the consecutive stretches of speech units uttered alternatingly by 1 and 2, i.e. $\alpha = v_1v_2v_3\ldots$ where $v_1 = t_1t_2\ldots t_{n_1}$ with $f(1) = \ldots = f(n_1) = 1$, $v_2 = t_{n_1+1},\ldots,t_{n_2}$ with $f(n_1+1) = \ldots = f(n_2) = 2$, and so on.

The first new concepts which will now be introduced are those of the fuzzy set G and the membership function g. The set G will consist of those speech units in which partners in the dialogue "state one's position", and g will be a function which to every speech unit t assigns the grade g(t) of membership in the set G.

Fixing now an arbitrary level ε, one can distinguish the set of all speech units t_j in α for which $g(t_j) \geq \varepsilon$. These are therefore units which contain information about the positions of the speakers.

Consider, for example, the following fragment of a dialogue

between two scientists:

v_1 = "Did you hear about results of Peter's experiments? They confirm his theory rather nicely".

v_2 = "Yes, I know his results. However, I have doubts about their validity, and moreover, I am not quite certain that they confirm his theory. Don't you think they could be explained by another hypothesis?"

Here all speech units in v_1 and v_2 have high grade of membership in G, since they reveal the attitudes of both speakers to Peter's results. However, if the reply to v_1 were:

v'_2 = "No, I haven't heard about them. What are his results?",

then $g(v'_2)$ would be rather small.

The next concept needed for describing eristic dialogues is designed to reflect the differences of opinions between the speakers. These differences may concern many aspects; to formalize this, one may proceed as follows. Let S denote the class of all strings of speech units with degree of membership in G at least ε, i.e. the set of all possible ways of "stating one's position" sufficiently clearly. Let now Δ be the set of binary relations $\delta \subset S \times S$. The symbol $u\delta v$ means that the points of view represented by strings of speech units u and v "differ" by δ.

It is natural to require that relations $\delta \in \Delta$ are symmetric, and that for all u,v there is at least one relation δ holding between u and v. Formally, these requirements may be stated as:

(1) if $u\delta v$, then $v\delta u$;

(2) $\forall\, u,v \;\; \exists \delta \in \Delta : u\delta v$.

Next, the relations δ of "differences of opinions" are ordered (at least partially); one also has to take into account the fact that for any given u, the utterance v need not concern all aspects covered by u. The formal representation of the latter fact requires introducing another ordering relation on Δ. Thus, it will be assumed that there are two relations, $\precsim$ and $\leq$, on Δ, such that

(3) $\delta \precsim \delta$ (reflexivity of $\precsim$);

(4) $\delta \leq \delta$ (reflexivity of $\leq$);

(5) if $\delta \precsim \delta'$ and $\delta' \precsim \delta''$, then $\delta \precsim \delta''$ (transitivity of $\precsim$);

(6) if $\delta \leq \delta'$ and $\delta' \leq \delta''$, then $\delta \leq \delta''$ (transitivity of $\leq$);

(7) if $\delta \leq \delta'$, then $\delta \precsim \delta'$ and $\delta' \precsim \delta$.

Conditions (3) - (6) require no comments. Condition (7) asserts that comparisons $\leq$ can be made only between relations δ and δ' which are equivalent in the sense of the equivalence relation $\sim$ induced by $\precsim$ (i.e. $\delta \sim \delta'$ if and only if $\delta \precsim \delta'$ and $\delta' \precsim \delta$).

The intended interpretation is such that $\delta \precsim \delta'$ means that every issue covered by δ is also covered by δ'. For $\delta \sim \delta'$, assume that $\delta \leq \delta'$, and $u\delta v$, $u\delta' v'$. This means that both v and v' concern the same set of issues as u, and v expresses point of view closer to u than v'.

Moreover, one can assume that in each equivalence class of the relation $\sim$ there is the minimal element 0_δ, representing the complete concensus on all aspects covered by δ.

The most natural interpretation here is when the elements δ are vectors, each coordinate referring to a different semantic

field (i.e. a set of interrelated semantic descriptors). Then $\delta \precsim \delta'$ would mean that δ' contains all coordinates of δ, while (for δ and δ' containing the same coordinates) $\delta \leq \delta'$ would mean that all coordinates in δ are less or equal to those in δ'.

To use an obvious example, one of the semantic fields, corresdonding to one (or perhaps several) coordinates, is the class of all modal frames with which the sentence may be uttered. Imagine, for instance, that one person says "I am absolutely certain that p", to which the other replies "I doubt it", or "I believe it, too". Letting B, B' and B'' be the semantic representations of "I am absolutely certain", "I doubt" and "I believe", the meaning of these utterances may be symbolically represented as B[p], B'[p] and B''[p], where [p] is the semantic representation of p. In this case, $B[p]\,B'[p]$ and $B[p]\delta' B''[p]$, where $\delta = (\delta_1, \ldots, \delta_m)$ and $\delta' = (\delta'_1, \ldots, \delta'_m)$ with $\delta_1 > \delta'_1$ and the remaining coordinates equal, so that $\delta > \delta'$, since the degree of disagreement between "absolutely certain" and "doubt" is more than between "absolutely certain" and "believe".

To illustrate the possible interpretation of other components of δ, let us start from a general scheme of an utterance in which the speaker presents his point of view.

One of the possible schematic representations of such sentences (strings of sentences) may be of the form

[MF] [D/E] [m(a,P)] [Arg]

where [MF] is the modal frame, [D/E] is the deontic or epistemic intention, [m(a,P)] is a schematic representation of an action-sentence "make a to be P", and [Arg] is the argumentation.

As an example, one can think of a sentence such as

$$\underbrace{\text{"I think that}}_{[MF]}\ \underbrace{\text{you should}}_{[D/E]}\ \underbrace{\text{test the hypothesis H again}}_{[m(a,P)]}$$

$$\underbrace{\text{because the results up to now are not conclusive".}}_{[Arg]}$$

In the above representation, [MF] and [D/E] stand for one of the functors about which it is assumed that they form a linguistic representation of motivation (see NOWAKOWSKA 1973a). This assumption is based on observations of functors appearing in sentences used for justifications, evaluations, explanations etc. of made or planned decisions.

Even a very superficial analysis of such sentences allows us to distinguish the subclasses, such as epistemic functors (I know, I am sure, I doubt, ...), emotional functors (I am glad that, ...), proper motivational functors (I want, I prefer, ...) and deontic functors (I ought to, I must, ...).

The functors of particular subclasses are assumed to form representations of different psychological continua, which are used for evaluations in decision situations.

For instance, the class of epistemic functors may be assumed to constitute a linguistic representation of the scale of subjective probability, i.e. different functors reflect various degrees of subjective probability.

Similarly, the class of proper motivational functors constitutes a representation of utility scale.

Some scales are "degenerate", e.g. the one referred to by "I ought to", having basically only two values, namely "I ought to" and "I must", and their synonymes.

Considerations of the scales of motivation leads to intro-

ducing the concept of motivational space. One can then assume that each alternative in a decision situation is evaluated on these scales. It may be conjectured that in some dialogues the persons display some "motivational consistency", understood as domination of one type of scales (e.g. always referring to what ought to be done, or referring to what the person likes, or wants, etc.).

Referring now to the third category in representation of sentences in eristic dialogues, the most useful in categorization seems to be that into approach and avoidance types. Generally, two "approach" actions or two "avoidance" actions, will be considered closer to one another than the actions of "opposite signs", the closeness being reflected in coordinates of δ.

Similarly, in cases of both objects and predicates one can think of "distance" as expressible in terms of coordinates of δ : for instance, two objects are close if they share many of their attributes.

Finally, as regards the argument [Arg], of great interest is the case when it has the form of an ethical, legal etc. rule. The "distance" between rules may be defined in terms of their level of generality, if one of these rules is a special case of the other; two rules are close if one is directly contained in the other.

It is clear that the idea sketched above, to construct the relations δ by distinguishing their components, gives only a crude approximation. In practice, one would have to use many more categories (coordinates), hence a richer classification. The most refined such classification suggested so far is the ingenious and thorough categorization given by Petöfi (1976).

There is no need to list his categories in this paper; denoting them by C_1, C_2,... (the ordering is arbitrary), one could assign to every pair of speech units (or strings of speech units) the set of categories which is relevant for them (i.e. establish the equivalence class to which the relation δ belongs), and then assign the "distances" between opinions on these categories, i.e. determine that $\delta = (\delta_{i_1}, \delta_{i_2}, \ldots)$ which holds between the pair in question.

As mentioned at the beginning, the construction sketched above refers only to those utterances in the eristic dialogue in which the speakers state their opinions. In between there may naturally occur other types of utterances, such as questions, paraphrasing in order to check the intended meaning, and so on.

6. A geometric interpretation of eristic dialogue

A geometric interpretation suggests itself for the above formalism. The vectors δ represent the "distances" between standpoints in the discussion, and the object is to reduce these distances to zero; if this is achieved, the dialogue is resolved. The consecutive distances may be changed by changing the standpoints of one or both of the partners; one can therefore visualize this situation by imagining two points moving in space, their distances decreasing or increasing, until possibly they meet.

The case of a resolution by compromise may be typified by political dialogues, e.g. between ministers of foreign affairs, gradually adjusting their standpoints until they are ready to sign a document specifying the common points of agreement. The extreme case, when there is no compromise, but the dialogue is resolved with one of the partners not changing his initial

standpoint at all, is typified by the argument in mathematics, when one person explains the proof of a theorem to another: if the dialogue is resolved at all, it may be only by one person accepting the truth of the other's theorem.

The geometric interpretation rests heavily on the concept of "distance", i.e. on the possibility of representing the degrees of disagreement, given in form of the vector in terms of the "overall distance".

This possibility, in turn, rests on the empirical problem as to whether the relations δ representing the differences with respect to various semantic fields may be chosen so as to satisfy the axioms of the so-called conjoint measurement (see, for instance, KRANTZ et al. 1971). There is no need to go into the details of axioms of conjoint measurement here; it is enough to say that they specify exactly what kind of "exchangeability" between the coordinates (i.e. distances within semantic fields) is needed for the existence of an overall measure (weighted combination of distances along coordinates).

7. A game-theoretic interpretation of eristic dialogue

Closely connected with the above geometric interpretation is the interpretation of a dialogue as a game described as follows: the game is played sequentially, the successive "moves" being utterances. The "position" in the game is represented by a pair of points in (semantic) space. The object of each player is to move the point corresponding to his opponent into some set in this space.

The strategy is then to use "arguments" ("moves") in such a way as to bring the opponent's point successively to regions in the space from which it is easier to "push" him into the desired set.

It may be debatable whether the theory of semantics is sufficiently developed as to permit the description which would enable one to apply formal solutions of game theory. Nevertheless, there is little doubt that the actual dialogue may be (at least intuitively) viewed as such games, and that reviewing in one's mind all possible arguments which might prove useful in some circumstances amounts to determining the strategy of the dialogue, in the technical sense in which the term "strategy" is used in game theory.

8. Concluding remarks

Clearly, the formalism presented in this paper is far from constituting an adequate theory of dialogues. It seems, however, that the line of research outlined may eventually prove fruitful.

One could argue that the main issue to be resolved first, before one attempts to build a theory of dialogues, is the construction of an adequate formalism for semantics of simpler units, such as words, sentences, texts, etc., along the lines envisaged by, for instance, PETŐFI (1974), since a complete theory of dialogues obviously presupposes a complete theory of semantics.

However, even before such a theory is constructed, one may analyse fruitfully some classes of dialogues which can be expected to have easily identifiable and formalizable issues. Tentative classes of such dialogues (apart from eristic ones) may be political discussions, e.g. between ministers trying to resolve some conflicts; examination and cross-examination of witnesses, especially in US courts, where the rules are laid

down precisely and in great detail; scientific dialogues (notably in form of sequences of articles instead of utterances), and the like.

For any class of such dialogues one may perhaps identify the semantic field of relevance, and introduce the relations δ, characterizing the "distances" between the opponent's opinions.

Finally, one may observe that the following formal analogy between semantics and languages (as defined in mathematical linguistics) may possibly be explored with some fruitful consequences.

Assume that an appropriate semantic theory is constructed, in the sense that every word or sentence has a representation of its meaning in form of a string of symbols of some kind. This would allow us to treat semantic representations in much the same way as one treats sentences, namely as admissible strings formed out of some "semantic vocabulary", say B, consisting of all semantic markers used in a given symbolism. Every meaning is then representable as a string of such markers, i.e. as an element from B^* (the class of all finite strings of elements from B). The meaning need not be unique, so that such a representation establishes a relation between L^{∞} (the class of all strings of speech units) and B^* (or, which amounts to the same, a function mapping L^{∞} into 2^{B^*}, which to every $u \in L^{\infty}$ assigns a set of elements in B^*). The set of all elements in B^* which are related to some $u \in L^{\infty}$ forms a subset of B^*, hence a language in the technical sense of the term. In other words, the "semantic language" would be defined as a subset of the monoid over the vocabulary of semantic formalism, consisting of all those strings which are well formed, and are representable by some strings of speech units. One can conjecture that the grammar of such "semantic language", even though this language may be constructed to some extent in an arbitrary way, will reflect the structural regularities on the non-linguistic reality, i.e. that there will exist some formal similarities between what is "spoken about" and its formal representation.

In particular, since semantic components have various degrees of generality, one can construct in this way a hierarchy of semantic languages, and investigate the properties of the natural language (as well as dialogues) through properties of embeddings of these languages one into another.

References

KRANTZ, D.H., LUCE, R.D., SUPPES, P., TVERSKY, A., Foundations of measurement. New York, Academic Press, 1971.

NOWAKOWSKA, M., Language of motivation and language of actions. The Hague, Mouton, 1973.

NOWAKOWSKA, M., Towards a formal theory of dialogue. Semiotics 17, 1976, 219-313.

PETÖFI, J.S., Some aspects of multipurposes Thesaurus. International Classification (Journal of Theory and Practice of Universal and Special Classification). Pullach b. München 1974.

PETÖFI, J.S., A formal semiotic text theory as an integrated theory of natural languages (methodological remarks). Preprint, Univ. of Bielefeld 1976.

Zur Verwendung der Quotiente in der Textanalyse

G. Altmann, Bochum

1. Zur Charakterisierung der Eigenschaften von Texten werden allzuoft Indizes (Kenngrößen, Maßzahlen) aufgestellt, deren Eigenschaften weiter nicht untersucht werden. Man benützt sie, rechnet mit ihnen, zieht intuitive Schlüsse aufgrund des optischen Eindrucks ("ist die Zahl groß oder klein?") und kommt zu Resultaten, die den reellen Zustand vollkommen verzerren können. Bedenkt man aber, daß das Messen, das Charakterisieren der Eigenschaften keineswegs ein Endziel ist, sondern lediglich eine Vorstufe für Inferenzen, für Ziehung von Schlüssen, für Generalisierungen, für Feststellung von Tendenzen, für Prüfung von Gesetzen, so sieht man sofort ein, daß zu jedem Index eine Auswertungsprozedur gehört, mit Hilfe deren man den gemessenen Wert objektiv interpretieren kann.

Die üblichste "Sünde" in der Linguistik ist die Bildung von irgendwelchen Quotienten ohne Rücksicht auf die linguistischen und die mathematischen Aspekte des gegebenen Indexes. Vor allem in den anfänglichen Stadien der quantitativen Forschung sind solche Indizes besonders beliebt. Zu dieser Sort gehört auch der rechnerisch äußerst einfache BUSEMANNsche Aktionsquotient, mit dem man einen Aspekt der "Aktivität" oder der "Deskriptivität" eines Textes zu charakterisieren pflegt. Er wird definiert als

$$Q = \frac{v}{a}$$

wobei v = Zahl der "aktiven" Verben

a = Zahl der Adjektive im Text ist.

2. Oberflächlich gesehen soll der Index angeben, ob ein Text (Autor) "aktiv" oder "deskriptiv" ist. Findet man mehr Verben als Adjektive, dann ist er aktiv und umgekehrt. Bei dieser Interpretation gerät man sofort in mehrere Schwierigkeiten:

(a) Wenn im Text gleichzeitig sehr viele Verben und sehr viele Adjektive vorkommen, d.h. $Q \approx 1$, dann ist der Text sowohl aktiv als auch deskriptiv. Wenn umgekehrt gleichzeitig wenige Verben und wenige Adjektive vorkommen, dann ist der Text weder aktiv noch deskriptiv, aber $Q \approx 1$. Daher beschreibt der Index <u>keine Eigenschaft</u> sondern einen <u>Gleichgewichtzustand</u> des Textes. Dabei ist offensichtlich, daß dieser Index nur einen einzigen Aspekt der "Aktivität - Deskriptivität" des Textes ausdrückt. Andere Maße sind ebenfalls möglich.

(b) Der Index nimmt Werte im Intervall $\langle 0,\infty)$ an. Der Wert 0 bedeutet "keine Aktivität", aber der Wert ∞ ist philologisch sinnlos. Er sagt nur, daß im Text keine Adjektive vorkamen. In dem Fall bedeutet aber schon das Vorhandensein eines einzigen Verbs eine "unendliche Aktivität", die sich nicht ändert, auch wenn weitere Verben im Text vorkommen.

(c) Die Entscheidung darüber, ob ein Text "aktiv" oder "deskriptiv" ist, folgt aus dem Index selbst nicht. Um wieviel größer als 1 muß Q sein, damit man von "Aktivität" sprechen kann, bzw. um wieviel kleiner als 1 muß Q sein, damit man von "Deskriptivität" sprechen kann? Es besteht keine Symmetrie hier, da bei "Deskriptivität" $Q \in \langle 0,1)$ und bei "Aktivität" $Q \in (1,\infty)$. Mit anderen Worten, die Zahlen haben keine klare Interpretation.

(d) Berechnet man den Index aus zwei Texten, so kann man keine direkten Vergleiche durchführen. Ein einfaches Beispiel zeigt, wie trügerisch Q ist. Sei in einem Text T_1

$$v_1 = 100 \text{ und } a_1 = 2, \text{ woraus } Q_1 = 50$$

und im Text T_2

$$v_2 = 100 \text{ und } a_2 = 4, \text{ woraus } Q_2 = 25.$$

Obwohl der Unterschied zwischen a_1 und a_2 nur 2 beträgt, ist der Unterschied zwischen Q_1 und Q_2 optisch "sehr groß". Die Quotienten sind nur dann direkt vergleichbar, wenn $a_1 = a_2$, d.h. wenn man a festsetzt. Wenn weiter Q_1 und Q_2 nicht gleich sind: wann kann man sagen, daß nicht nur eine zufällige Schwankung in der Stichprobe vorliegt, sondern daß ein signifikanter, systematischer Unterschied zwischen ihnen besteht?

3. Um diesen immerhin interessanten und inhaltsvollen Index anwendbar zu machen und gleichzeitig einige Probleme zu zeigen, die mit Indizes verbunden sind, werden wir einige Aspekte des Aktionsquotienten untersuchen (für einige andere Eigenschaften der Indizes vgl. GALTUNG 1967; SCHEUCH-ZEHNPFENNIG 1974).

Es ist immer ratsam, jedoch nicht unbedingt nötig, die Werte eines Indexes auf das Intervall $\langle 0,1\rangle$ zu beschränken. Dies kann man mit einer geeigneten Transformation durchführen, z.B.

$$I' = I/(1+I)$$

oder

$$I' = \frac{I - I_{min}}{I_{max} - I_{min}}$$

oder

$$I' = I/I_{max}$$

usw. je nach Umständen. Im Falle des Aktionsquotienten wäre eine Möglichkeit

$$Q' = \frac{v}{a + v} .$$

In diesem Fall würden wir $a + v = n$ als die ganze Stichprobe betrachten und Q' wäre lediglich die Proportion von v, d.h. $Q' = p_v$.

Daher würde sich Q' im Intervall <0,1> bewegen und man könnte die ganze Statistik für das Testen der Proportionen anwenden. Q' mißt ebenso wie Q das "aktiv-deskriptive Gleichgewicht", das bei Q' = 0,5 vorhanden ist.

Folgende Hypothesen sind möglich:

(a) Die Hypothese der "erhöhten Aktivität":

$$H_A: p > 0,5$$

gegen die Nullhypothese, daß keine solche Tendenz besteht:

$$H_o: p = 0,5 \quad \text{oder} \quad H_o: p \leq 0,5.$$

In diesem Falle ist zu berechnen

$$P(X \geq v) = \sum_{x=v}^{n} \binom{n}{x} \left(\frac{1}{2}\right)^n . \qquad (3)$$

Setzt man das Signifikanzniveau z.B. auf α = 0,05, so kann man aus (3) folgende Schlüsse ziehen:

Wenn $P(X \geq v) \leq 0,05$, dann ist die Nullhypothese abzulehnen. Man kann annehmen, daß signifikant erhöhte Aktivität vorliegt.

Wenn $P(X \geq v) > 0,05$, dann kann man nicht sinnvoll auf erhöhte Aktivität schließen, d.h. man nimmt H_o an.

(b) Hypothese der "erhöhten Deskriptivität":

$$H_D: p < 0,5$$

gegen die Nullhypothese, daß keine solche Tendenz besteht:

$$H_o: p = 0,5 \quad \text{oder} \quad H_o: p \geq 0,5.$$

In diesem Falle ist zu berechnen

$$P(X \leq v) = \sum_{x=o}^{v} \binom{n}{x} \left(\frac{1}{2}\right)^n . \qquad (4)$$

Wenn $P(X \leq v) \leq 0,05$, dann ist die Nullhypothese abzulehnen. Man kann annehmen, daß signifikant erhöhte Deskriptivität vorliegt.

Wenn $P(X \leq v) > 0,05$, dann kann man nicht sinnvoll auf erhöhte Deskriptivität schließen.

(c) Die Hypothesen (a) und (b) sind einseitig. Jedoch sind einseitige Fragestellungen nur dann korrekt, wenn eine begründete Richtungshypothese vor der Datenerhebung aufgestellt werden kann. Daher handelt es sich in den meisten Fällen um die Hypothese des "aktiv-deskriptiven Ungleichgewichts":

$$H_U: p \neq 0,5$$

gegen

$$H_o: p = 0,5.$$

Wenn Q' > 0,5, d.h. wenn v > a, dann ist (3) zu berechnen und wenn Q' < 0,5, d.h. wenn v < a, dann ist (4) zu berechnen.

Wenn nun $P(X \geq v) \leq 0,025$ oder $P(X \leq v) \leq 0,025$, dann kann man H_o ablehnen und von einem signifikanten aktiv-deskriptiven Ungleichbewicht sprechen. Sonst nimmt man Gleichgewicht an.

Dies ist der übliche Binomialtest, wie man ihn in jedem Lehrbuch der Statistik findet.

Beispiele. (i) Man vermutet in einem gegebenen Text erhöhte Aktivität (in Bezug auf Verben und Adjektive). Die Zählung ergibt v = 20, a = 10. Um die Hypothese zu testen, berechnet man

$$P(X \geq v) = P(X \geq 20) = \sum_{x=20}^{30} \binom{30}{x} \left(\frac{1}{2}\right)^{30} = 0,0494.$$

Da 0,0494 < 0,05, lehnen wir H_o ab und nehmen an, daß im Text signifikant erhöhte Aktivität vorhanden ist.

(ii) In einem Text, für den erhöhte Deskriptivität angenommen wird, findet man v = 13 und a = 17. Um diese Hypothese zu testen, berechnen wir

$$P(X \leq v) = P(X \leq 13) = \sum_{x=0}^{13} \binom{30}{x} \left(\frac{1}{2}\right)^{30} = 0{,}2933.$$

Da 0,2933 > 0,05, nehmen wir die H_0 an, daß keine signifikant erhöhte Deskriptivität vorliegt.

(iii) In einem Text wurde v = 20, a = 10 gefunden. Man teste, ob hier Ungleichgewicht vorhanden ist. In (i) haben wir festgestellt, daß $P(X \geq 20) = 0{,}0494$. Da dieser Wert größer als 0,025 ist, nehmen wir die H_0 an, d.h. es besteht kein Ungleichgewicht. An diesem Beispiel sieht man, daß sich bei der Unkenntnis der Richtung des Ungleichgewichts das Signifikanzniveau halbiert.

4. Die folgende Überlegung zeigt, daß bei ziemlich einfach aussehenden Indizes wie Q die Ableitung eines Tests nicht ganz einfach ist. Manchmal, wenn wir Glück haben, läßt sich die ganze Prozedur beträchtlich vereinfachen.

Sei v die Zahl der Verben in einem gegebenen Text, a die Zahl der Adjektive und r die Zahl aller restlichen Wortarten. Weiter sei die (unbekannte) Wahrscheinlichkeit des Vorkommens von Verben p_v, die von Adjektiven p_a und die der restlichen Wortarten p_r. Es gilt

$$v + a + r = n$$

$$p_v + p_a + p_r = 1. \qquad (5)$$

Philologisch gesehen, stellen wir die Frage, ob im Text ein Gleichgewicht besteht. Übersetzt in die Sprache der Statistik, testen wir die Hypothese

$$H_0 : p_v = p_a$$

gegen

$$H_U : p_v \neq p_a .$$

Zu diesem Zweck stellen wir das sogenannte Likelihood-Ratio

$$\lambda = \frac{L(\omega)}{L(\Omega)}$$

auf, wobei $L(\omega)$ die Maximum-Likelihood Funktion mit den Schätzern $\hat{p}_v$ und $\hat{p}_a$ von p_v und p_a unter der Nullhypothese, und $L(\Omega)$ die Maximum-Likelihood Funktion mit den Schätzern $\hat{p}_v$ und $\hat{p}_a$ von p_v und p_a im ganzen Parameterraum (unter der Menge der Alternativhypothesen) ist. Die Wahrscheinlichkeit, daß wir bei der Untersuchung eines gegebenen Texts genau v Verben, a Adjektive und r restliche Wortarten in der gegebenen Reihenfolge vorfinden, ist

$$L = p_v^v p_a^a p_r^r . \qquad (6)$$

Mit der Nullhypothese "$p_v = p_a$" folgt daraus

$$L = p_v^{v+a} p_r^r . \qquad (7)$$

Da laut (5) $p_r = 1-p_v-p_a$, so wird aus (7)

$$L = p_v^{v+a}(1-2p_v)^r . \qquad (8)$$

Zur Bestimmung eines "besten Schätzers" $\hat{p}_v$ von p_v geht man von dem Bestreben aus, $\hat{p}_v$ so zu bestimmen, daß die in der Stichprobe vorgefundene Wertekonstellation (hier: a, v, r) die wahrscheinlichste ist. Dies ist gerade dann der Fall, wenn L bezüglich p_v ein Maximum annimmt, was mit der üblichen Extremalaufgabe $\frac{\partial L}{\partial p_v} = 0$ zu lösen ist. Es ist jedoch einfacher, statt dessen mit $\frac{\partial \ln L}{\partial p_v} = 0$ zu rechnen, was erlaubt ist, da der Logarithmus seine Extrema an denselben Stellen wie die Funktion hat:

$$\ln L = (v+a)\ln p_v + r \ln(1-2p_v)$$

$$\frac{\partial \ln L}{\partial p_v} = \frac{v+a}{p_v} - \frac{2r}{1-2p_v}$$

$$0 = \frac{v+a}{\hat{p}_v} - \frac{2r}{1-2\hat{p}_v},$$

woraus

$$\hat{p}_v = \frac{v+a}{2n} \qquad (9)$$

folgt. Dasselbe gilt auch für $\hat{p}_a$, wenn man in (7) statt p_v überall p_a schreibt und die Schätzung berechnet.

Setzt man (9) in (8) ein, so kann man schreiben

$$L(\omega) = \left(\frac{v+a}{2n}\right)^{v+a} \left(1 - \frac{v+a}{n}\right)^r . \qquad (10)$$

Die Schätzung von p_v und p_a unter Annahme der Alternativhypothesen finden wir auf dieselbe Weise. Durch Logarithmieren von (6) bekommen wir

$$\begin{aligned} \ln L &= v \ln p_v + a \ln p_a + r \ln p_r \\ &= v \ln p_v + a \ln p_a + r \ln(1-p_v-p_a). \end{aligned}$$

Die Ableitung nach p_v bzw. p_a ergibt

$$\frac{\partial \ln L}{\partial p_v} = \frac{v}{p_v} - \frac{r}{1-p_v-p_a} = \frac{v}{p_v} - \frac{r}{p_r},$$

bzw.

$$\frac{\partial \ln L}{\partial p_a} = \frac{a}{p_a} - \frac{r}{1-p_v-p_a} = \frac{a}{p_a} - \frac{r}{p_r} .$$

Setzt man diese Ausdrücke gleich 0, so erhält man laut (5)

$$\hat{p}_v = \frac{v\hat{p}_r}{r} \qquad (11)$$

bzw.

$$\hat{p}_a = \frac{a\hat{p}_r}{r} . \qquad (12)$$

Addiert man (11) und (12), so bekommt man

$$\hat{p}_v + \hat{p}_a = 1 - \hat{p}_r = \frac{(v+a)\hat{p}_r}{r}, \qquad (13)$$

woraus

$$\frac{\hat{p}_r}{r} = \frac{1}{n} . \qquad (14)$$

Setzt man (14) in (11) und (12) ein, so wird

$$\hat{p}_v = \frac{v}{n} \qquad (15)$$

und

$$\hat{p}_a = \frac{a}{n} . \qquad (16)$$

Daraus folgt nach Einsetzung von (15) und (16) in (6)

$$L(\Omega) = \left(\frac{v}{n}\right)^v \left(\frac{a}{n}\right)^a \left(1 - \frac{v+a}{n}\right)^r . \qquad (17)$$

Aus (10) und (17) bilden wir das Likelihood-Ratio als

$$\lambda = \frac{L(\omega)}{L(\Omega)} = \frac{\left(\frac{v+a}{2n}\right)^{v+a} \left(1 - \frac{v+a}{n}\right)^r}{\left(\frac{v}{n}\right)^v \left(\frac{a}{n}\right)^a \left(1 - \frac{v+a}{n}\right)^r} = \frac{\left(\frac{v+a}{2}\right)^{v+a}}{v^v \quad a^a} . \qquad (18)$$

Die Funktion $-2 \ln \lambda$ ist ungefährt wie Chiquadrat, hier mit 1 Freiheitsgrad verteilt (vgl. KENDALL, STUART 1967, II: 240f.), d.h.

$$\begin{aligned} -2 \ln \lambda &= 2v \ln v + 2a \ln a - 2v \ln \frac{v+a}{2} - 2a \ln \frac{v+a}{2} \\ &= 2 \sum_x x \ln \frac{2x}{v+a} \approx \chi^2_{(1)}, \quad x = a,v. \end{aligned} \qquad (19)$$

Diese Funktion kann bereits für Testzwecke verwendet werden, da 2x ln x tabelliert ist (vgl. KULLBACK, KUPPERMAN, KU 1962). Setzt man jedoch

$$\frac{x - \frac{v+a}{2}}{\frac{v+a}{2}} = \delta$$

so ist

$$\frac{x}{\frac{v+a}{2}} = 1 + \delta$$

d.h. (19) wird zu

$$2 \sum x \ln(1+\delta).$$

Entwickelt man $\ln(1+\delta)$ in eine Taylorreihe und nimmt man nur die ersten zwei Glieder, so bekommt man

$$2\Sigma x \ln(1+\delta) = 2\,\Sigma\, x(\delta - \frac{\delta^2}{2})$$

$$= 2\Sigma\,[(x - \frac{v+a}{2}) + \frac{v+a}{2}]\,(\delta - \frac{\delta^2}{2}),$$

woraus unter der Bedingung, daß v und a groß sind, folgt

$$X^2_{(1)} = \sum_x \frac{\left(x - \frac{v+a}{2}\right)^2}{\frac{v+a}{2}}, \qquad x = a, v. \qquad (20)$$

Entwickelt man das Binom und führt die Summation durch, so ergibt sich schließlich

$$X^2_{(1)} = \frac{(v-a)^2}{v+a}. \qquad (21)$$

Dieses einfache Kriterium ist für einen zweiseitigen Test bei großen Stichproben leicht zu verwenden.

Beispiel. Wenn v = 20 und a = 10, so ergibt sich nach (21)

$$X^2 = \frac{(20-10)^2}{20+10} = 3{,}33.$$

Der kritische Wert $X^2_{0,05;(1)} = 3{,}84$. Da das berechnete X^2 kleiner als 3,84 ist, nehmen wir die Hypothese des "aktiv-deskriptiven Gleichgewichts" an.

Man merke, daß wir den Test ohne die Bildung des Quotienten durchgeführt haben. Will man jedoch den Quotienten Q' bilden und dann testen, so kann man bei großen Stichproben auch folgendermaßen verfahren: Da $\hat{Q}' = \hat{p}$, so ist $E(\hat{Q}') = p$ und $V(\hat{Q}') = pq/n$. Transformiert man $\hat{Q}'$ auf eine Normalvariable, dann ergibt sich

$$z = \frac{\hat{Q}' - E(\hat{Q}')}{\sqrt{V(\hat{Q}')}} = \frac{\hat{Q}' - p}{\sqrt{pq/n}}.$$

Da unter der Nullhypothese p = 0,5, so bekommt man

$$\frac{\hat{Q}' - 0{,}5}{\sqrt{\frac{0{,}5(0{,}5)}{n}}} = (2\hat{Q}' - 1)\sqrt{n} = z \qquad (22)$$

was mit der Wurzel von (21) identisch ist, wenn man $\hat{Q}' = \frac{v}{v+a}$ und n = v+a einsetzt.

Beispiel. Sei wieder v = 20 und a = 10, so wird laut (22)

$$\left[2\left(\frac{20}{30}\right) - 1\right]\sqrt{30} = 1{,}83,$$

ein nichtsignifikantes Resultat. Es ist leicht nachzuprüfen, daß $(1{,}83)^2 = 3{,}33 = X^2_{(1)}$, was wegen $z^2 = X^2_{(1)}$ übereinstimmen muß. Bei diesem kleinen Stichprobenumfang ist jedoch die berechnete Wahrscheinlichkeit vom exakten Wert beträchtlich abweichend.

5. Eines der oft aufgegriffenen Probleme bei der Textanalyse ist der Vergleich zweier Indizes. Die einfache Feststellung, daß beispielsweise bei einem Verfasser $\hat{Q}_1 = 6{,}75$ und bei einem anderen $\hat{Q}_2 = 7{,}05$ ist, reicht nicht aus, um den zweiten Verfasser als "aktiver" einzustufen. Ein statistischer Test ist hier unentbehrlich. Wiederum zeigen wir, wie man in diesem einfachen Fall verfahren kann.

Ohne einen Index zu bilden, bewerten wir die Proportionen von Verben in zwei Stichproben. Wir wissen aus § 3, daß das Vorkommen von Verben einer Binominalverteilung folgt. Unsere Nullhypothese lautet

$$H_0 : p_1 = p_2,$$

wobei p_i den Parameter der Binominalverteilung im Text T_i darstellt. Es kann wiederum drei sinnvolle Alternativhypothesen (wie oben) geben, und zur Entscheidung kann man z.B. mit Hilfe des Fisherschen exakten Tests gelangen (vgl. z.B. SIEGEL 1956: 96).

Bei großen Stichproben ergibt der Vergleich zweier Proportionen, d.h. von $\hat{Q}'_1$ und $\hat{Q}'_2$ einfach

$$z = \frac{(v_1 n_2 - v_2 n_1)\sqrt{n_1 + n_2}}{\sqrt{n_1 n_2 (v_1 + v_2)(a_1 + a_2)}} \qquad (23)$$

wo $n_1 = v_1 + a_1$, $n_2 = v_2 + a_2$. Formel (23) folgt aus

$$z = \frac{\hat{Q}'_1 - \hat{Q}'_2}{\sqrt{V(\hat{Q}'_1 - \hat{Q}'_2)}}$$

durch Einsetzung von $\hat{Q}'_i = \hat{p}_i = v_i/(v_i + a_i)$ und

$$V(\hat{Q}'_1 - \hat{Q}'_2) = V(\hat{p}_1 - \hat{p}_2) = \hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

wo $\hat{p} = (v_1 + v_2)/(n_1 + n_2)$.

Benutzt der Philologe jedoch Q und testet direkt die Differenz $\hat{Q}_1 - \hat{Q}_2$, so muß er vorsichtig verfahren. Will man bei großen Stichproben auf die Normalverteilung übergehen und unter der Nullhypothese das Kriterium

$$\frac{\hat{Q}_1 - \hat{Q}_2}{\sqrt{V(\hat{Q}_1 - \hat{Q}_2)}} = z \qquad (24)$$

verwenden, so muß er den mittleren Fehler einsetzen, der mit $\sqrt{V(\hat{Q}'_1 - \hat{Q}'_2)}$ nicht identisch ist. Es ist

$$V(\hat{Q}_1 - \hat{Q}_2) = V(\hat{Q}_1) + V(\hat{Q}_2).$$

$V(\hat{Q})$ leitet man durch Linearisierung in der Umgebung des Erwartungswerts E als

$$V(\hat{Q}) = V\left(\frac{v}{a}\right) = \left(\frac{\partial Q}{\partial v}\right)_E^2 \cdot V(v) + \left(\frac{\partial Q}{\partial a}\right)_E^2 \cdot V(a) + 2\left(\frac{\partial Q}{\partial v}\right)\left(\frac{\partial Q}{\partial a}\right)_E \cdot Cov(v,a) \qquad (25)$$

ab, wenn man nur die ersten Glieder der Entwicklung benutzt.

Greift man zurück auf die ursprüngliche Multinomialverteilung, wo

$$E(v) = np_v;\ E(a) = np_a$$

$$V(v) = np_v q_v;\ V(a) = np_a q_a$$

$$Cov(v,a) = -np_v p_a,$$

so wird (25) zu

$$V(\hat{Q}) = \left(\frac{1}{np_a}\right)^2 \cdot np_v q_v + \left[\frac{-np_v}{(np_a)^2}\right]^2 \cdot np_a q_a + \frac{2n^2 p_v^2 p_a}{(np_a)^3}$$

$$= \frac{Q^2}{n}\left(\frac{1}{p_v} + \frac{1}{p_a}\right). \qquad (26)$$

Setzt man hier die Schätzung $\hat{p}_v = v/n$ und $\hat{p}_a = a/n$ ein, so bekommt man

$$V(\hat{Q}) = Q^2\left(\frac{1}{v} + \frac{1}{a}\right). \qquad (27)$$

Daraus ergibt sich

$$V(\hat{Q}_1 - \hat{Q}_2) = Q_1^2\left(\frac{1}{v_1} + \frac{1}{a_1}\right) + Q_2^2\left(\frac{1}{v_2} + \frac{1}{a_2}\right) \qquad (28)$$

Da unter der Nullhypothese $Q_1 = Q_2$ ist, schreiben wir

$$V(\hat{Q}_1 - \hat{Q}_2) = Q^2\left(\frac{1}{v_1} + \frac{1}{a_1} + \frac{1}{v_2} + \frac{1}{a_2}\right),$$

so daß (24) zu

$$z = \frac{\hat{Q}_1 - \hat{Q}_2}{Q\sqrt{\frac{1}{v_1} + \frac{1}{a_1} + \frac{1}{v_2} + \frac{1}{a_2}}} \qquad (29)$$

wird. Q selbst kann man als das geometrische Mittel beider Quotienten abschätzen, d.h. als

$$\hat{Q} = \sqrt{\hat{Q}_1 \hat{Q}_2},$$

so daß schließlich

$$z = \frac{\hat{Q}_1 - \hat{Q}_2}{\sqrt{\hat{Q}_1 \hat{Q}_2}\sqrt{\frac{1}{v_1} + \frac{1}{a_1} + \frac{1}{v_2} + \frac{1}{a_2}}} \qquad (30)$$

Beispiel. (i) Sei im Text T_1 $v_1 = 60$, $a_1 = 40$ und in T_2 $v_2 = 50$, $a_2 = 50$. Daraus ergeben sich $\hat{Q}_1 = 60/40 = 1{,}5$ und $\hat{Q}_2 = 50/50 = 1$. Laut (23) ergibt sich

$$z = \frac{[60(100) - 50(100)]\sqrt{200}}{\sqrt{100(100)110(90)}} = 1{,}42$$

und laut (30)

$$z = \frac{1{,}5 - 1}{\sqrt{1{,}5(1)}\sqrt{\frac{1}{60} + \frac{1}{40} + \frac{1}{50} + \frac{1}{50}}} = 1{,}43.$$

Im Falle einer zweiseitigen Hypothese wäre die Gleichheit beider $\hat{Q}_i$ anzunehmen, da $1 - P(-1{,}42 \leq z \leq 1{,}42) = 2\phi(-1{,}42) = 0{,}15560$.

(ii) Setzt man jedoch $v_1 = 40$, $a_1 = 60$, $v_2 = a_2 = 50$, so bekommt man $\hat{Q}_1 = 40/60 = 0{,}67$, $\hat{Q}_2 = 50/50 = 1$ und

$$|\hat{Q}_1 - \hat{Q}_2| = 0{,}33,$$

d.h. "optisch" einen anderen Unterschied als im Beispiel (i), wo $|\hat{Q}_1 - \hat{Q}_2| = 0{,}5$ war. Berechnet man jedoch z, so bekommt man genau dasselbe Resultat wie im Beispiel (i).

6. Diese Überlegungen sollten zeigen, daß das Charakterisieren eines Textes mit einem Index erst dann sinnvoll und weiterführend ist, wenn

(a) der Index eine eindeutige philologische Interpretation hat, d.h. wenn man weiß, was die Zahlen aussagen.

(b) Die philologische Interpretation ist aber nichts anderes als die Übersetzung der statistischen Interpretation in die Sprache des Philologen. Die statistische Interpretation ist lediglich die Überprüfung des Gewichts eines Indexes z.B. in Wahrscheinlichkeitsbegriffen. Daher gehört zu jedem Index eine statistische Prozedur, mit deren Hilfe man den Index testet und über ihn Aussagen macht.

Anhand dieses sehr einfachen Falles wollten wir vor einer allzu leichtfertigen Berechnung irgendwelcher Maßzahlen warnen. Ein Index an sich sagt nichts, er stellt nur eine Verschlüsselung einer von uns angesetzten Qualität dar. In der Textanalyse braucht man letzten Endes keine zahlenmäßige Charakteristiken, sondern objektive, exakte Aussagen über die Eigenschaften des Textes. Die exaktesten Aussagen kann man jedoch eben mit Zahlen machen, die die feinsten Abstufungen zulassen, in ein mathematisches Modell einsetzbar sind und die maximal erreichbare Objektivität gewährleisten.

Literatur

GALTUNG, J., Theory and methods of social research. Oslo, Universitetsforlaget 1967.

KENDALL, M.G., STUART, A., The advanced theory of statistics. London, Griffin 1967.

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 B, 1962, 217-243.

SCHEUCH, E.K., ZEHNPFENNIG, H., Skalierungsverfahren in der Sozialforschung. In: KÖNIG, R. (Ed.), Handbuch der empirischen Sozialforschung. Stuttgart, Enke 1974, Bd. 3a, 97-203.
SIEGEL, S., Nonparametric statistics for the behavioral sciences. New York, McGraw-Hill 1956.

wc
fc
al
hy
vc
Di
bz
Se
bz
Ad
wo

DISTRIBUTIVE-STATISTICAL TEXT ANALYSIS: A NEW TOOL FOR SEMANTIC AND STYLISTIC RESEARCH

Wolf Moskovich, Ruth Caplan, Jerusalem

INTRODUCTION

During the last fifty years a strong and definite tendency has developed in text linguistics, semantics, stylistics and related fields - the building of a theory of text based on the notion of co-occurrence of text units.

In 1931 the great Russian linguist L.V. ŠČERBA called for the creation of a new type of syntactic description that takes into consideration "not only the rules of syntax, but what is more important, the rules of meaning combinations that produce not just sums of meaning but new meanings, rules that are regretfully very little studied by scholars though they are intuitively known to all good stylists" (1, p. 131).

In the structural syntax of A. DE GROOT a clear dividing line is drawn between two independent structures - the sequence of words and syntax. The same point of view that distinguishes between the structural order of text units and their linear order is shared by L. TESNIÈRE. This distinction is clearly expressed in the works of the London School of structural linguistics. J.R. FIRTH writes about colligation - juxtaposition of mutually connected grammatical categories and collocation - context of concrete words. The notion of collocation was defined by FIRTH rather vaguely and was subsequently more exactly defined in quantitative terms by M.A.K. HALLIDAY as "the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur at n removes (a distance of n lexical items) from an item x, the items a, b, c" ...(2).

In the school of American descriptive linguistics the need to use co-occurrences of text elements as a major source of information about covert connections between these elements was clearly realized and became the basis of the whole philosophy of this school. During the 30s and 40s American descriptivists developed and applied more or less effective procedures for identification and classification of text elements. But these procedures used as information only the fact of co-occurrence or non co-occurrence of elements in a text and not data on the frequency of co-occurrence. Though in theory American descriptivists understood the necessity to extract data on co-occurrences out of a large corpus of texts, in practice they relied upon the linguistic intuition of the researcher.

In the theory of text grammars much attention was given to the analysis of the linear structure of text, as opposed to its hierarchical structure (3).But only recently did the notion of co-occurrence and corresponding statistical procedures start to be introduced into the theory of text grammars (4).

It is the theory of distributive-statistical analysis (DSA) that provides the most suitable framework and effective techniques for the systematic study of co-occurrences of text elements. This type of text analysis consists of algorithmic procedures with a wide use of statistics and is based only on quantitative information on objectively delimited text elements.

The results of the application of DSA open new vistas for various fields of literary studies and linguistics, and especially for semantics and stylistics.

MILESTONE IN THE HISTORY OF DISTRIBUTIVE-STATISTICAL ANALYSIS

DSA by nature is a deciphering procedure. The type of questions DSA may help to solve may be characterized in a way suggested by J.G. MEUNIER (5):

Given the linguistic hypothesis of the dependence of a lexical form on its immediate context for the structuring of its meaning;

Given also the linguistic hypothesis of the organic nature of a text, i.e. of the dependence of a lexical form of the text as a whole for the structuring of its meaning;

How can we:

1) Discover contextual and textual dependencies that are pertinent for the meaning of a given lexical form;
2) Characterize its various contextual dependencies in a way that reveals the taxonomy of certain semantic dimensions;
3) Represent in an arbitrary form the meaning of a lexeme in a text, either by indicating its semantic constituents or the semantic field to which it belongs?

DSA was used as a basis for deciphering messages long before it was used for other research purposes. It is difficult to access the level of its development and the scope of its use in cryptography as the corresponding data are rarely published.

In search of the first publications on the use of DSA for scholarly purposes we discovered that apparently the paper by

A.L. BALDWYN, an American psychologist, published in 1942, marks the beginning of scientific publications in this area (6).BALDWYN used the values of co-occurrence of words in the letters written by a female patient as indicators of corresponding connections among ideas in her mind.

The introduction of computers in the early 50s produced an increased interest to co-occurrence analysis, especially in the USA. In 1955 an electric network of associative links between words and documents was exhibited in the USA. The network was built as a result of analysis of co-occurrences of 250 terms in 100 documents (7).This network EDIAC (The electronic display of indexing association and content) was built by a private firm on contract for the information service of the U.S. Air Force. Very soon the importance of co-occurrence analysis for information retrieval purposes became clear and various groups of researchers started working on the improvement of the techniques of co-occurrence analysis. L.B. DOYLE has done many co-occurrence counts on the material of various samples of text. He studied co-occurrence of words belonging to different parts of speech and situated at different distances from one another in the text (8).Special techniques for transition from the discovery of syntagmatic links among co-occurring words to paradigmatic links between them were developed by H.E. STILES (9).

Starting in 1957 intensive research on co-occurrence analysis has been conducted in Great Britain, at the CAMBRIDGE LANGUAGE RESEARCH UNIT (10).In these studies co-occurrence of words in only one interval of text - the text as a whole - is analysed. The mathematical apparatus used in these studies is based on techniques suggested by T.T. TANIMOTO (11).

Since 1960 the number of papers on DSA has increased rapidely. Publications came almost exclusively from three countries - the USA, the USSR and Great Britian. In the period 1960-1965 DSA was enriched by the introduction of the procedures of factor analysis and latent class analysis (12,13).

Most of the research done in this period was concerned with investigating the feasibility of the use of DSA for the building of automatic term structures of the thesaurus type for information retrieval purposes. The results obtained by different research teams seemed to be promising, though it became clear that huge financial expenses are connected both with the input of large bodies of texts into the computer and with subsequent machine co-occurrence analysis, which frightened off many researchers. After 1965 the number of papers on DSA decreases. Some fundamental research projects based on large samples of texts were carried out in the period between 1965 and 1970 (14,15 16,17,18).

In their review of research on automatic term structures carried out in this period K.S. JONES and M. KAY conclude the following: "From the linguistic viewpoint, the interesting aspect of statistical associations is the character of the derived list of classes, compared with those typical of manually organized vocabularies. In general ... words brought together by statistical criteria are linked by very varied semantic relations, and sets of manually associated words are not usually dependent on single or well-defined meaning relationships. Further, many of these relationships are highly local, that is, collection specific" (19, pp. 169-170).

Though most of efforts in the development and application of DSA in Western countries was concetrated in information retrieval, extremely important research in the period 1965-1970 was also done in linguistics proper. In one such study synonymous word pairs were successfully distinguished from non-synonymous pairs by their patterns of co-occurrence with other words (20). In other studies co-occurrences of syntactically connected words in the text were analysed and sets of words with common contextual properties exhibited a fair degree of correspondence with intuitively delimited semantic fields (21,22,23). K. HARPER counted the co-occurrence of nouns with their governing and subordinate words in sentences in a sample of 120,000 words of Russian text on physics. Semantic distance between any two nouns was defined as the ratio of the quantity of governing and subordinate words common for both of them to the product of the frequencies of these nouns in the text. Semantic groups of nouns were discovered which showed a good approximation to intuitively felt groups of semantically related words. In some recent studies, co-occurrence of nouns with the same verbs as subjects (or objects) was used as a measure by which nouns and verbs were grouped into classes. These classes largely corresponded to semantic classes derived manually (24,25).

The development of DSA as a purely linguistic method is characteristic for the research conducted in the USSR. For different reasons that we shall not discuss in this paper a pragmatic approach to automatic term classification was not followed in the USSR from the very beginning and has begun only recently. Instead only theoretical linguistic research was conducted. In 1959 N.D. ANDREEV suggested and tested several

statistical procedures for measuring the syntagmatic and paradigmatic proximity of linguistic units (26). Afterwards ANDREEV's ideas were developed mainly in the direction of DSA of morphology.

The very term "distributive-statistical analysis" was coined by A.J. ŠAJKEVIČ in 1963 and since that time has been widely accepted in Soviet linguistics. ŠAJKEVIČ's first studies on DSA started in the early 60s with experiments on the discovery of semantic fields of language on the basis of statistical analysis of co-occurrence of words in text (27). This work was of great methodological importance as the high level of its subsequent citation in Soviet linguistic literature shows. ŠAJKEVIČ and other Soviet linguists widely used DSA for lexico-semantic and stylistic studies (28,29,30,31,32,33,34,35). Two findings seem to be important results of these studies: 1) the most effective distributive-statistical procedure is one that is based not only on data on the frequency of co-occurrence of text elements but on the measurement of the degrees of deviation of observed co-occurence from theoretically expected co-occurrence of text elements; 2) with the widening of the interval of text in which co-occurrence data are registered the character of the linguistic facts being discovered changes in a certain direction.

Though the development of DSA is an international achievement, it was in the USSR that the experiments and supporting theoretical work were carried out as a purely linguistic enterprise.

Since the early 70s a similar tendency has been observable in Western countries. In France it was the research group at the Ecole Normale Supérieure de Saint Cloud that started to use DSA for the analysis of the contents of political texts (36,37). A distributive-statistical procedure connected with the theoretical framework of FIRTH's ideas on collocation was developed by G.L.M. BERRY-ROGGHE in Great Britain(38). In Canada two teams of researchers are working on the analysis of philosophical texts with the help of DSA (5,39,40). In Germany B. RIEGER has applied the theory of fuzzy sets to the interpretation of semantic networks obtained as a result of DSA (4).

In contrast to the USSR, where due to the efforts of ŠAJKEVIČ, ANDREEV and other linguists DSA is being developed as a strict and coherent linguistic theory, research done in the West is more heterogeneous and is still often regarded as a simple extension of experiments in artificial intelligence and information retrieval.

THE STANDARD DISTRIBUTIVE-STATISTICAL PROCEDURE

The information about text elements and their distribution in the text may be divided into two categories:

1) information on the distribution of elements relative to the text itself and its parts (textual distribution of elements);
2) information on the distribution of an element relative to another element (reciprocal distribution of elements).

In order to analyse the textual distribution of elements the text is divided into a number of segments. The frequency of a given text element is counted for every segment. Then the number of segments in which this element appeared 0, 1, 2, 3, ...,n times is established (during this operation the original sequence of text segments is broken). Generally the most basic parameters of the distribution (such as the mean and dispersion)

suffice for a classification of text elements. For example, in the texts of U.S. patent formulas published in the bulletin "Official Gazette" (sample I, 262,925 words) among the most frequent words two classes are discovered. The first class is characterized by a high dispersion and includes names of various objects: machines and apparatus (pump, etc.), their parts and details (axis, walls, cylinder, port, etc.), physical phenomena and functions (combustion, pressure, etc.), substances (air, fuel). The second class consists of words with a low dispersion. Almost all of them are grammatical words: articles, prepositions, copulas, conjunctions.

Some specific form-words characteristic for the language of patent claims fall into the same category: words indicating the reciprocal position of details (connected, disposed, opposite), their movement (movement, rotation), specific patent 'pro-forms' (means, device, etc.).

Patent texts have a highly peculiar grammar that is based not only on regular morphological expression, but on semantic properties of some specific patent form-words as well (33).

In W. Shakespeare's comedies (sample II, 276,000 words) on the other hand the picture of the distribution of the most frequent words is quite different. In the group of words with a high dispersion we find articles, pronouns, names of persons (father, wife, daughter, son) and almost all names of characters with the exception of characters to whom all of the other characters of a comedy often refer (e.g. Angelo in 'Measure for Measure' of Falstaff in 'The Merry Wives of Windsor').

In order to check the possibility of delimiting names of persons in a text of belles lettres words of high frequency and high dispersion were analyzed in some of A. Pushkin's poems. In 'Evgeny Onegin' the words ya ('I'), on ('he'), ona ('she'), Tatyana, Olga, Lensky and in 'Ruslan and Ludmila' the words ya, on, ona, Ruslan, knyaz' 'prince', Naina, Knyažna 'a princess', xan 'khan', karla 'dwarf', golova 'head' have a high dispersion. But the names of the central heroes of the poems have a low dispersion - Evgeny, Onegin, Ludmila. Thus this method of delimiting semantic groups of words seems to be rather weak and it allows one to divide the whole list of words into only two groups (35).

Another way of analyzing the textual distribution of elements is based on the counts of occurrences of elements in certain positions in a text. This way of analysis may be called 'positional' analysis. In the text some evident segments are established (the beginning or the end of a segment). The frequencies of occurrence of elements in these positions are established which are then compared to their mathematical expectation (in other words the conditional probability of occurrence of an element in a certain position is compared to the independent probability of its occurrence in this position).

This method of analysis is widely used by ANDREEV to discover morphological classes in Russian and other languages(26).Four statistical-combinatorial classes of words were discovered in Russian: 1) substantives and cardinal numerals; 2) declinable adjectives, participles and those pronouns that have different forms for the three genders; 3) verbs, verbal adverbs, predicative adjectives and participles; 4) invariable words.

These parts of speech established by DSA show a striking similarity to the traditional classification of parts of speech by Russian grammarians.

Positional analysis is successful in text segments that have a clear structure (e.g. in sentences). In larger texts the semantic structure is not as evident and the results are not as good (35).

For example, in Shakespeare's comedies (sample II), two positions - the beginning of a comedy (the first thousand words) and the end of the comedy (the last thousand words) were checked. The beginning of the play, where the exposition is given and the main heroes and place of action are introduced are characterized by the following words: father 20 (10), brother 14 (6), there 35 (22), in 139 (118) (the first figure indicates the average frequency of a word per 10 thousand words in the beginning of a play and the second - in parenthesis - its frequency in the whole text). In the beginning of plays there is sometimes a story about preceding events told by the heroes which explains the high frequency of pronouns we, us, ours - 109 (67).

The end of Shakespeare's comedies is more dynamic than their beginning. Dialogue here prevails over monologue, and this is reflected in the high incidence of the pronouns you, your, yourself - 323 (286). The most characteristic feature of Shakespeare's comedies - one that is connected with comedies in general - is the concentration of words connected with weddings: husband 10 (5), wife 17 (6), hand 16 (7), ring 23 (4), (where the first figure refers to the average frequency of a word per 10 thousand words at the end of a play and the second to the frequency in the whole text). Though positional analysis produces results interesting for stylistics and semantics, the analysis of mutual distribution of text elements is much more promising a procedure.

The standard procedure of DSA used in most of the Soviet studies is described here in the form suggested by ŠAJKEVIČ (35).

Semantic associations between words in a text are discovered by comparing the real frequency of their co-occurrence (x) in a certain interval of text with the mathematical expectation of their co-occurrence in the same interval. Mathematical expectation is computed with the assumption of independence of the words in question

$$m_{ab} = \frac{f_a \ f_b}{n} \qquad \text{(I)}$$

where f_a, f_b - frequencies of words a and b, and n - the number of intervals into which the text is divided. The relation of x_{ab} (real co-occurrence) to m_{ab} makes it possible to measure the strength of connection between two words. There may be different quantitative expressions of this connection. It is simpler to measure the strength of the connection by the relation x/m. But such an index is inconvenient because under small values of m the relation x/m may be large even if no statistical significance exists between x and m - it is the same when x = 700, m = 200 and when x = 7 and m = 2, though in the first case the difference between x and m is very large and in the second it is statistically negligible. That is why usually not the strength of the connection but the statistical significance of such a connection is taken into account.

The excess of x over m is expressed by

$$z = \frac{x - np}{\sqrt{npq}}, \qquad \text{(II)}$$

where n is the number of intervals, $p = \frac{f_a f_b}{n^2}$, $q = 1 - p$ and $np = m_{ab}$. This is the usual transformation of a binomial variable to the standard normal variable (on the use of formula II in some Western distributive-statistical studies see 36, 38).

Since n is usually great and p very small, one can safely use the Poisson distribution which is the limit of the binomial and compute the probability of the observed or a more extreme value of x_{ab} as

$$P\,(X \geq x_{ab}) = \sum_{x=x_{ab}}^{\infty} \frac{m_{ab}^{x} e^{-m_{ab}}}{x!}$$

where e is the known mathematical constant with the approximate value of 2,71828.

At small values of m tables of the Poisson distribution may be used. If m is great, say m>30, then, again, the Poisson variable can be transformed to the normal variable by means of the formula

$$z = \frac{x_{ab} - m_{ab}}{\sqrt{m_{ab}}} \qquad \text{(III)}$$

The probabilities of some values of z are as follows

z	$1 - \Phi(z) = P$
1	0,1587
2	0,0228
3	0,00135
4	0,000032
5	0,00000029

We choose the following arbitrary decisions for the connection between a and b:

Interval	Connection
$1 \leq z < 2$	very weak
$2 \leq z < 3$	weak
$3 \leq z < 4$	medium
$4 \leq z < 5$	strong
$z \geq 5$	very strong

Let $[1 - \Phi(z)]/2 = 1 - F(x_{ab}-1) = P(X \geq x_{ab})$, where $\Phi(z)$ is the cumulative distribution function of the normal variable and F that of the Poisson variable. Then, for example, at $m_{ab} = 1{,}0$ x_{ab} and z are connected approximately in the following way

z	x_{ab}
1	3
2	4
3	6
4	8
5	10

This means that when we expect $m_{ab} = 1{,}0$ but come accross $x_{ab}=10$, the connection is very strong. At $m_{ab} = 100$ to receive the same strength of connection x_{ab} had to be higher than 150. For example, in Shakespeare's comedies the frequency of burn is 41, of fire - 88, of hot - 37, of light - 117 and n = 370000. If the interval is the line then we have

$$m_{burn,fire} = \frac{41(88)}{37000} = 0{,}10.$$

Since the real co-occurrence is $x_{burn,fire} = 5$ then

$$P(X \geq 5) = 0{,}0000000767 = [1-\Phi(z)]/2$$

from which $z \approx 5{,}12$, i.e. these words are statistically very strongly connected.For the two other pairs we find: $m_{burn,hot} = 0{,}04$, $x_{burn,hot} = 2$, $z \approx 3$; $m_{burn,light} = 0{,}13$, $x_{burn,light} = 3$, $z \approx 3$, the strength of these connections is medium.

Two text elements are connected paradigmatically if both of them are syntagmatically connected with some third text elements. It is reasonable to assume that the strength of a paradigmatic connection increases with the rise of the number and intensity of common syntagmatic connections of the two text elements under consideration.

The transition from syntagmatic to paradigmatic connections may be achieved in a number of alternative ways. One of the formulas successfully tested so far is the following:

$$T_{ab} = z_{ab} + \sum_i \lg(z_{ac_i} \cdot z_{bc_i}) \qquad \text{(IV)}$$

where a and b are the text elements (morphemes, words of fixed word combinations) whose paradigmatic connection is measured and c_i - the text elements with which both a and b are connected syntagmatically.

It was established experimentally that for the text intervals between 2 and 50 words the first member of the right side of this equation may be omitted and the following variant of the formula may be used (33):

$$T_{ab} = \sum_i \lg(z_{ac_i} \cdot z_{bc_i}) \qquad \text{(IVa)}$$

Two limitations for the use of formula IV are suggested: $(z_{ac_i} \cdot z_{bc_i}) > 1$ and the number of common syntagmatic connections between a and b should be more than two, i.e. $n_{c_i} > 2$ (33).

Let us illustrate the use of this formula on the following example. For the words lower and upper the list of common words c_i (in the interval of 2 words in technological texts) includes: edge, end, ends, plate, portion, surface, wall.

The values of z_{ac_i} and z_{bc_i} for the strength of syntagmatic connections of each of these words with the words lower (a) and upper (b) are the following:

	lower	upper
edge	8,5	4,1
end	10	10
ends	7,5	10
plate	4,1	3,4
portion	10	10
surface	6,3	3,6
wall	4,7	4,1

Then, T_{ab} is 11,2. The derivation and a more detailed discussion of the formula IV may be found in (33).

The first stage of DSA produces the so-called first-generation profiles of text elements reflecting both syntagmatic and paradigmatic connections between them. The second stage produces the so-called second-generation profiles where only paradigmatic connections between text elements are retained.

The strength of association between the text elements a and b has been measured in several other ways P. JONES and R. CURTICE (41) have shown that almost all (if not all) of the corresponding formulas are based on the single theoretical foundation offered by the 2 x 2 contingency table.

"The argument behind a typical association measure, when developed along statistical or theoretical lines, offers little or no basis for distinguishing among them. The reasoning suggests comparing the number of observed co-occurrences with the calculated number of expected co-occurrences. Given two formulas, it will generally be found that substantially the same supporting rationale is preferred for both; there are many ways of measuring statistical surprise or the unexpectedness of an observation, and a large number of the available formulas can responsibly claim to do so" (41, p. 153).

According to P. JONES and R. CURTICE the following formulas are mutually reversible:

$$A_{ab} = x_{ab} \qquad (V)$$

$$A_{ab} = x_{ab} - \frac{f_a f_b}{n} \qquad (VI)$$

$$A_{ab} = \frac{x_{ab} - \frac{f_a f_b}{n}}{\sqrt{\frac{f_a f_b}{n}}} \qquad (VII)$$

$$A_{ab} = \log_{10} \frac{\left(|nx_{ab} - f_a f_b| - \frac{n}{2}\right)^2 \cdot n}{f_a f_b (n-f_a)(n-f_b)} \qquad (VIII)$$

$$A_{ab} = \frac{x_{ab}}{f_a + f_b - x_{ab}} \qquad (IX)$$

These formulas are discussed in detail in (41). The derivation of VIII is given in (9). The formula IX represents the number of co-occurrences of two text elements normalized by the number of occurrences where each of these two elements appears separately(41).

The results of calculations of the strength of connection according to these formulas of the term Rocket Motor Case with other terms in a sample of 100 thousand documents are shown in Table 1.

Table 1. 15 terms the most closely connected with the term Rocket Motor Case according to different evaluations.

No.	Formula V	Formula IV	Formula VII	Formula VIII	Formula IX
1	Case	Case	Case	Case	Case
2	Motor	Motor	Motor	Motor	Motor
3	Rocket	Rocket	Winding	Winding	Winding
4	Rocket Engine	Rocket Engine	Filament Winding	Rocket	Filament Winding
5	Solid	Solid	Deep Draw	Filament Winding	Filament
6	Propellant	Propellant	Rocket	Stretch Forming	Closure
7	Steel	Steel	Stretch Forming	Deep Draw	Fiberglass
8	Fabri- cation	Fabri- cation	Hydrotest	Hydrotest	Glass Fiber
9	Winding	Winding	Filament	Filament	Stretch Forming
10	Filament	Filament	Rocket Engine	Rocket Engine	Stretch
11	Solid Prop.	Solid Prop.	Closure	Closure	Solid Prop.
12	Filament Winding	Filament Winding	Fiberglass	Fiberglass	Rocket Engine
13	Titanium	Titanium	Stretch	Stretch	Reinforced Plastic
14	Fiber	Fiber	Spiral Wrap	Steel	High Strength
15	Glass	Glass	Steel	Fabrication	Fabrication

The table shows 15 terms most closely connected with the term Rocket Motor Case that are ranked according to the order of their strength of connection with this term calculated with the help of different formulas. It is obvious that different formulas provide similar results - the lists of terms are very

similar and there are certain differences in their ranking. The pairs of formulas V and VI and VII and VIII provide two slightly different sets of results. The list of terms received with the formula VIII differs more significantly from the lists obtained with V and VI. Technical terms like Hydrotest, Deep Draw and Closure begin to be included. But it is extremely similar to VII with only minor permutations of terms. It may be assumed that formulas of the type VI, VII, VIII and IX that are based on the evaluation of the statistical significance of the excess of the observed co-occurrence over expected co-occurrence are more preferable than the formula V based solely on the value of the observed co-occurrence as they provide a procedure for filtering the most significant connections. Among the different formulas of this type there seems to be very little difference as to the results obtained but conclusive empirical evidence is still lacking.

DISCOVERY OF LATENT SYNTACTIC AND SEMANTIC CLASSES BY DISTRIBUTIVE-STATISTICAL TECHNIQUES

An important state of affairs encourages the wide use of distributive-statistical analysis in linguistics and literary analysis: text intervals of different size give different information.

Text interval means the length of a piece of text fixed for a certain stage of analysis and serving as a basis for all calculations of mathematical expectation and frequency of co-occurrence of text elements.

Transition from one interval to another makes it possible to obtain information of a new kind on every interval.

A tentative list of optimal intervals contains the following:

1) the microinterval;

2) the minimal interval;

3) the small interval;

4) the medium interval;

5) the large interval;

6) the maximum interval (cf. 47).

The microinterval coincides with a word in a text, and the maximum interval - with the text itself as a whole. Between these two intervals tentative arbitrary divisions are made: minimal interval: 2 - 5 words on a text, small interval: 5 - 50 words, medium interval: 50 - 500 words, large interval: 500 - 2000 words. If the text element chosen for analysis is a graphic word, then in the microinterval we obtain morphological information, on the minimal interval - mainly syntactical and phraseological information, in the small interval - lexico-semantic information, in the medium interval - lexico-semantic and thematic information, on the large and maximal interval - thematic and stylistic information.

The algorithm of the distributive-statistical analysis is iterative: once the information on a larger interval is obtained, the algorithm returns to a smaller interval and the results obtained there are checked and corrected after comparison with the results obtained on a larger interval. The iteration is finished only then when it ceases to give new information. Usually it is one to two iterations. For example, in the micro-interval tentative paradigms are formed on the basis of analysis of inner strings of letters in words, then distributive classes of words are studied on the minimal interval; and after that the

information on distributive classes is used for the correction of paradigms. In the microinterval DSA is able to discover morphemes, regular paradigms and some classes of stems and affixes.

In the minimal interval distributive classes of words are obtained and fixed word combinations are discovered. The feasibility of the use of DSA for discovery of morphological paradigms was shown many times on the material of various languages (26,33 42,43). In contrast to the classical distributive analysis where similarity (or difference) of distribution of two groups of words is established on the basis of the fact of their co-occurrence or of the absence of their occurrence, in DSA we ask another question: 'How often do these groups of words co-occur?'. The results of the pure distributional procedure of DSA are usually not perfect but quite satisfactory.

In experiments in the microinterval experimental paradigms are formed on the basis of the analysis of strings of letters and are defined as sets of affixes co-occurring with the same stems. In the minimal interval distributional classes of words are discovered. Such distributional classes roughly correspond to syntactical classes and/or parts of speech of traditional linguistics. The experimental paradigms may be further checked and corrected in accordance with the distributional classes. If identical suffixes belonging to different experimental paradigms are distributionally similar, they may be united into a common paradigm. In one of the experiments 43 paradigms of the English language were reduced to 23 after comparison with their distributional characteristics (44).

DSA in the minimal interval is often used for studies in grammar with co-occurrences of words being used as criteria for establishing syntactic classes and syntactic connections. In some cases the text intended for analysis is first subjected to pre-computer text marking. Three types of markings are used: 1) dividing lines between morphemes; 2) grammatical and semantic indicators of word forms; 3) indicators of syntactic roles and relations. On the basis of such preliminary markings (syntactic class, subclass, governing word, the type of syntactic connection between words) an automatic grammar for English texts was build (45). In one of a number of similar studies, researchers tried to clarify the following question: do the semantics of the components and the type of syntactic links between them condition the morphological structure of noun combinations in modern Ukrainian? For the configuration N Nominative case plus N Genitive case, when the governing component designates various appliances and is characterized by suffixes - -ač or -tor, the following syntactic-semantic links between components are predicted: 1) 'to have as an object of action' - 70%; 2) 'to be a part' - 20%; 3) 'to be an instrument' - 10% (46). The use of pre-computer text marking broadens the possibilities of DSA and links it to other linguistic techniques. But the introduction of preliminary marking into the text to be subjected to DSA contradicts the very nature of DSA as a discovery procedure and shifts more work from the computer to the researcher. That is why in cases where there are no limitations on computer processing time DSA is used as a discovery procedure on an unmarked text and the results of the procedure are subsequently subjected to interpretation and evaluation in terms of this or another grammatical classification.

In the minimal interval mostly syntagmatic links between words are discovered. For example, in modern English prose the

adjective black significantly co-occurs with such nouns as eyes, hair, smoke, etc. and the adjective grey with nouns hair, beard, moustache, etc. On the basis of this information second-generation profiles reflecting paradigmatic links between words can be successfully discovered (30). In the case of black and grey a strong paradigmatic relation is established reflecting a marked overlap between the lists of nouns with which each of these adjectives co-occurs.

In the small interval paradigmatic links are already discovered regularly in first-generation profiles. Complex macrosemantic maps can be built and the structure of the lexicon as a whole and its individual semantic fields in particular studied. Let us analyze a macrosemantic map of English adjectives built as a result of DSA in the interval 'a line of text' on a sample of English poetry and drama (Sample III). This sample of English texts has a length of 2 million words (more than 250 thousand lines) and includes words of Chaucer, Spenser, Shakespeare, Milton, Wordsworth, Shelley, Arnold and Marlowe. In these texts co-occurrences of adjectives with an absolute frequency of 15 and higher are studies on the condition that at least two authors use this adjective. Out of 1078 adjectives subjected to analysis 442 were discovered to have some semantic links with other adjectives. In general 763 semantic links were discovered (27).

Before analyzing this material the influence of outside factors that may introduce a certain degree of distortion should be eliminated. In the Sample III such a factor is alliteration. In order to eliminate the influence of alliteration, the figures of co-occurrences of adjectives in which first letters are the same were reduced by 1/4. After the introduction of this correction the number of adjectives having some kind of semantic links was reduced to 426 (with 715 semantic links). The more frequent an adjective is, the more semantic links with other adjectives it has. If we divide the whole list of adjectives into three parts: a) with a frequency of 100 and more; b) with a frequency of 40 to 100; c) with a frequency of 15 to 40, we discover that the number of adjectives having semantic links in the first group is 248 out of 324 (or 77 %), in the second group - 120 out of 297 (or 40 %) and in the third - 60 out of 411 (or 15 %). The number of words having semantic links in individual authors is smaller than in the unified list (Wordsworth and Shelley - each 197, Spenser - 213, Shakespeare - 272), while in the unified list of eight authors the number is 426.

Only a small part of the macrosemantic map is organised into absolutely isolated pairs and groups of words; cheerful - cheerless; Christian - pagan; female - male; fleshly - ghostly; inward - outward; Greek - Latin; almighty - vengeful, etc. Most of the map represents semantic groups of words having strong connections between elements of the group and weak connections outside the group. Some examples of the semantic groups are: 1) semantic field of *LIFE* with central words dead and alive; 2) *HIGHNESS* with central words high and low; 3) *GENERAL SIZE* with central words great, huge and small; 4) *THINNESS*: thin, thick and subtle; 5) *PRIDE*: proud and humble; 6) *IMMORTALITY*: mortal and immortal; 7) *STUBBORNNESS* and *SULLENNESS*: sullen; 8) *DUMBNESS*: dumb, mute, deaf; 9) *VOLUME*: deep, hollow, wide, broad, large; 10) *HEAVINESS*: dull, heavy, light, slow, swift; 11) *TEMPERATURE*: and *DAMPNESS*: cold, hot, dry, wet, moist; 12) *WEAKNESS*: faint, feeble, weary, sick, whole, tedious; 13) *SOFTNESS*: tender, sharp, soft; 14) *COLOUR*: blue, white, black, red, yellow, green, grey,

soft; 14) *COLOUR*: blue, white, black, red, yellow, green, grey,
brown, pale purple.

In the small interval all the essential paradigmatic connections between words are easily discovered: synonymy (dumb - mute; weak - faint - feeble; vile - base - mean; foul - filthy - loathsome; gentle - mild), antonymy (mortal - immortal; new - old; great - small; dead - alive; strong - weak) and chains of words with related meaning (green - fresh - new; lofty - divine - earthly - heavenly - sacred; heroic - brave - honourable - noble).

The semantic network of 442 words is divided into 53 groups and 33 pairs of words. The division of the network into individual semantic fields is connected with considerable difficulties as the techniques of processing such complex graphs are not yet satisfactorily developed. Nevertheless the criterion for the delimitation of semantic fields always remains the same: the quantity and the combined strength of links inside the group must be greater than the quantity and the combined strength of outward links. A common semantic denominator cannot be found for all semantic fields though semantic links among these words may be intuitively justified.In the Sample I of technological texts these results were fully corroborated. In the small interval a number of semantic fields were discovered: 1) a group of adverbial geometrical indicators: axially, radially, transversely, vertically, horizontally, circumferentially, laterally, longitudinally, angularly, forwardly, respectively, rearwardly; 2) a group of adjective geometrical indicators: angular, axial, radial, lateral, longitudinal, pivotal, relational; 3) designators of degree and measure: predetermined, low, high, constant, ambient; 4) a group of movement and position: movement, direction, displacement, position, side, axis, rotation, motion, plane; 5) a group of 'matter': air, gas, fluid, liquid, fuel, water; 6) a group of 'openings and passages': chamber, line, opening, conduit, aperture, bore, passage, passageway, port, nozzle, duct, etc. 312 words out of 2468 words in this sample are connected by strong paradigmatic links (the general number of second-generation paradigmatic links is 527) (33).

Because the limits of the small interval (5 - 50 words) apparently coincide with the length of the most natural unit for the study of co-occurrence of words - the sentence, this interval has been tested experimentally much more thoroughly than any other intervals. Since all the syntagmatic and most of the paradigmatic relations between words are actualized within the sentence it is to be expected that the small interval provides the main part of all the co-occurrence information which can be extracted from a text.

The medium interval gives the same kind of information as the small interval but its first-generation word profiles reveal more vividly and fully paradigmatic links between words. As the limits of this interval correspond to those of a paragraph or an abstract the so-called thematic information connected with the subject of texts is discovered in some detail(32).

The large and maximum intervals provide information of thematic and stylistic character and will be analyzed in the following section of this paper.

DSA discovers classes of linguistic units that may be called (following B.L. WHORF) "covert" linguistic classes. But what is the quality of such classes discovered by DSA and to what reality

do they correspond? To provide an answer to this question is extremely difficult for two reasons. First, at this stage of the development of linguistics there exist no objective criteria for the evaluation of the quality of "covert" classes. Second, the results of DSA strongly depend upon the character of text chosen for analysis. Various texts strongly differ from one another by their lexicon and lexico-semantic links among words. That is why in principle DSA should give different semantic groupings for the same list of words depending on the character of the analyzed text. If we study texts that are divided, for example, into 500 different groups, we may expect that DSA will reveal 500 semantic pictures similar to one another and at the same time having certain differences. Let us compare the results of computations carried out with the use of the same formula, but on three different samples of texts(30).Three different pictures of colour names in modern English were obtained with points of coincidence and difference between them. In all three pictures the words black and gray, red and purple were linked, in two of the pictures the word white was connected with yellow, red, green, pink and the word golden with yellow. But in other features the three pictures are different from one another. For example, in one of them golden is linked with yellow, white and rosy, but in another with dark, black and yellow. This is explained by the different thematic character of the samples of text: in the first of them, a poetic text, the scope of collocation of the word golden with different nouns is extremely wide. In the second sample, a prose text describing everyday life, the scope of collocation of golden is relatively narrow and is limited to words of the type hair, map, head, etc. The typical attributes of these nouns are the adjectives dark, black, yellow, golden.

All three pictures of the semantic field of colour discovered by DSA seem to correspond to our intuitive ideas of the structure of this field in English. And at the same time each of them is specific and is linked to the subject of the analyzed text.

A deeper analysis of this problem may be found in (47). Here it is necessary to repeat the main conclusion of this analysis: DSA gives results similar to results obtained by intuitive groupings of words, but often discovers important semantic links rarely thought of during logico-intuitive analysis.

An important problem is the interpretation of semantic groupings discovered by DSA. This area still remains a field of impressionistic exercise as it is extremely difficult to disentangle and interpret multiple multidirectional connections, cycles impasses and loops in the labyrinths of words.

The representation of semantic maps is still limited by the necessity to draw on paper unidimensional semantic maps which do not optically reveal all the complexities of interrelations among words. Semantic maps are sometimes called lexicographs (36,37) and there have been recent attempts at representing them as three dimensional structures.

In working with the complicated circuitry of semantic units it is necessary to have criteria for the delimination of smaller groups of units that correspond to semantic fields. Usually the following procedure is used: words are added to one another in such a way that the addition of every new word should not increase the sum of outward links of the group of words formed.

DISTRIBUTIVE-STATISTICAL ANALYSIS IN STYLISTICS

The study of distributive-statistical characteristics of text elements makes it possible to solve different tasks important for linguistic and literary research without recourse to semantic criteria. In this way semantic and stylistic connections of text elements may be discovered, various classifications of these elements may be built and connections established not only between text elements but also between the texts themselves.

Data relevant for stylistic research are detected in all the intervals of text but particularly valuable for automatic stylistic classification and grouping of texts are results obtained at the large and maximum intervals.

The large interval provides specific information connected with the subject and the expressive tonality of texts that cannot be obtained either on the minimal or on the maximum intervals. In the sample IV consisting of 1 million words of text of English writings of the Elizabethan and early Stuart periods (1560-1640) five relatively isolated groups of words were discovered (35).

The first group consists of 42 words (with the combined frequency exceeding 5000): great, other, same, whole, small, said, certain, chief, English, diverse, ancient, present, former. 10 of these words are diagnostic words of the expository factual style of speech.

The second group consists of 7 words (with the combined frequency exceeding 2000): amorous, fair, sweet, gentle, goodly, lovely, wondrous. The common semantic feature of these words is their connection with the 'lyrical' component of sense.

The third group includes 24 words (with the combined frequency of 1500): golden, happy, bright, heavenly, blessed, sacred, pure, free, famous, learned, perfect, silver, virtuous, chaste, red, earthly, celestial, pleasing, warm, etc. This group represents the 'elevated' component of sense. The fourth group embraces 40 words (with the combined frequency of 3000): noble, warlike, bloody, cruel, proud, royal, brave, mightly, princely, gracious, deadly, worthy, base, glorious, valiant, dreadful, fatal, etc. This group of words corresponds to the 'civil' or 'military - feudal' component of sense.

The fifth group has 51 words (with the combined frequency of 2600): poor, foul, sad, black, cruel, wicked, heavy, vain, mortal, wretched, low, bitter, weary, woeful, etc. The semantic component represented here is connected with melancholy. It may be called an 'elegiac' component.

The proportion of the combined frequency of words of each group in the combined general frequency of all the groups may be considered as a measure of this component for each individual text. This provides a possibility for classification of texts according to the interrelation of separate components. A group of texts where the expository component is predominant includes words by Francis Bacon, The King James translation of the New Testament, constitutional documents.

The predominance of the 'lyrical' component is noticeable in John Lyly's 'Love's Metamorphosis', 'Endymion', in Shakespeare's 'Sonnets', 'A Midsummer Night's Dream', and 'As You Like It', and in Thomas Dekker's 'The Shoemaker's Holiday'.

The 'elevated' component is characteristic for such works as 'The Hymns' by Edmund Spenser, 'The Holy Sonnets' by John Donne,

'Measure for Measure' and 'King Lear' by Shakespeare, and 'Antonion and Mellida' by John Marston.

The 'civil' component has a high frequency in 'Poly-Olbion' by Michael Drayton, 'Catilina' by Ben Johnson, 'Philaster, or Love Lies A-Bleeding' by Francis Beaumont and John Fletcher, 'Richard III', 'Henry V', 'Anthony and Cleopatra', 'Julius Caesar' and 'Macbeth'.

For some texts a combination of several components is characteristic. For example in Shakespeare's 'Venus and Adonis' and J. Fletcher's 'The Faithful Shepherdess' all the 'emotional' components (the 'lyrical', 'elevated' and 'elegiac') are well represented.

A comparison of the coefficients of correlation among different components reveals that the 'expository' component is negatively correlated with all the rest, while the latter ones are practically independent of each other with correlation coefficients being positive but negligible. The 'expository' component has a thematic character while other components are indicators of stylistic tonality.

In the large interval words are interrelated in a different way than in the small interval. For example, in the small interval the adjectives of colour are interconnected forming the semantic field of colour. In the large interval these words are connected with components, golden and red - with the 'elevated' component, black - with the 'melancholy' one. The words fair and foul, in the small interval representing a connected pair, fall into separate groups in the large interval: 'lyrical' and 'elegiac'. In the large interval considerably less antonymic pairs of words are discovered than in the small interval: 4 % of all links versus 13 % of all links.

The maximum interval that embraces entire long texts still remains largely unexplored. A few experiments conducted in this interval however brought extremely interesting results. In the sample IV semantic information of two kinds was discovered: for belles lettres it has a mixed subject-oriented and stylistic character and for texts of other genres - a pure thematic character. 233 texts of this sample are interconnected by 1124 links (an average of 9,6 links per text). For example, Shakespeare's 'Romeo and Juliet' ist strongly connected with 'The Comedy of Errors', 'Henry IV', 'The Taming of the Shrew', 'Richard II', and 'Venus and Adonis' by the same author as well as with 'The Sad Shepherd' by Ben Jonson.

On the bais of these links three major groups of texts may be isolated: A) poetic texts and some dramatic texts (by Th. Kyd, Chr. Marlowe, W. Shakespeare and J. Marston); B) dramatic texts, mainly comedies and J. Lyly's 'Euphues: The Anatomy of Wit' and 'Eupheus and His England'; C) factual expository (words by F. Bacon, 6 descriptions of voyages, constitutional documents, etc.).

On the basis of this tentative stylistic classification of texts the delimination of the most characteristic diagnostic features of each style can be made. Among adjectives the most characteristic positive diagnostic features of the group of texts A are the words bloody, cruel, deadly, fair, foul, goodly, heavy, lovely, mighty, proud, warlike, weary, of the group B - dear, excellent, fine, gentle, good, honest, ill, mad, noble, old, own, poor, pretty, sweet, young, and of the group C -

ancient, certain, diverse, divine, excellent, great, other, particular, present, said, same, small, whole. Negative diagnostic features among adjectives for the group of texts A are the words: good, great, honest, old, other, own, said, whole, for the group B - certain, English, great, high, other, said, same, small, and for the group C - dear, fair, gentle, poor, sweet.

The set of positive and negative features for texts of the group A is called index A, for B - index B and for C - index C. A very conventional semantic interpretation of these indices may look as follows: A - index of 'poetics', B - index of 'dialogue', C - index of 'expository' style.

The study of these indices may cast some light on the development of the system of styles in written English of the period under consideration. This period may be divided into three parts: 1) before 1590; 2) 1590-1610; 3) after 1610. For each of these periods 24 representative texts are chosen (8 texts for each of the groups A, B and C) and the rank correlation coefficient between indices A, B and C is calculated according to the formula:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n} \qquad (X)$$

where n is the number of pairs of texts compared, d_i - the differences in ranks of texts in the groups A, B and C (the texts are ranked in each group on the basis of the corresponding indices A, B and C). The coefficients obtained are shown on Table 2.

Table 2. Coefficients of rank correlation between different stylistic indices for three parts of the Elisabethan and early Stuart period of English literature.

	before 1590	1590 - 1610	after 1610
A - B	+ 0,46	+ 0,16	+ 0,20
A - C	- 0,89	- 0,67	- 0,47
B - C	- 0,60	- 0,59	- 0,75

These coefficients reveal the development of the style of the dialogue and in connection with this changes in the hierarchy of opposition among styles. In the period before 1590 the main opposition was that between A and C, with the upcoming B being still close to A, and having very few features common with C. In the final period after 1610 the interrelation of styles is quite different. Style B is now well separated from style A and is maximally distant from style C (a detailed analysis of the corresponding findings is given in (48)).

DSA in the maximum interval shows that statistical indices of occurrence of words and other language elements vary more in texts of different styles than in texts of different authors belonging to the same styles. That means that statistical studies of the individual style of an author, the establishment of the authorship, etc. should be preceded by the quantitative study of style differences.

It would be interesting to compare the five components of text classification: 1) 'expository', 2) 'lyrical', 3) 'elevated', 4) 'civil', 5) 'elegiac' discovered in the large interval with the three stylistic indieces obtained in the maximum inter-

val. The coefficients of association between the components and the indices are shown in Table 3.

The coefficient of association is defined by

$$Q = \frac{N\delta}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} \qquad (XI)$$

where $\delta = (AB) - \frac{(A)(B)}{N}$.

The capitals A, B in this formula denote corresponding attributes (each of the five components and each of the three stylistic indices taken pairwise for the evaluation of their degree of association); (A), (B) - numerical values of A and B, (AB) - numerical value of AB, i.e. the number of texts possessing at the same time both the characteristics of the component A and the stylistic index B; α, β - the absence of the attributes A, B; (α) (β) - corresponding numerical values; (αB), $(A\beta)$ and $(\alpha\beta)$ - numerical values for these combinations of attributes; N - the number of texts under consideration; (the formula and its derivation are given in (49, p. 30).

A very important conclusion can be drawn from the figures in Table 3: the 'expository' component coincides with the functional style C (the 'expository' style). This component is a style differentiating means. All the rest of the components are proven to be separate entities, though there exists a certain degree of correlation between style B ('dialogue') and the components 2 ('lyrical') and 5 ('elegiac') as well as between style A ('poetic') and the components 3 ('elevated') and 5 ('elegiac') (a detailed interpretation of these findings is given in (34)).

Table 3. Coefficients of association between stylistic components and indices of functional styles for the English writings of the Elisabethan and early Stuart period.

Components	A	B	C
1	- 0,87	- 0,63	+ 0,90
2	+ 0,27	+ 0,61	- 0,73
3	+ 0,48	+ 0,01	- 0,44
4	+ 0,59	+ 0,30	- 0,45
5	+ 0,44	+ 0,54	- 0,51

Analysis of the results of DSA obtained in various intervals of text shows the qualitative differences in the kind of information obtained in every interval. For example, in the minimal interval the adjective huge entered the field of 'size' being connected with the adjectives great and small. In the large interval it entered the component 'elegiac', while great and small entered the component 'expository'. The semantic field of 'pride' discovered in the small interval fell apart in the large interval: the adjective proud entered the 'expository' component and the adjective humble - the 'elegiac' component. The semantic field of 'moral and esthetical judgement' was split between three components at the large interval: fair belongs to the 'lyrical' component, foul - to the 'elegiac' one and base to the 'expository' one (34).

Experiments with the use of DSA in various intervals of text has produced significant results. Macrosemantic maps for texts of various authors were created and interesting tendencies of

changes in semantic connections with differences in time and literary periods were discovered. For example, by comparing two macrosemantic maps of English adjectives (one for English poetry of the 16^{th} to 18^{th} centuries, and another of the poetry of English Romantics) it was established that in the poetry of the Romantics many antonymic connections are weakened or even disappear. This is a sign of a considerable semantic-stylistic change: a decrease of semantic contrast in a line of the poetic text. If at the centre of the poetry of the 16^{th} to 18^{th} centuries was man and the traditional symbolism of colour, in Romantic poetry the semantic field of colour mingles with the semantic field of light. This is apparently due to the turning of the Romantics towards descriptions of nature as a gateway to mystical awareness; for as William James pointed out in 'Varieties of Religious Experience', a sudden sense of light is common to many who enter a state of mystic experience.

The examples given above leave no doubt as to the extremely important potential value of DSA for stylistic studies. It provides the style researcher with a powerful discovery procedure revealing new aspects of stylistic features of text and making it possible to replace the vague notions of functional stylistics with a more exact set of notions of stylistic components based on statistical characteristics of texts.DSA is a useful tool for analyzing the style of individual authors and the changes in the manner of writing and the ideas of an author during his creative lifetime. It is especially valuable in the analysis of literature of the past. Its most obvious advantage is that it permits the thorough searching of much larger quantities of text that can be retained in the mind of at least most human researchers. It facilitates the weighing of discordant and conflicting evidence and reveals the complex nature of the phenomenon under investigation showing the multiple links of a notion with other notions in a semantic network. It is also able to indicate the extent to which particular elements within the complex are due to one source or another (39).

Long before the introduction of DSA, Andrei BELYJ in his famous book on Gogol'(50) explained the change in Gogol's Weltanschauung and hence built the periodization of his creative work on the basis of the changes of frequencies of colour words (the fall in the frequency of use of the designations of red and the rise of the designations for yellow)(50). DSA might have yielded much more evidence as it shows not only the changes in the frequencies of individual words but at the same time - changes in their semantic links with other words.

The immense potentials of DSA in this area were shown in A. McKINNON's works on S.A. Kierkegaard and G.W.F. Hegel. Having subjected texts written by these philosophers during various periods of their life to DSA, McKINNON revealed and interpreted changes in their philosophies during their lifetimes. One of McKINNON's findings was that contrary to the standard 'melancholy Dane' tradition in Kierkegaard scholarship the frequency of eight standard images for hope, including the word for hope is almost double of that for various images of despair. This was the case except for a relatively brief period around the year 1849 when the situation is almost exactly the reverse. The same study attempted to determine with the help of DSA whether Kierkegaard had ever resolved what was once a conflict in his mind between the conception of Christ as Model and Christ als Redeemer. DSA

tended to support a conclusion that by around 1850 Kierkegaard had come to see this distinction itself as untenable (51).

The number of serious studies in the application of the distributive-statistical method in stylistic research is still limited. The global description of the vocabulary of language carried out with the help of this method necessitates the development of new theoretical ideas on the ontology of semantic space in language. The interrelations of semantic units appear to be more varied and multidirectional than was hitherto assumed.

The interpretation of semantic networks is an extremely complicated task. A useful notion that may help to find explanations for the features of individual semantic networks may be the notion of semantic centres. Semantic centres are clusters of keywords most characteristic for an analyzed text. Certain parts of a semantic network produced as a result of DSA may be isolated as semantic centres. Semantic networks are differentiated from each other on the basis of their semantic centres and of the configuration and intensity of links within semantic centres and among them. The notion of semantic centres helps to interpret semantic networks in terms of generalized semantic groups and not just concrete keywords.

It is in the evaluation of semantic networks for individual authors and literary movements that the notion of semantic centres of keywords is highly applicable. For example, for the poetry of the French Romantics (V. Hugo, A. de Musset, Th. Gauthier, M. Desbordes-Valmore, G. de Nerval, A. de Vigny) three semantic centres are characteristic: 1) motion; 2) motivation of motion (attempt, aim, possibility, cause, discovery, etc.); 3) faith and destiny. In contrast to the poetry of the Romantics the poetry of the Parnasian movement (Leconte de Lisle, J.M. de Heredia, S. Prudhomme, C. Mendès, F. Coppée, L. Dierx) concentrates on other semantic centres; 1) war; 2) forces; 3) the historical past. In the verses of French Symbolists (S. Mallarmé, A. Rimbaud, P. Verlaine, Ch. Baudelaire, P. Valery, P. Claudel) there is a vast variety of semantic centres reflecting both the wide scope of symbolist vocabulary and the main feature of this movement - the expression of various sensations: 1) water; 2) planets and stars; 3) sensations; 4) thought; 5) indefiniteness; 6) the body; 7) cleanliness; 8) wine; 9) holiday; 10) size (52). The semantic worlds of individual authors reflect similar features. In the poetry of G. de Nerval the main feature of Romantic verse-contrast in description of events, the celebration of their darker, more somber aspect - is reflected in a highly specific manner. For his works the opposition of heat and cold, light and dark, of the cycle of the year's seasons and of repetition is characteristic. For A. Rimbaud the wide use of prosaic words in poetry as a protest against the traditional poetic vocabulary is highly indicative. His verses are unmatched both in the richness and the frequency of the prosaic stratum of the vocabulary (52).

It appears that by using comparatively simple distributive-statistical techniques, valuable materials for stylistic studies may be produced. Such techniques provide an objective picture of the distribution of semantic units in texts and lay a foundation for an objective approach to the study of stylistic phenomena. Statistical discovery procedures of this type require as their basis exhaustive and exact data on the frequency and distribution of text elements. This necessitates a further effort in

compilation of concordances to texts of the most important authors of different literary periods. For the material contained in a concordance is a sine qua non for further distributive-statistical evaluation of the statistical significance of individual collocations.

CONCLUSION

At first glance the results obtained by DSA cannot be equal in quality with semantic descriptions and thesauri built by experts. One cannot expect that computers could immediately provide anything that could replace the work and intuition of a scholar (whose intuition is needed all the same for the selection of units for analysis and interpretation of the results of DSA). But hopefully DSA may reveal groups of words which remained previously unnoticed, but which are potentially useful.

Associative maps of text elements provided by DSA have some characteristics that make them absolutely indispensable for the kind of objective text study that modern semantics and stylistics purport to strive for:

1) they are obtained directly from the text and are not creations of a subjective judgement; they include only elements found in the text and no other elements that might be taken from other sources;
2) they are generated by an algorithmic procedure than can be repeated by a computer on any new corpus of texts;
3) they reveal various factes of the meaning of a word in a given body of texts;
4) they detect not only semantic links between text elements but statistical links as well; the latter are more

varied and cannot be detected by any other method.

Difficulties that stand in the way of a wider application of the DSA are threefold:

1) the comparative novelty of the method and the need for further development and testing of its techniques;
2) the high cost of input and computer processing of large text corpora;
3) the psychological difficulty for users involved in the use of any new technique.

In view of the complexity of the field and the need for an immense outlay of human and financial resources involved in the application of DSA to analysis of large samples of text, closer international cooperation in the field remains a priority. Work on DSA is still scattered throughout many disciplines (artificial intelligence, computational linguistics, psychology, philosophy, information retrieval, literary studies) and there is little knowlegde of work done in various countries and a very negligible exchange of ideas.

We hope that this paper will draw more attention to the scope and potential of the field.

REFERENCES

1. ŠČERBA, L.V., O trojakom aspekte jazykovyx javlenij i ob eksperimente v jazykoznanii. Izvestija AN SSSR, Otdel obščest-vennyx nauk 1931.
2. HALLIDAY, M.A.K., Categories of the theory of grammar, Word 17, 1961, 241-292.

3. VAN DIJK, T.A., Some aspects of text grammars. The Hague, Mouton 1972.
4. RIEGER, B., Theorie der unscharfen Mengen und empirische Textanalyse. Aachen, MESY 1976.
5. MEUNIER, J.G., Analyse contextuelle et analyse conceptuelle. Projet ALTPO, Doc. JGM 3, Université de Québec à Montréal, Fevrier 1976.
6. BALDWYN, A.L., Personal structure analysis: a statistical method for investigating the single personality. Journal of Abnormal and Social Psychology 37, 1942, 163-183.
7. TAUBE, M. et al., Studies in coordinate indexing. Washington 1965, vol. 3.
8. DOYLE, L.B., Semantic road maps for literary searchers. Rept. No SP - 199, System Development Corporation, Santa Monica, California, January 23, 1961.
9. STILES, H.E., The association factor in information retrieval. Journal of the ACM 8, 1961, 271-279.
10. NEEDHAM, R.M., Research on information retrieval classification and grouping, 1957-1961. Ph. D. Thesis, Cambridge, England 1961.
11. TANIMOTO, T.T., An elementary mathematical theory of classification and prediction. IBM Corporation, New York, November 1958.
12. BORKO, H., The construction of an empirically based mathematically derived classification system. Rept. No. SP - 585, System Development Corporation, Santa Monica, California, October 26, 1961.
13. BAKER, F.B., Latent class analysis as an association model for information retrieval. In: STEVENS, M.E. (ed.), Statistical

association methods for mechanized documentation. NBS Miscellaneous Publication 269, Washington, D.C. 1965.

14. JONES, P.E., GUILLIANO, V.E., CURTICE, R.M., Papers on automatic language processing, vols. 1-3. Arthur D. Little Inc., Cambridge, Mass. 1967.

15. JONES, P.E. et al., Application of statistical association techniques to the NASA document collection. Arthur D. Little Inc., Cambridge, Mass. 1968.

16. CAGAN, C., A highly associative document retrieval system. Journal of the ASIS 21, 1970, 330-337.

17. VASWANI, P.K.T., CAMERON, J.B., The National Physical Laboratory experiments in statistical word associations and their use in document indexing and retrieval. Teddington, National Physical Laboratory, Middlesex, 1970.

18. SPARCK JONES, K., JACKSON, D.M., Current approaches to classification and clump finding at the Cambridge Language Research Unit. The Computer Journal 10, 1967, 29-37.

19. SPARCK JONES, K., KAY, M., Linguistics and information science. N.Y. and London 1973.

20. LEWIS, P.A.W., BAXENDALE, P.B., BENNETT, J.L., Statistical discrimination of the synonymy/antonymy relationship between words. Journal of the ACM 14, 1967, 20-44.

21. HARPER, K.E., Measurement of similarity between nouns. Rm - 4532 - PR, Rand Corporation, Santa Monica, California 1965.

22. NEEDHAM, R.M., Automatic classification in linguistics. The Statistician 17, 1967, 45-54.

23. SPARCK JONES, K., A small semantic classification experiment using co-occurrence data. Cambridge Language Research Unit 1967.

24. HIRSHMANN, L., GRISHMAN, R., SAGER, N., Grammatically-based automatic word class formation. Information Processing and Management 11, 1975, 39-57.

25. SAGER, N., Computerized discovery of semantic word classes. In: GRISHMAN, R. (ed.), Directions in artificial intelligence; natural language processing. N.Y., Courant Institute of Mathematical Sciences, New York University 1975, 27-48.

26. ANDREEV, N.D., Modelirovanie jazyka na baze jego statističeskoj i teoretiko-množestvennoj struktury. In: Tezisy soveščanija po matematičeskoj lingvistike. Leningrad 1959, 16-22.

27. ŠAJKEVIČ, A.J., Raspredelenie slov v tekste i vydelenie semantičeskix polej jazyka. In: Inostrannye jazyki v vysšej škole 1963, 14-26.

28. PRITSKER, A.J., Distributivno-statističeskij analiz semantičeskogo polja. In: Problemy formalizacii semantiki jazyka. Tezisy dokladov. Moscow 1964, 130-132.

29. MOSKOVICH, W., Opyt kvantitativnoj tipologii semantičeskogo polja. Voprosy jazykoznanija 4, 1965, 80-91.

30. MOSKOVICH, W., Statistika i semantika. Moscow, Nauka 1969.

31. IVANOVA, N.S., Ustanovlenie smyslovyx svjazej meždu slovami na osnove statističeskoj metodiki. In: MOSKOVICH, W. (ed.), Voprosy lingvostatistiki i automatizacii lingvističeskix rabot, vol. 1. Moscow, Patent 1967, 17-20.

32. IVANOVA, N.S., MOSKOVICH, W., Automatic compiling of thesauri on the basis of statistical data. In: Information retrieval among examining patent offices. VII annual meeting. BIRPI, Geneva 1968.

33. IVANOVA, N.S., ŠAJKEVIČ, A.J., Distributivno-statističeskoje

opisanie amerikanskix patentnyx tekstov. In: MOSKOVICH, W. (ed.), Voprosy lingvostatistiki i automatizacii lingvističeskix rabt, vol. 4. Moscow, Patent 1970, 77-221.

34. ŠAJKEVIČ, A.J., Korreljacionnyj analiz v lingvostatistike i ponjatie intervala teksta. In: MOSKOVICH, W. (ed.), Voprosy lingvostatistiki i automatizacii lingvističeskix rabot, vol. 2. Moscow, Patent 1970, 254-274.

35. ŠAJKEVIČ, A.J., Distributivno-statističeskij analiz v semantike. In: Principy i metody semantičeskix issledovanij. Moscow, Nauka 1976, 353-378.

36. GEFFROY, A. et al., Lexicometric analysis of co-occurrences. In: AITKEN, A.J. et al. (eds.), The computer and literary studies. Edinburgh 1972, 113-133.

37. GEFFROY, A., Analyse lexicométrique des co-occurrences et formalisation. In: Les applications de l'informatique aux textes philosophiques. CNRS 1970, 8-83.

38. BERRY-ROGGHE, G.L.M., The computation of collocations and their relevance in lexical studies. In: AITKEN, A.J. et al. (eds.), The computer and literary studies. Edinburgh 1972, 103-112.

39. McKINNON, A., Two techniques for identifying word association. CIRPHO 2, Fall, 1974, 49-64.

40. MEUNIER, J.G., L'analyse lexicale des texts par ordinateurs. Paradigmes et definitions. Projet ALTPO, Doc. JGM 4, Université de Québec à Montréal, 1976.

41. JONES, P.E., CURTICE, R.M., A framework for comparing term association measures. American Documentation 18, 1967, 153-161.

42. ANDREEV, N.D., Statistiko-kombinatornye metody v teoretičeskom i prikladnom jazykoznanii. Leningrad, Nauka 1967.

43. ANDREEVA, L.D., Statistiko-kombinatornye tipy slovoizmenenija i razrjady slov v russkoj morfologii. Leningrad, Nauka 1969.

44. ŠAJKEVIČ, A.J., Vydelenie klassov slov i paradigm posredstvom distributivno-statističeskogo metoda. Prikladnaja lingvistika 18, 1976, 96-134.

45. MOSKOVICH, W., Quantitative linguistics. In: WALKER, D., KARLGREN, M., KAY, M. (eds.), Natural language in information science. Stockholm, Scriptor 1977, 57-74.

46. SKOROXOD'KO, E.F., Pro zastosuvannja elektronnyx cyfrovyx mašyn v lingvistyčnyx doslidžennjax. In: Ukrains'ka respublikans'ka naukova konferencija z pytan' metodologii movoznavstva. Kiev 1964, 49-50.

47. MOSKOVICH, W., Distributivno-statističeskij metod postroenija tezaurusov: sovremennoe sostojanie i perspektivy. Moscow 1971.

48. ŠAJKEVIČ, A.J., Opyt statističeskogo vydelenija funkcional'nyx stilej. Voprosy jazykoznanija 1968, 64-76.

49. YULE, C.U., KENDALL, M.G., An introduction to the theory of statistics. N.Y. 1950

50. BELYJ, A., Masterstvo Gogolja. Moscow-Leningrad 1934.

51. McKINNON, A., Computer-assisted instruction with a difference. In: Programmed learning. London 1972, 18-21.

52. ABRAMOVA, N.I. Poetičeskaja leksika francuzskogo jazyka. Moscow 1974.

Überlegungen zur Interpretation des Zipfschen Gesetzes am Beispiel der frühen Kindersprache

Wolfgang Marx und Elisabeth Schürer-Necker,
München

ZUSAMMENFASSUNG

Das ZIPFsche Gesetz läßt prinzipiell zwei verschiedene Interpretationsstrategien zu:

1) Die beobachtete Gesetzmäßigkeit ist das Ergebnis eines Optimierungsprozesses und damit einer Entwicklung, die gesteuert wird durch ein ökonomisches Prinzip.
2) Die beobachtete Gesetzmäßigkeit ist die notwendige Konsequenz der Wahrscheinlichkeitsgesetze.

Die vorliegende Arbeit soll eine Entscheidung zwischen diesen beiden Erklärungs-Modellen an Hand von theoretischen Erwägungen und empirischen Daten herbeiführen. An Hand der empirischen Daten, die aus dem Bereich der frühen Kindersprache gewonnen worden sind, ergab sich eine eindeutige Verwerfung des statistischen Modells zugunsten des ökonomischen Modells.

1. FRAGESTELLUNG

Das ZIPFsche Gesetz beinhaltet zunächst nicht viel mehr als eine Aussage über die Verteilung von Worthäufigkeiten im System konkreter Sprachen. Ordnet man die einzelnen Wörter nach der Häufigkeit ihrer Verwendung (p_n) in eine Reihe mit den Rangplätzen n (wobei der kleinste Rang der größten Häufigkeit zugeordnet wird), so gilt die Beziehung

$$p_n \times n = c,$$

d.h. das Produkt von Freqeunz und Rangplatz eines Wortes ist jeweils eine konstante Größe.

Zur graphischen Veranschaulichung des Gesetzes trägt man die Häufigkeiten der einzelnen Wörter über ihren Rängen auf, wobei man üblicherweise das von ZIPF bevorzugte doppelt logarithmische Koordinatensystem verwendet. Man erhält auf diese Weise eine Gerade mit der Steigung von a = - 1.

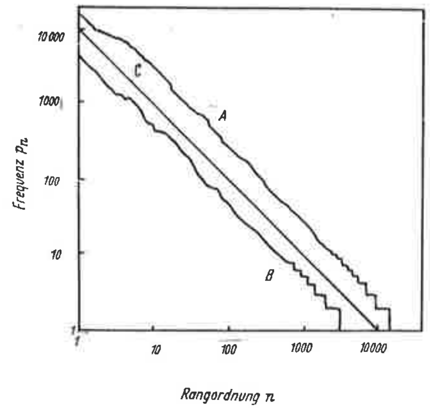

ABBILDUNG 1. Die Ranghäufigkeit der Wörter. A : James JOYCE "Ulysses"; B: amerikanisches Zeitungsenglisch; c: hypothetische Idealgerade. Aus C.K. ZIPF, 1949, S. 25

Eine Reihe von empirischen Zählungen, die an unterschiedlichem Material verschiedener Sprachen durchgeführt wurde, ergab, daß die von ZIPF gefundene Formulierung in der Regel eine brauchbare Annäherung an die vorgefundenen Häufigkeitsverhältnisse darstellt. Dieses Faktum wird heute auch von den Kritikern ZIPFs nicht bestritten, ungeachtet der Tatsache, daß es im Detail Änderungsvorschläge gibt. Strittig hingegen ist die Interpretation seiner Befunde. In diesem Zusammenhang bieten sich grundsätzlich zwei verschiedene Interpretationsstrategien an:

1) Die beobachteten Regelmäßigkeiten sind das Ergebnis eines Optimierungsprozesses und damit einer Entwicklung, die gesteuert wird durch ein ökonomisches Prinzip.
2) Die beobachteten Regelmäßigkeiten sind die notwendige Konsequenz der Wahrscheinlichkeitsgesetze.

ZIPF selber argumentierte im Sinne der ersten Hypothese. Seiner Ansicht nach wäre das Sprechen vom Standpunkt des Sprechers aus dann am einfachsten, wenn die Sprache nur aus einem Wort bestünde. Aus der Sicht des Hörer wäre dagegen die Sprache dann am rationellsten, wenn für jede zu unterscheidende Bedeutung auch ein eigenes Wort vorhanden wäre. Es stehen sich hier also zwei gegenläufige Tendenzen gegenüber: die Tendenz nach "unification", die in die Richtung zielt, die Zahl der verschiedenen Wörter auf 1 zu vermindern, während die Häufigkeit dieses Wortes auf 100 % anwächst. Umgekehrt zielt die Tendenz nach "diversification" in die Richtung, die Zahl der verschiedenen Wörter zu erhöhen, während die durchschnittliche Auftretenshäufigkeit gegen 1 absinkt. Die gefundene Beziehung drückt nun das Gleichgewicht der beiden entgegengesetzten Kräfte aus. Dieses Gleichgewicht ist das Ergebnis eines Optimierungsprozesses, der nach dem Prinzip des geringsten Aufwandes ("least effort") erfolgt.

MANDELBROT (zitiert nach CHERRY 1963: 145f.), der die Sprache mehr aus informationstheoretischer Sicht betrachtet, kommt aufgrund rein mathematischer Überlegungen zu ähnlichen Schlüssen. Er geht dabei vom Begriff der Kosten aus. Alle Elemente verursachen dem Sender einen gewissen Aufwand. Es läßt sich nun zeigen, daß man den Zeichensequenzen ("Wörtern") eines beliebigen

Sprach-Systems Auftretenswahrscheinlichkeiten so zuordnen kann, daß die Gesamtkosten im Mittel ein Minimum annehmen, wobei ihr mittlerer Informationsgehalt invariant bleibt. Dabei zeigte sich, daß die sich ergebende Beziehung zwischen den Auftretenswahrscheinlichkeiten und den Rängen der Wörter dem ZIPFschen Gesetz entspricht. Dieses beschreibt also auch in informationstheoretischer Sicht ein ökonomisches Optimum. (Eine analoge Überlegung findet sich auch bei HÖRMANN 1970: 92.)

Andere Wissenschaftler, z.B. GUIRAUD (1963) und MILLER (1965), haben im Sinne der 2. Hypothese argumentiert. Es läßt sich zeigen, daß bei einer rein zufälligen Verteilung von Leerzeichen (bzw. "Wortbegrenzungsmarken") über einen aus zufällig gezogenen Zeichenfolgen ("Wörter") bestehenden Text Verteilungen von Worthäufigkeiten entstehen, die der empirisch gefundenen ZIPF-Verteilung ähnlich sehen. Alle diese Verteilungen lassen sich durch folgende allgemeine Gleichung darstellen:

$$\lg p_n = b - a \lg n \qquad (2)$$

mit a und b als Konstanten und $a \approx 1$.

Die für das ZIPFsche Gesetz relevante Größe in dieser Gleichung ist die Konstante a, die den Anstieg der Geraden bestimmt. Der Ausprägungsgrad dieser Konstanten wiederum ist bestimmt durch die Größe der Parameter p_x für die Auftretenswahrscheinlichkeiten des Leerzeichens und der Zeichen ("Buchstaben") des Systems. Dabei gilt folgende allgemeine Beziehung:

Je größer der Wert p_x für die Auftretenswahrscheinlichkeit des Leerzeichens, desto steiler verläuft die Gerade. Beim Grenzwert $p_x = 1.00$ ergibt sich eine Senkrechte. Umgekehrt wird der Verlauf der Geraden flacher, je kleiner p_x für das Leerzeichen wird. Etwa von der Größenordnung $p_x = \frac{1}{7}$ an ist die Annäherung an den anderen Extremverlauf, nämlich eine Gerade mit dem Anstieg von a = -1, praktisch erreicht.

Nun liegen tatsächlich in den europäischen Sprachen die Wahrscheinlichkeiten für die Wortzwischenräume in der Größenordnung, die Verteilungen von Worthäufigkeiten erwarten lassen, die sich der ZIPFschen Gerade stark annähern. Für die deutsche Sprache beispielsweise ist dieser Wert gut $\frac{1}{7}$ (.152). Was liegt also näher, als anzunehmen, daß ein rein statistisches Modell zur Erklärung der ZIPFschen Befunde ausreicht?

Modellrechnungen auf der Basis dieser Überlegungen führen in der Tat zu den erwarteten Ergebnissen, wobei allerdings anzumerken ist, daß diese Berechnungen in der Regel von dem für sie günstigsten Fall der Gleichwahrscheinlichkeit der Zeichen ausgehen. (Bei einer gleichbleibenden Anzahl von Elementen und konstanter Wahrscheinlichkeit für das Leerzeichen wird der Verlauf der Geraden nämlich dann am flachsten und das heißt: nähert sich der idealen Geraden mit a = -1 am meisten an, wenn alle Elemente gleich wahrscheinlich sind.)

Eine empirische Entscheidung zwischen den beiden Interpretations-Ansätzen herbeizuführen, setzt voraus, daß es gelingt, unter bestimmten Bedingungen aufgrund der verschiedenen Modelle zu verschiedenen Vorhersagen zu kommen und zu prüfen, welche Vorhersage sich als zutreffender erweist. Wir haben für eine solche Überprüfung das Beispiel der frühen Kindersprache gewählt, weil sich für diesen Bereich in der Tat aufgrund der unterschiedlichen Modelle unterschiedliche Erwartungen formulieren lassen.

Vorausgesetzt, das statistische Modell ist zutreffend, ist zu erwarten, daß sich das ZIPFsche Gesetz auch im Bereich der

Kindersprache nachweisen läßt. Wenn man davon ausgeht, daß der Parameter p_x für die Auftretenswahrscheinlichkeit des Wortzwischenraums wahrscheinlich größer ist als bei der Erwachsenensprache, schlichter formuliert, daß die durchschnittliche Wortlänge in der Kindersprache kleiner ist als bei der Erwachsenensprache, ist zu erwarten, daß die Gerade eher etwas steiler ausfällt als bei der Erwachsenensprache.

Dieselbe Erwartung müßte man übrigens auch von der ZIPFschen Theorie her formulieren, wenn man bedenkt, daß ein steilerer Verlauf als Ausdruck eines Überwiegens der Tendenz zur "unification" und damit einer gewissen "Egozentrizität" der Kindersprache gewertet werden kann. Informationstheoretisch gesehen würde das bedeuten, daß die Kindersprache redundanter ist als die Erwachsenensprache, was von HOFSTÄTTER (1957: 276) auch über die Frühphase der Phylogenese der Sprachentwicklung vermutet wird.

Im Rahmen unserer Arbeit möchten wir jedoch von einer anderen Hypothese ausgehen, und zwar aufgrund folgender Überlegungen. Informationstheoretisch gesehen sind Entwicklungsprozesse in der Regel durch Abnahme von Entropie gekennzeichnet. Wenn das - wie wir erwarten - auch für die Sprachentwicklung zutrifft, dann sollte die Kindersprache nicht redundanter, sondern umgekehrt gerade weniger redundant sein als die Erwachsenensprache. Das wiederum bedeutet, daß wir erwarten, daß die Kurven bei der Kindersprache systematisch nicht steiler, sondern flacher verlaufen als bei der Erwachsenensprache.

Zum Verständnis dieser Überlegungen möge man sich verdeutlichen, daß die Entropie bzw. der mittlere Informationsgehalt einer Sprache dann ein Maximum erreicht, wenn alle m Wörter die gleiche Auftretenswahrscheinlichkeit haben, nämlich p = 1/m. In diesem Falle würde die ZIPFsche Kurve genau waagerecht verlaufen. Ein Minimum würde die Entropie einer Sprache dann aufweisen, wenn nur 1 Wort die Auftretenswahrscheinlichkeit p = 100 hätte. In diesem Falle würde die ZIPFsche Kurve genau senkrecht verlaufen. Die empirisch ermittelte Steigung von a = -1 stellt - wie oben ausgeführt - nun das optimale Gleichgewicht zwischen vermittelter Information und verursachtem Aufwand dar.

Wenn nun in früheren Stadien der Sprachentwicklung - wie wir vermuten - mehr Entropie besteht als im Endstadium, so bedeutet das, daß im Verhältnis zur vermittelten Information der geleistete Aufwand in diesen früheren Stadien höher ist als später. Ein solches Zuviel an Aufwand (beispielsweise an Krafteinsatz) im Verhältnis zum Ergebnis beobachten wir auch bei anderen Entwicklungsprozessen, beispielsweise bei der Entwicklung von feinmotorischen Fertigkeiten wie dem Schreibenlernen. Hier denke man etwa an den extrem hohen Schreibdruck der Erstklässler.

Wenn wir diese Überlegungen auf unser ökonomisches Modell zurückbeziehen, kommen wir zu der Annahme, daß das ZIPFsche Gesetz das Ergebnis eines Optimierungsprozesses und damit einer Entwicklung ist. Wir vermuten, daß dieser ausgezeichnete Zustand nicht von allem Anfang an besteht, sondern erst am Ende des Prozesses erreicht wird. Daraus läßt sich die Erwartung ableiten, daß das ZIPFsche Gesetz in der frühen Kindersprache noch nicht gültig ist, sondern daß wir hier einen Zustand vorfinden, der von dem Endzustand noch deutlich abweicht.

Während nun das statistische Modell zu der Erwartung führt, daß der Verlauf der Geraden für die Kindersprache eher steiler sein sollte als bei der Erwachsenensprache, erwarten wir aufgrund

unserer Überlegungen genau das Gegenteil, nämlich daß der Verlauf der Geraden flacher wird als bei der Erwachsenensprache. Sollte sich unsere Erwartung bestätigen, wäre das statistische Modell eindeutig widerlegt, da - wie auch immer die Parameter p_x für die Auftretenswahrscheinlichkeiten des Zwischenraumes und der Zeichen gewählt werden - in keinem Falle der Wert von a größer als -1 werden kann.

2. DURCHFÜHRUNG DES EXPERIMENTS

Da das ZIPFsche Gesetz universell gültig sein soll, ist die einzige Bedingung, die wir beachten müssen, eine gewisse Mindestgröße der Sprach-Stichproben einzuhalten. Was die Vpn anbetrifft, so beschränken wir uns auf Kinder, die sich nach der klassischen Terminologie im Stadium des "Zweiwortsatzes" befinden. Die Zuordnung der Kinder zu dieser Kategorie erfolgte an Hand eines Vortestes, bei dem die MLU-Werte (mean length of utterance) der Kinder bestimmt wurden. Kinder, die ein MLU $\geq$ 1,2 und $\leq$ 2,5 Morphemen pro Äußerung aufwiesen, wurden in die Stichproben aufgenommen. Insgesamt wurden 14 Kinder untersucht, 8 Jungen und 6 Mädchen, deren Lebensalter zwischen 1,87 und 5,25 lag.

Wenn man die Faustregel anwendet, daß ein Kind als sprachlich retardiert gelten kann, wenn sein Alter größer ist als sein 1 1/2-faches MLU, kann man genau 7 der Kinder als sprachlich normal entwickelt einstufen und 7 der Kinder als sprachlich retardiert.

Mit jedem Kind wurden 3 Sitzungen durchgeführt. Die 1. Sitzung fand bald nach dem Vortest statt und dauerte 45 Minuten. Die 2. und 3. Sitzung wurden ca. 6 Wochen später aufgenommen und dauerten je 30 Minuten. Die Kinder befanden sich dabei mit ihrer Bezugsperson in einem Raum, wo sie an einem Tisch mit einer Reihe von Spielsachen spielen konnten. Alle Kinder hatten die gleichen Spielsachen zur Verfügung. Um das Sprachverhalten möglichst wenig zu beeinflussen, wußten weder die Kinder, noch die Bezugspersonen, daß es bei der Untersuchung um Sprache ging. Den Bezugspersonen wurde mitgeteilt, daß das Spielverhalten des Kindes beobachtet würde. Sie bekamen folgende Instruktion:

> ***"Spielen Sie mit dem Kind, was Sie wollen, aber nur mit den Sachen in der Kiste. Wir wollen nur beobachten, wie sich das Kind beim Spiel verhält."***

Alle Sitzungen wurden mit Videorecorder bzw. Kasettenrecorder aufgenommen. Die Bänder wurden anschließend von je 2 unabhängigen Auswertern abgehört und - möglichst lautgetreu - transkribiert. Für nicht verstandene Äußerungen wurde ein Querstrich geschrieben, Äußerungen, die unmittelbar wiederholt wurden, wurden mit "W" klassifiziert, Imitationen mit "I". Als Imitation galt, wenn das Kind eine Äußerung der Bezugsperson nachsprach und dabei die gleiche Reihenfolge der Wörter beibehielt, nichts hinzufügte oder veränderte. Die Imitation mußte aber nicht vollständig sein.

Nachdem die Sitzung von beiden Auswertern abgehört worden war, wurde eine Liste des gemeinsam Gehörten erstellt, wobei für nichtverstandene oder unterschiedlich verstandene Äußerungen Zeilen freigelassen wurden. Anschließend wurden die Bänder erneut von beiden Auswertern abgehört. Dabei wurde das gemeinsam Gehörte nochmals kontrolliert, und es wurde versucht, die leeren Stellen zu ergänzen. Danach wurden die beiden neuen Protokolle verglichen; gleich Verstandenes wurde hinzugefügt, unterschiedlich Verstandenes zählte als nicht verstanden. Diese Listen, die nur aus dem gemeinsam Verstandenen (durchschnittlich

78,7 % der Äußerungen der Kinder; Variationsweite 70,7 - 88,6 % pro Kind) bestanden, waren das endgültige Material der weiteren Auswertung.

Bei den Auszählungen der Listen der einzelnen Sitzungen gingen wir von ZIPFs Einheit aus, nämlich der Wortform. Dabei werden bekanntlich gebeugte Formen desselben Wortes als getrennte Ereignisse gewertet. Im Gegensatz zu ZIPF haben wir jedoch Homonyme als unterschiedliche Ereignisse betrachtet, da sie ja unterschiedliche Bedeutungen haben und sicher als eigene Wörter gelernt werden müssen.

Die Imitationen des Kindes wurden beim Auszählen nicht berücksichtigt, weil es uns sinnvoll erschien, nur das zu werten, was das Kind spontan äußerte. Wiederholungen, Ausrufe und paralinguistische Ausdrücke haben wir hingegen - entsprechend der Praxis ZIPFs - mitgewertet.

Von jeder Sitzung eines Kindes wurde dann eine graphische Darstellung der Verteilung der Worthäufigkeiten über den Rängen angefertigt. Dabei wurde entsprechend der ZIPFschen Methode ein doppelt logarithmisches Koordinatensystem verwendet. Als beste Anpassung einer Geraden an die empirischen Punkte wurde die Regressionsgerade berechnet. Ihre Steigung soll im Idealfall genau $a = -1$ sein.

Tatsächlich wird dieser Wert in den allermeisten Fällen nicht exakt erreicht, was bei empirischen Meßwerten auch nich weiter verwunderlich ist. Die Berechnung der Anstiegswerte für unsere 42 Geraden (14 Kinder x 3 Sitzungen) ergab eine Variationsweite von $a = -1{,}12$ bis $a = -0{,}652$ mit einem Mittelwert von $a = -0{,}879$. An Hand der Abbildungen 2 und 3 kann abgeschätzt werden, in welchem Ausmaß die empirischen Verläufe sich den theoretisch erwarteten Geraden anpassen. Abbildung 2 zeigt eine relativ gute, Abbildung 3 eine relativ schlechte Anpassung.

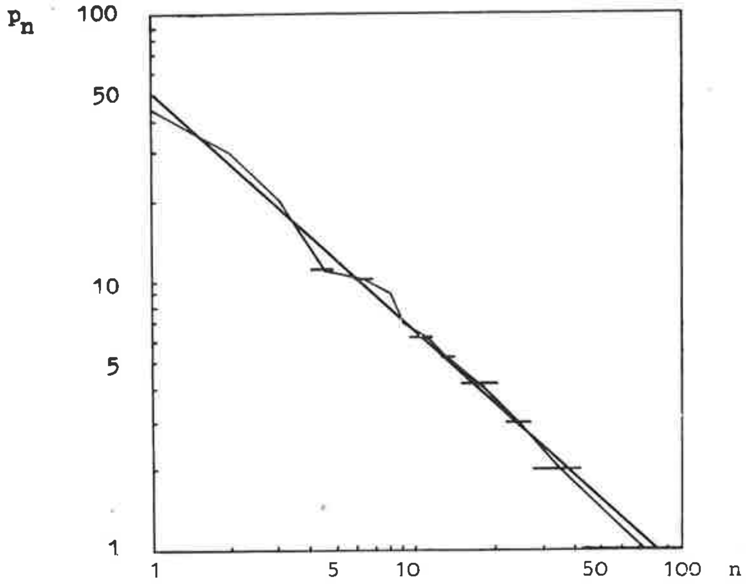

ABBILDUNG 2. Empirischer und theoretischer Verlauf für die Vp U.E. Beispiel für eine relativ gute Anpassung.

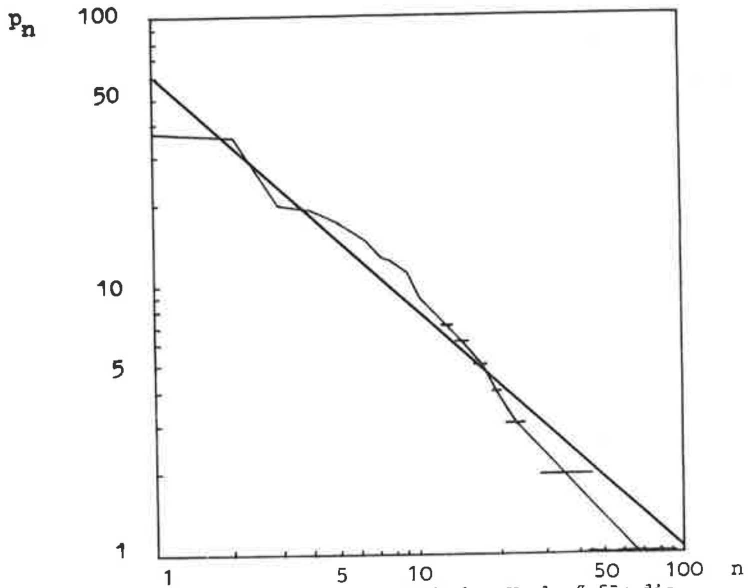

ABBILDUNG 3. Emprischer und theoretischer Verlauf für die Vp S.P. Beispiel für eine relativ schlechte Anpassung.

Ein sehr einfacher statistischer Test unserer Hypothese läßt sich nun aufgrund folgender Überlegungen ableiten: Wenn die beobachteten Abweichungen lediglich zufälliger Natur sind, dann sollten wir etwa gleich häufig Abweichungen nach oben wie nach unten feststellen, d.h. es sollten etwa gleich viele Geraden steiler bzw. flacher verlaufen als die Idealgerade.

Die Auszählung bei unseren 42 Geraden ergab jedoch ein Verhältnis von 36 (flacher) : 6 (steiler). Die Wahrscheinlichkeit dafür, ein solches Verhältnis durch reinen Zufall zu finden, läßt sich an Hand des Binomialtests auf

$$p = .000007 < p = .01$$

berechnen. (Bei zweiseitiger Fragestellung.)

Entsprechende Auszählungen für die Untergruppen Jungen und Mädchen, sowie normale und retardierte Sprachentwicklung ergaben in allen Fällen dasselbe Ergebnis: ein hochsignifikantes Überwiegen des flacheren Geradenverlaufs. In dieser Hinsicht gibt es zwischen den genannten Gruppen keine Unterschiede.

Bereits diese einfache statistische Analyse zeigt, daß bei unseren Daten eine statistisch signifikante systematische Abweichung von der Idealgerade vorliegt; und zwar geht diese Abweichung eindeutig in Richtung auf einen flacheren Verlauf. Das System der frühen Kindersprache weist also mehr Entropie auf als das System der Erwachsenensprache. Entwicklung bedeutet also auch im Bereich der Sprachentwicklung Abnahme von Entropie. Das wiederum bedeutet entsprechend unserer Überlegungen, daß ein ökonomisch betrachtet ungünstiges Verhältnis zwischen Aufwand und Ergebnis (hier: der vermittelten Information) besteht. Auch dieses ein typisches Charakteristikum von Entwicklungsprozessen im Frühstadium.

Dieses Ergebnis impliziert nun die Verwerfung einer rein statistischen Interpretation des ZIPFschen Gesetzes, da - wie wir bereits ausgeführt haben - bei reiner Zufallsbedingtheit der Anstieg der Geraden in keinem Falle flacher als a = -1 sein kann. Wir halten es daher für sinnvoller, die beobachtete Regelmäßigkeit als Ergebnis eines auf Optimierung hinzielenden Entwicklungsprozesses zu betrachten, wobei unseres Erachtens die rein auf informationstheoretischen Überlegungen basierende Analyse von MANDELBROT den bisher überzeugendsten Beitrag in dieser Richtung darstellt.

Die Deutung ZIPFs hingegen erscheint nach unseren Ergebnissen eher weniger überzeugend. Der gefundene flachere Verlauf der Geraden würde nämlich im Sinne der ZIPFschen Terminologie darauf schließen lassen, daß in der Kindersprache die Tendenz zur "diversification" überwiegt, was eine besondere Berücksichtigung der Bedürfnisse des Hörers bedeutet. Das erscheint nach allem, was wir über frühkindliches Sprachverhalten wissen, wenig wahrscheinlich.

Abschließend kann noch angemerkt werden, daß die Daten, die ZIPF selber zur frühen Kindersprache vorgelegt hat (ZIPF: 1949) ebenfalls einen Mittelwert für a aufweisen, der größer als -1 ist, wenn auch die Abweichung nicht so deutlich ist wie bei unseren Daten. Immerhin verlaufen jedoch auch rund 39 % seiner Geraden flacher als die Idealgerade. Auch diese Daten sprechen also gegen die Beibehaltung des statistischen Modells, das ja nur Variablität in Richtung auf einen steileren Verlauf zuläßt.

Literatur

CHERRY, C., Kommunikationsforschung - eine neue Wissenschaft. Frankfurt/Main 1963.

GUIRAUD, P., Structure des répertoires et repartition fréquentielle des éléments: La statistique du vocabulaire écrit. In: MOLES, A.A., VALLANCIEN, B. (Eds.), Communications et langages. 1963, 35-48.

HOFSTÄTTER, P.R., Psychologie. Fischer Lexikon Nr. 6, Frankfurt/Main 1957.

HÖRMANN, H., Psychologie der Sprache. Verb. Neudruck, Berlin, Heidelberg, New York 1970.

MILLER, G.A., Introduction. In: ZIPF 1965, IV-X.

SCHÜRER, E., Das Zipfsche Gesetz in der frühen Kindersprache. unveröffentl. Dipl.-Arbeit, München 1976.

ZIPF, G.K., Human behavior and the principle of least effort. Cambridge (Mass.) 1949.

ZIPF, G.K., The Psycho-Biology of Language. An introduction to dynamic philology. 2. Aufl. Cambridge (Mass.) 1965.

DISKUSSION

W. Matthäus:

Der Befund, daß die Worthäufigkeits-Verteilung bei Kindern systematisch vom ZIPFschen Gesetz abweicht, ist interessant und verlangt nach Interpretationen. M.E. ist aber der Stand der Einsicht noch nicht so weit entwickelt, daß man das Spektrum möglicher Interpretationen auf die Entscheidungsalternative "statistisch vs. ökonomisch" einengen könnte. Es gibt zunächst einige begriffliche Schwierigkeiten mit dieser Alternative.

1. Die Tatsache, daß der Variationsbereich des Anstiegskoeffizienten a erst bei +1 beginnt, ist nicht eine Konsequenz der Wahrscheinlichkeitsgesetze überhaupt, sondern eines speziell konstruierten Zufallsmodells.

Die Konstruktion dieses Modells müßte übrigens dem Leser zugänglich gemacht werden, wenn er die Argumentation beurteilen soll. Eine gute Quelle ist MILLER/CHOMSKY (1965: 456-464). Dort findet man auch die Formel, aus der die von MARX beschriebene Abhängigkeit des Koeffizienten a von 2 Faktoren hervorgeht, und zwar vollständig und verständlich abgeleitet. In dieses Modell nun gehen sehr spezielle Annahmen ein (Wörter aufgebaut nach demselben Mechanismus wie strings einer Zufallsziffern-tafel). Es gibt aber <u>andere stochastische</u> Modelle für Worthäufigkeits-Verteilungen, in denen der Anstiegskoeffizient zwischen 0 und 1 liegen kann, z.B. das von H.A. SIMON (1956) entworfene mit seinen Varianten, das nicht auf einer Optimierung, sondern eher auf einer Durchschnittsbildung beruht.

Also "a kleiner als 1" ist kein Beweis gegen einen Zufallsprozeß, sondern gegen den speziell von MILLER/CHOMSKY konzipierten.

2. Das ökonomische Modell ist im Grunde ja selbst ein Zufallsmodell. Die Ökonomie besteht doch darin, daß ein <u>statistisches</u> Maß des Aufwandes für die Information eines <u>Zufallsprozesses</u> (Redundanz) optimiert wird. Schließlich ist ja das MILLER/CHOMSKY-<u>Zufallsmodell</u> einem <u>optimalen</u> Kode (im Sinne von minimum redundancy) äquivalent (M/C: 462).

Die Gegenüberstellung von "statistisch" und "ökonomisch" ist auch deswegen nicht ganz überzeugend.

3. Die Beziehung "Aufwand - Entropie - Redundanz" bleibt etwas unklar. Wieso bedeutet steigende Redundanz (mit der Entwicklung) abnehmenden Aufwand bei der Informationsübertragung? Ist nicht Redundanz eigentlich der <u>überflüssige</u> Aufwand, den man sich bei der Kodierung zu minimieren bemüht (soweit "noise" das zuläßt) ?

4. Zum Ansatz insgesamt möchte ich bemerken:
Auf dem Abstraktionsniveau, auf das man sich begeben muß, um die Ranghäufigkeits-Funktion zu untersuchen, kann man identische "Phänotypen" von durchaus verschiedenen "Genotypen" (Mechanismen, Erzeugungsprozessen) erhalten. M/Cs Warnung vor spekulativen Kurzschlüssen (S. 462) erscheint mir nicht überholt: Zwar sind minimum-redundancy code und natürliche Sprache durchs ZIPFsche Gesetz zu beschreiben, doch ist die natürliche Sprache nach M/Cs Meinung weit von minimaler Redundanz entfernt.

Hieran möchte ich die Frage anschließen: Ist das <u>statistische</u> Informationsmaß für natürliche Sprachen so relevant, daß seine Optimierung den Inhalt eines psychologischen Entwicklungsprozesses bilden könnte? M.E. sind Alternativerklärungen, die die Entwicklung zum ZIPFschen Gesetz hin als Nebenprodukt erscheinen lassen, plausibler.

Zum empirischen Vorgehen beim Entscheidungsversuch: Alternativerklärungen des flacheren Anstiegs müßten ausgeschlossen werden, ehe man sich an tiefere theoretische Diskussionen gibt. Mir fallen auf Anhieb zwei ein (vielleicht nur in Unkenntnis dessen, daß sie längst erledigt sind):

a) Sind vielleicht die Wörter auf dem untersuchten Entwicklungsstand der Sprache noch keine abgegrenzten oder wenigstens stabilen Einheiten, so daß die aus der Erwachsenensprache übernommene Segmentierung den in der Kindersprache zu zählenden Gebilden nicht entspricht?

b) MANDELBROT (1965) weist auf Anomalien bei den ganz häufigen und den ganz seltenen Wörtern hin. Bei allen Kurven auf S. 555, 557 und 558 ist der Oberteil der Funktion abgeflacht. Man müßte die Erklärung ausschalten, daß die Abflachung bei den

Kindern daher rührt, daß nur die 70 "häufigsten" Rangplätze der üblichen 5000 bis 1000 besetzt sind.

W. Marx, E. Schürer-Necker:

Ad. 1: Tatsächlich ist es durchaus denkbar, stochastische Modelle zu entwickeln, die einen flacheren Geraden-Verlauf als a = -1 ermöglichen. Jedoch ist dieses im Grunde gar nicht der springende Punkt unserer Untersuchung. Was gegen die Annahme von nur auf Zufallsprozessen aufbauenden Modellen spricht, ist das Faktum, daß sich die Verteilung für die Kindersprache systematisch von der Verteilung für die Erwachsenensprache unterscheidet. Wenn tatsächlich nur Zufallsprozesse für diese Worthäufigkeitsverteilungen verantwortlich wären, ist nicht einzusehen, warum es zu einem solchen Unterschied und gerade in dieser Richtung (auf einen flacheren Verlauf der Kindersprache hin) kommen sollte. Dieses Ergebnis kann auf der Basis eines rein stochastischen Modells weder vorhergesagt noch erklärt werden.

Man könnte sich jetzt darauf zurückziehen, daß für die Kindersprache und die Erwachsenensprache eben unterschiedliche Parameter gelten. Das wäre jedoch nur eine andere Beschreibung des empirischen Befundes, keine Erklärung dafür. Gerade um solche unterschiedlichen Parameter zu erklären, wird es unseres Erachtens notwendig, Modellannahmen zu machen, die über das bloße Wirken eines Zufallsgenerators hinausgehen. Damit sind wir jedoch bereits beim nächsten Punkt angelangt.

Ad. 2: Richtig ist sicher, daß in beiden Modellklassen Zufallsprozesse eine grundlegende Rolle spielen. Das von uns als "ökonomisch" bezeichnete Modell verfügt jedoch nicht nur über einen Zufallsgenerator, sondern darüber hinaus über einen Selektionsmechanismus, der nun gerade nicht zufällig selegiert, sondern nach ganz bestimmten Prinzipien, in diesem Falle nach dem Prinzip der Optimierung des Aufwandes für den Benutzer der Sprache.

Es geht dabei - wenn man den Benutzer der Sprache in den Kalkül miteinbezieht - tatsächlich um effektiven Aufwand, um Kosten (im Sinne von Speicherkapazität und Energieaufwand der Zeichenproduktion).

Der Einwand, daß beide Modelle sich in ihren Vorhersage- bzw. in ihren Erklärungsmöglichkeiten nicht unterscheiden, gilt nicht mehr nach unseren Befunden.

Ad 3: Der Aufwand des Benutzers der Sprache (im Sinne der unter 2 genannten effektiven Kosten) steht - wie die Studie von MANDELBROT (1954) gezeigt hat - in keiner linearen Beziehung zur Entropie des verwendeten Sprachsystems. In den Extrembereichen sehr hoher und sehr niedriger Entropie besteht ein ungünstigeres Verhältnis von Aufwand zu vermittelter Information als im Mittelbereich.

Im Sinne dieser Analyse besteht der Mehraufwand der Kindersprache darin, daß sich die Auftretenshäufigkeiten von kurzen und langen Wörtern nicht so stark unterscheiden wie bei der Erwachsenensprache. Anders formuliert, Kinder benützen lange (und damit relativ aufwendige) Wörter relativ häufiger als Erwachsene. Dadurch wird der Informationsgehalt dieser Wörter bei gleichbleibenden Kosten für den Sender reduziert.

Ad 4: Prinzipiell besteht das genannte Problem in der Tat sehr oft, zwischen verschiedenen Modellen zu unterscheiden, die zu gleichen oder doch ähnlichen Vorhersagen führen. In dieser Situation empfiehlt es sich, nach Bedingungen zu suchen, unter denen unterschiedliche Erwartungen formuliert werden können, die zumindest eine Alternative ausschließen. Genau diesen Weg haben wir in unserer Arbeit zu gehen versucht.

Das schließt natürlich nicht aus, daß noch andere Alternativ-Hypothesen gefunden werden, die unsere Ergebnisse auch erklären können. In diesem Falle muß dann erneut nach Differenzierungsmöglichkeiten gesucht werden. Bevor wir auf die beiden hier vorgelegten Alternativerklärungen eingehen, möchten wir noch kurz bemerken, daß wir die Optimierung des Energieaufwandes für die Benutzer von Sprachen in der Tat für so wichtig erachten, daß wir die Annahme eines Entwicklungsprozesses, der auf diese Optimierung abzielt, für sehr wahrscheinlich halten.

Ad 4a: Daß die Wortbegrenzungsmarken von Anfang an stabil sind, erscheint anhand der Beobachtung des konkreten Sprachverhaltens von Kindern sehr wahrscheinlich. Ein Kind des hier untersuchten Stadiums, das in der Lage ist, "Mamma Ball" zu sagen, wird niemals "Mamm Aball" oder "Mammab All" sagen. Die Wörter sind schon sehr früh eindeutig abgegrenzte Einheiten.

Ad 4b: Tatsächlich ist es so, daß in dem hier untersuchten Stadium des Zweiwortsatzes gerade die normalerweise häufigsten Wörter nicht vorkommen, also die Artikel, Konjunktionen und ähnliche Formwörter. Dieser besonders flache Kurvenanteil entfällt also bei unserer Sprachstichprobe. Die geäußerte Vermutung erweist sich also als nicht stichhaltig.

Literatur

MANDELBROT, B., Structure formelle des textes et communications; deux études. Word 10, 1954, 1-27.

MANDELBROT, B., Information Theory and Psycholinguistics. In: WOLMAN, B. (Ed.), Scientific Psychology. New York 1965.

MILLER, G.A., CHOMSKY, N., Finitary Models of Language Users. In: Handbook of Mathematical Psychology II, New York 1965.

SIMON, H.A., On a Class of Skew Distribution Functions. Biometrika 43, 1956, 425-440.

THE INFLUENCE OF DEMOGRAPHIC VARIABLES ON LINGUISTIC PERFORMANCE*

Ora Rodrigue-Schwarzwald, Ramat Gan

Foreword

Considerable research has been devoted in recent years to the different languages used by different social classes. There is a wide divergence among social dialects, reminiscent at times of the disparities among dialects of different geographic areas, but without the differentiation of registers that one finds in the speech of one person in different circumstances (HASAN 1973). It is this area of social dialects which will be our prime concern in the present paper. We shall describe the language used by the advantaged classes, as opposed to the disadvantaged, taking into consideration demographic variables of age and comparing the language of various age-groups - children vs. older groups.

Most of the studies examining language between advantaged and disadvantaged children were based on an analysis of vocabulary and syntactic structures (BALGUR 1974, BERNSTEIN 1971, CAIS 1977, LAWTON 1968, MINKOVITZ 1969, NIR 1974, STAHL 1977). Very little work has been done in connection with the morphological and phonological aspects of grammar. While some studies in developmental psychology have dealt with the phonetic side of acquiring an inventory of phonemes as manifested in different social groups (IRWIN 1948, TEMPLIN 1957), none has extended to general descriptions of these phenomena. Moreover, many of the descriptions have been limited to written language, without reference to spoken, uttered language. Insofar as they dealt with spoken language, this was purely for the sake of syntactic and lexical analysis. The study described in the present paper includes a linguistic analysis from the standpoint of phonology and morphology, based on utterances of advantaged and disadvantaged students of various ages as they read out lists of sentences. Its vantage point has been linguistic rather than educational or sociological (cf. papers by CAIS 1977, DAVIS 1976, STAHL 1977), its many implications for these areas notwithstanding.

Much has been written about the differences between the language of the advantaged and the disadvantaged with respect to syntax and vocabulary. The present study will not try to determine whether the language of the disadvantaged is inferior or the language of the advantaged superior (BEREITER 1965) nor will it concern itself with the position whereby each of the two is a separate entity which cannot be compared with the other so that there can be no question of preference or superiority of one over the other (LABOV 1970). Whatever the vantage point, it is clear that the two languages - of the advantaged and the disadvantaged - do differ. Our study too revealed differences and we shall elaborate on these below. For lack of any other criteria, we have resorted to the accepted morphological standards and have used them as a yardstick for determining the degree of deviation in each of the two populations. Without studying each of the languages separately, we compared the two, using a single tool for our comparison. This is not to say that we have shown a preference for the linguistic norm which we chose to use. We shall try to explain the derivation of the differences and to characterize the deviations.

The study focused on an examination of the verb - a closed,

structured morphological category. We have examined the following questions:

a) What is the difference between the advantaged and the disadvantaged populations in the performance of verbs?

b) What are the differences between the performance of uncommon and common verbs?

c) Are there any differences among age-groups in the performance of verbs?

The study assumed differences between the populations and between the performance of common and uncommon verbs. It will be seen that the performance of both populations is better in the case of the common verbs than the uncommon ones and the disparity in the performance of the two populations is smaller in the case of the common verbs than the uncommon ones. Furthermore, the number of errors will be seen to decrease with age, i.e., the younger subjects will commit more errors than the older ones.

Description of the Study

Three hundred and twenty subjects took part in the study, taken from both populations, with four age-groups from each: 10-11, 13-14, 17-18 and 20+. The two populations were advantaged and disadvantaged. The samplings of the advantaged population were chosen from among fifth and eighth graders in advantaged schools (as defined by the Ministry of Education), twelfth graders from regular high schools and university students. The sample groups of the disadvantaged population were taken from among fifth and eighth graders at disadvantaged schools (as defined by the Ministry of Education), working youths who study one day a week in a school for apprentices (aged 17-18) and students in adult education classes at night school, in the process of completing their elementary or high school education (aged 20 +). There was a total of eight groups, each consisting of forty subjects. Two lists of 120 sentences were drawn up for the purposes of the study and each student was asked to read one of the lists out loud. He was told that the aim of the study was to verify whether partially vocalized texts facilitate reading[1]. The reading was recorded on tape. Half of the subjects in each group read one list and half read the other. The sentences were brief, consisting of 3-4 words. Each contained a Hebrew verb conjugated in the third person singular past, present or future. All of the words in the sentences were partially vocalized, so as to facilitate their identification.

As we have stated, we limited our study to the verbs. In order to examine the spirantization rule[2] in Modern Hebrew, we chose only those verbs which lend themselves to spirantization. Namely, we selected from the Hebrew dictionaries all the verbs which include orthographic <p>, <b> or <k> as first and/or second radical. These amount to 3635 verbs, multipled by the three tenses to a total of 10,905 verbal forms. Three hundred and thirty-four verbs were common (with a frequency of no less than 5 in a random sampling of 200,000 words). The other 3301 verbs were uncommon. This ratio of common (10%) and uncommon (90%) verbs was retained in the construction of the lists. Twenty-four verbs out of the 240 verbs (12 in each list) were common; 216 verbs were uncommon[3]. There were 124 verbs with first radical <p, b, k> and 134 verbs with second radical <p, b, k>[4]. Seventy-four verbs occured in the past tense, 83 in the present, and 83 in the future tense. The common verbs were randomly scattered among the uncommon verbs.

The other words in the sentences were designed only to provide the context for the verbs and they were selected in keeping with the content of the given verb.

It will be recalled that most of the verbs were uncommon and that only the essential vocalization was added to them. This rendered the linguistic stimulus somewhat vague since, although the letters are familiar and the relevant vocalization is provided, the word is not part of the vocabulary of most of the speakers and its meaning is therefore not familiar to them. A speaker who has been exposed to vague linguistic stimuli is bound to try to clarify them; i.e., to shape them in accordance with those inner linguistic processes which he possesses.[5] He will apply the linguistic rules and the vocabulary at his disposal in order to minimize the vagueness of the stimuli. Thus his performance will shed light on his linguistic sets. The performance of the uncommon vs. the common elements, in which the linguistic stimulus is assumed to be more familiar, will be described below.

Methodological problems

The act of reading out loud is an artificial form of speech. Free conversation is more conducive to revealing the natural phonological and morphological phenomena of language. The drawback of free conversation is its very randomness. The researcher has no control over the utterances. While he may be able to exercise some control over the content of the conversation, the researcher has no way of ensuring the subjects' use of those grammatical forms which he wishes to examine and describe. General phonetic and other random grammatical phenomena can be described on the basis of free conversation but they are not controlled by the researcher. Our study was controlled in the sense that it examined the verbal aspect. The researcher exerted control over the data by virtue of the fact that each subject was required to use the verbs on his list of sentences. The errors in these verbs do not detract from the validity of the test, since such errors are just as apt to occur in free speech.

A test based on reading may be said to have another drawback. One might argue that the experiment tests reading ability rather than linguistic performance. A gap between age-groups might therefore be attributed to the fact that the younger subjects do not yet read well while the older ones are proficient at reading. The same may be said of the difference between the advantaged and the disadvantaged populations. Here too the gap is due to a disparity in reading ability and not in language . The advantaged read well whereas the disadvantaged do not. This argument is not entirely well-founded either. First, the subjects in the study were fifth graders long past the age at which children first learn to read. Secondly, all of the subjects chosen for th experiment - including those of the disadvantaged population - were proficient readers, capable of forming letters into units of meaning. Third, insofar as the experiment tested reading, it did so to the same extent for both the advantaged and the disadvantaged populations, bearing in mind that 90% of the verbs were uncommon. Furthermore, the pronounciation revealed clear morphological and phonological tendencies. There were very few complaints about explicit difficulties in reading and these few occurred to almost the same extent in all of the groups.

The conditions of the experiment were artificial in another respect as well. The subjects were forewarned that they might not understand some of the words they would be reading. Such a

situation is reminiscent of an exam and may give rise to some nervousness which, in turn, is liable to affect performance. Since this factor was shared by all of the subjects, its influence was uniform and may be ignored. Moreover, although it is well-known that there is an influence of the exam situation on its results, I doubt if there is conclusive evidence concerning the dependence between population class and anxiety level on exams. Furthermore, in this study the experiments explained that this was not an exam, and that the results would be given only to the investigator[6].

Another problem which needs consideration bears upon generalization. Can the findings be replicated if one applies different language material, or samples other subjects from the same population? In reference to this issue it should be emphasized that similar significant results were obtained in other linguistic studies on independent grounds (SCHWARZWALD 1978a, 1978c). Examples drawn from free conversation as well as planned techniques demonstrated the same linguistic phenomena discussed below. Secondly, one must keep in mind that in the present study two independent lists were read by different subjects, and in spite of that, similar results were found. This can be seen as a double-cross validation which enhances the validity of our study.

<u>Findings</u>

The recordings were analysed and the results were drawn up in writing. This analysis set out to determine whether there was any deviation from the performance regarded as normative. Normative performance is the formal pronunciation and morphophonemic formation considered correct by the Hebrew Language Academy and taught in grammar lessons at Israeli schools.

The average errors in each group are given in table 1.

Table 1: Average errors in common and uncommon verbs by age and population class[7].

Verb type	Population	Age			
		10-11	13-14	17-18	20
Common	Advantaged	1.18	0.90	1.10	0.55
	Disadvantaged	3.30	3.32	2.38	2.32
Uncommon	Advantaged	4.38	3.50	2.58	2.50
	Disadvantaged	6.45	6.05	5.85	4.58

The errors were also calculated in terms of percentages for each group of verbs, common vs. uncommon, in each age-group and in each type of population (Table 2).

Table 2: Average percentage of errors in common and uncommon verbs by age and type of population.

Verb type	Population	Age				Average by Population	Average by Verbs
		10-11	13-14	17-18	20		
Common	Advantaged	9.8	7.5	9.2	4.6	7.8	15.7
	Disadvantaged	27.5	27.7	19.8	19.4	23.6	
Uncommon	Advantaged	36.5	29.2	21.5	20.8	27.0	37.4
	Disadvantaged	53.8	50.4	48.8	38.1	47.8	
Overall Average by Age		31.9	28.7	24.8	20.7		

A three-way analysis of variance with repeated measures of the dependent variables was then conducted. The results are presented in Table 3.

Table 3: Analysis of variance for the number of errors with repeated measures on common and uncommon verbs.

Source	SS	df	MS	F
Between subjects				
Age (A)	161.16	3	53.72	14.77
Population class (B)	772.21	1	772.21	212.31
A x B	6.95	3	3.32	<1.00
Subjects within groups	1134.82	312	3.64	
Within subjects				
Verb type (C)	1084.21	1	1084.21	549.84
A x C	23.99	3	7.99	4.05
B x C	14.10	1	14.10	7.15
A x B x C	26.99	3	8.99	4.56
C x subjects within groups	615.22	312	1.92	

Several conclusions emerge from the results presented in these tables:

a) The incidence of errors in the uncommon verbs (M=37,4%) is significantly higher (F (1,312)=549.84, p<.001) than that of the common verbs (M=15.7%).

b) The disparity among age groups is also significant [F (3,312) 14.77,p<.001]. As the age increases, the average of errors decreases. The highest incidence of errors occurs among the fifth graders (M=31.9%), declining steadily and reaching its lowest point at the age of 20+ (average 20,7%).

c) The difference between populations is significant [F (1,312) =212.31, p<.001], with that of the disadvantaged population (M=23.6% for the common and 47.8% for the uncommon verbs) being higher than that of the advantaged population (7,8% average for the common and 27.0% for the uncommon verbs.)

d) There is a significant three-way interaction between common and uncommon verbs, age groups, and population types [F(3,12)= 4.46, p<.01]. It is attributable to both of the factors enumerated below.

e) There is a significant interaction between the common and the uncommon verbs and between population types [F(1,312)=7.15, p<.001], owing to the fact that the discrepancy between disadvantaged and advantaged populations is higher for the uncommon than for the common verbs (see Fig. 1). The discrepancy is smaller for the common verbs (average discrepancy of 15.8%) than for the uncommon ones (average discrepancy of 20.8%). The interaction also results from the differences between the very similar averages of the advantaged population (ranging from 9.8% to 4.6%) for the common verbs, versus the gradual decreases of averages (ranging from 36.5% to 20.8%) for the uncommon verbs. The disadvantaged population shows constant decrease of errors in both types of verbs (ranging from 27.5% to 19.4% for the common verbs and 53.8% to 38.1% for the uncommon verbs), although the rate of decrease varies among the age groups.

f) There is also some interaction between verb types and age-groups [F(1,312)=4.05, p<.01], deriving from the fact that the average of errors in the uncommon verbs gradually declines with age (see Fig. 1 and the discussion below) while the average of errors for the common verbs in the advantaged group increases (M= 9.2%) in the twelfth grade after declining in the eighth (M=7.5%) and declines once again at the age of 20 + (M=4.6%). In the disadvantaged population, on the other hand, the incidence or errors declines significantly at the age of 17-18 (from M=27.5%, 27.7% to M=19.8%).

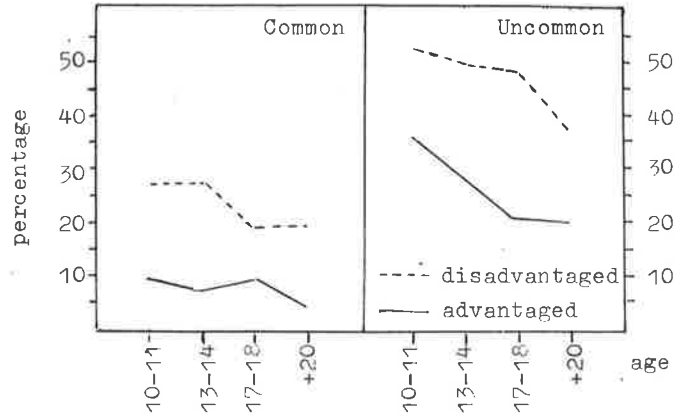

Fig.1. Average percentage of errors in common and uncommon verbs by age and type of population

Discussion

The above findings support the assumptions of the study: less errors in morphophonemic performance of common verbs than of uncommon ones for both populations; differences between the two populations, with a higher incidence of errors in the disadvantaged than in the advantaged population; differences among age groups, with the fifth graders accounting for the highest incidence of errors and the adults for the lowest.

Every child is born with the ability to generate language by constructing the rules of grammar. The child has some inner, general conceptions about language. When he encounters data from a language he is able to formalize the rules and generate the correct sentences of the language. The mistakes he commits are a result of a false generalization regarding the rules of grammar. After facing new data, he re-constructs the rules and generates the sentences properly (CHOMSKY 1969, CHOMSKY 1965, CHOMSKY & HALLE 1968). The child is capable of generating an enormous variety of sounds and syllables; yet being exposed to a limited number of phones in his language, he learns to use only these. Some of the universal, innate concepts of language are extinct due to a lack of active use.

Any child masters the spoken language before he gets to elementary school. He needs no study of all the basic rules of the language. The task of the language teacher is mainly a stylistic one. He exposes the students to other possible syntactic structures, morphemes, and words, and thus he enables them to enlarge their generative ability. Therefore the more educated the speaker the wider his language control.

The linguistic difference between an advantaged and a disadvantaged population is due to differences in education. Both groups start elementary school at unmatched levels. The advantaged child often hears formal and educated language at home. He is exposed to books, and his vocabulary is larger than that of the disadvantaged child (BALGUR 1974, LAWTON 1968). The latter is not so frequently talked to: in the main, the only utterances he hears are brief commands, and the vocabulary which he hears is very poor (STAHL 1977). With students at such varied levels, the language teacher has entirely different tasks for each population. With the advanced population he can broaden the linguistic knowledge which already exists in the students. With the disadvantaged population he must first of all teach the basic rules of grammar and try to widen the students' vocabulary.

This problem is especially acute in Israel. Hebrew is the official language spoken in Israel; however, many of the speakers are of a foreign origin, so that even advantaged chilren may hear

wrong forms of Hebrew at home. Moreover, the norms for the present grammar are anchored in the past, especially in the morphophonemics of Biblical Hebrew. The students learn the basic morphophonemic rules of the old language, and this of course does not necessarily reflect their own language.

Therefore there are two issues in language competence: one has its sources in the general difference between the populations, the other derives from the special state of Hebrew, which has not yet acquired new norms.

The performance shown in the results presented in this paper strongly reflects these issues, particularly because the mistakes were measured according to the grammatical norms taught in schools. There is no doubt that the ability to generate the rules for the obscure verbs increases with education. Many of the rules required for the formation of the verbs are not internalized in early childhood. They need much practice before they are mastered by those who complete a formal education. The number of mistakes prove this point. The disadvantaged students acquire less rules; therefore, their rate of errors is high. Furthermore, the size of the vocabulary plays an important role in the morphophonemic performance. The advantaged population masters a larger vocabulary, including verbs. The advantaged have already been exposed to many verbs in various verbal patterns, and therefore they generate the forms better. Contrarily, the disadvantaged, whose vocabulary is very poor and do not use many verbal patterns, commit very many normative mistakes. It is worth noting that advanced students of the third and fourth age groups (17-18 and 20+ years old) make almost the same rate of mistakes in the uncommon verbs as do their disadvantaged mates in the common verbs.

verbs having deviant and unusual morphological conjugations (e.g., [hofia, mexin] 'appeared, prepare') (of the weak verb category), their formation was readily mastered by virtue of their frequency. The uncommon forms were not familiar to the speakers, who used varicus ways of clarifying the vague stimuli at their disposal. The average of errors, even for the uncommon verbs was less than 50%. In other words, there was a tendency to pronounce them according to normative rules, but the impact of knowledge and learning was evident. The incidence of errors was lower in the advantaged population with its greater exposure to a large vocabulary and to formal learning of the rules of grammar, than in the disadvantaged group, but was still relatively high. The average incidence of errors in the performance of uncommon verbs in the advantaged groups was even higher than the average of errors in the performance of the common verbs by the disadvantaged groups.

The disparity between the advantaged and the disadvantaged populations was manifested in the performance of both types of verbs but, as we have seen, it was smaller in the case of the common verbs. The maximal disparity in the case of the common verbs was found at the eighth grade level (M=7.5% vs. 27.7%; i.e., a disparity of 20.2%) and the minimal disparity was found in the 17-18 age-group (M=9.2% vs. 19.8%; i.e., a disparity of 10.6%). On the other hand, the maximal disparity in the case of uncommon verbs was found in the 17-18 age-group (M=21.5% vs. 48.8%; i.e., a disparity of 27.3%) and the minimal disparity occurred among the fifth graders (M=36.5% vs. 53.8%; i.e., a disparity of 17.3%). This is a significant difference. The common verbs were familiar to all the subjects, for which reason the incidence of errors was low in both populations. The uncommon verbs were unfamiliar and it was here that the disparity between the two popu-

It is worth noting that errors occurred in the morphophonemic performance of both the common and the uncommon verbs in all of the groups. While the incidence of errors in the common verbs was low, there were nonetheless some common verbs with a relatively high incidence of errors; namely, [yizaxer] 'will be recalled', [yimaxer] 'will be sold', [yexave] 'will be hidden', It is no mere coincidence that all three of these verbs are in the <u>nifʕal</u> conjugation, future tense. While all three - [nizkar, nimkar and nexba] -'was recalled, sold, hidden' are common verbs, their usage has not been examined in all tenses. They are more frequently used in the basic forms (past and present tenses) than in the future tense, which involves forms that are less common. Moreover, other examinations conducted as part of our study revealed that, on the whole, performance of the future tense was poorer than that of the past and present tenses. It was also found independently that the future is less frequently used, which accounts for the rather high incidence of errors in these verbs in all of the groups in the study. A very high incidence of errors (approximately 50%) occurred in the disadvantaged population in the case of the verbs [yitba] 'will drown' and [yəcape] 'will expect' as well, both of them in the future tense. The verb [yitba] was most often pronounced [yitva], by analogy with the past and present forms [tava, tovea]. The performance of [yəcape] involved changes of conjugation, primarily to the <u>kal</u> conjugation (i.e., [yicpe]) or a spirantization, of the stop [yəcafe]. It would appear that this word is not in common use among the disadvantaged, possibly because of its somewhat abstract meaning.

There was a low average of errors in the common verbs because of their familiarity and frequent occurrence. Even in the case of lations was most marked. In the case of the common verbs, the average of errors was very low in the advantaged population. The differences among the various age-groups in this population were very slight and insignificant. In the disadvantaged population, there was an immaterial difference in performance between the fifth and the eighth graders and between the 17-18 and 20 age-groups. The marked decline in the number of errors occurred between the ages of 13-14 and 17-18.

A different pattern was found to exist in the case of the uncommon verbs. The decline in the incidence of errors in the advantaged population occurred until the age of 17-18. There was almost no difference between the 17-18 age-group and the 20 + group in the performance of uncommon verbs, whereas in the disadvantaged population the decline in the incidence of errors was very slight until the age of 17-18, increasing only between the ages of 17-18 and 20+ (see Fig. 1). It is worth noting that the disparity between the 10-11 age-group and the 20+ group was the same for both populations (M=36.5% vs. 20.8% in the advantaged population and 53.8% vs. 38.1% in the disadvantaged population; i.e., a disparity of 15.7% in both.)

The decline in the incidence of errors in uncommon verbs in the advantaged population may be attributed to the following factors: the students acquire a formal education and an extensive knowledge of grammar until the twelfth grade of high school. They acquire a large vocabulary and are aware of the formal and normative rules, so that they commit progressively fewer errors until the age of 17-18. This incidence then remains stable and is not significantly different at the age of 20+, unlike the case in the disadvantaged population where the vocabulary of the subjects did not increase to the same extent. Namely, in the case of the

disadvantaged population, most of the uncommon words remained as vague at a later age. One must bear in mind that the subjects in the 17-18 age-group were working youths, enrolled in programs of compulsory schooling one day a week. They had not made a free choice to study any of the subjects being taught in the program, including language and grammar and that accounts for the high incidence of errors. This situation changes at the age of 20+, as youths who had previously dropped out of school for various reasons resume their studies in an attempt to complete their education. In a process centered on self-learning, they expand their vocabulary and acquire a somewhat greater knowledge of formal grammar, so that the incidence of errors in their performance declines. Nonetheless, the incidence of errors in the 20+ group of the disadvantaged population remains slightly higher than that of the fifth graders in the advantaged population - a noteworthy result.

As stated above, there were no significant differences among the age groups in the advantaged population performing the common verbs. In the disadvantaged group too the improved performance of common verbs occurred at the age of 17-18, in contrast to the pattern for the uncommon verbs. These results apparently reflect the poor reading ability of the fifth and eighth graders in this population, which is manifested in the case of the uncommon verbs as well. However, while a knowledge of grammar is of no help in the performance of uncommon verbs by the 17-18 group, the performance of the common verbs is more correct by virtue of their familiarity.

Linguistic Phenomena Revealed in the Study

Both the numerical results and the graph reveal several interesting points with respect to the differences among populations, among verbs and among age-groups. Also worth noting are the types of errors and the manner of their occurrence in the subjects' performance. In the present section, we shall focus on some of the findings which emerge from the examination of the performance of verbs, so as to learn more about their use in practice:

A. Phonology:

1. Verbs containing pharyngeal and guttural consonants acted as weak verbs[8], indicating that the consonants [h ʔ ʕ] have become weaker. On the other hand, verbs containing the letters <?> [?] or <k> [k,x] were treated entirely as pharyngeal verbs, e.g., [macbia] instead of of [macbi] 'recruit', [maxalim] instead of [maxlim] 'insult'. The guttural sounds of modern Hebrew have been weakened. The confusion in the performance of these sounds may be attributed to a growing ambiguity with regard to the explicit phonetic conditions required for observance of the rules of pharyngeal pronunciation.

2. There was a difference between the first and the second radicals in the application of the spirantization rule (SCHWARZWALD 1978b), the first being more productive than the second. In the first radical, it was pronounced as a stop word initially and in the hitpaʕel conjugation, while in all of the remaining cases it was a fricative. In the second radical, it was generelly performed as a stop. Its pronunciation as a fricative was most often found in the kal conjugation (past and present), in denominative verbs and in verbs whose third radical is fricative.

3. There were many instances of metatheses. In metathesis, there is a transposition of the consonants in the verb. The prevalent tendency was to replace an uncommon verb with a common one; e.g., [ʔušpar] 'was finished' became [ʔufšar] 'was made

possible' and [nivlaš] 'was investigated' became [nilbaš] 'was worn'.

4. In some conjugations of the verb (present kal and future nifʕal conjugations and all tenses of the piʕel and the hitpaʕel), the vowel of the second radical sometimes became [a] or [ə] rather than [e] - a morphophonemic tendency which merits special attention. It could be explained in two ways: a. The influence of other forms, since the vowel of the second radical is [a] in the other conjugation, or b. perhaps the vowel of the second radical throughout the verbal system is a mid-central vowel /ə/ and all its actualizations are phonetically close, namely, [ə], [a], or [e]. This tendency requires separate study in free conversation.

5. From the standpoint of phonetics, both assimilation and dissimilation were found to occur, depending on the circumstances: in quadrilateral and geminate verbs, the identical consonants were performed in the same manner (both as fricatives or both as stops). On the other hand, in homorganic roots, in which the first and second radicals are the same, there was a tendency towards dissimilation: consonants were replaced with other consonants or else their phonetic quality as stops or fricatives was altered.

B. Morphology:

1. The incidence of errors in verbs conjugated in the passive conjugation was higher than in the different types of active conjugations (SCHWARZWALD 1978c). This finding reinforces the observation whereby the uses of the passive are less common, which in turn accounts for their lower frequency and the higher incidence of errors in their performance (cf. LAWTON 1968: 109,

BERNSTEIN 1971: 118).

2. Of all the tenses, the future was marked by the highest incidence of errors. As argued above, the future tense is less often used. The incidence of errors in its use may, however, also be due to its spelling. The spelling <ypʕl> is typical of five different conjugations (yifʕal, yipaʕel, yəfaʕel, yəfuʕal, yufʕal). The partial vocalization notwithstanding, a certain amount of confusion occurred with respect to these different forms.

3. There were many instances of conjugations being interchanged, especially when the given spelling may be accored several meanings, in spite of the partial vocalization. The prevalent tendency was to replace a less common verbal conjugation with a more common one of the same root. There were relatively few replacements of the hitpaʕel and hifʕil, since the spelling of these is unequivocal. The nifʕal and the hitpaʕel were interchanged on several occasions, perhaps because of their semantic affinity as conjugations which express reflexiveness and reciprocity.

4. There were some instances of weak verbs being interchanged; i.e., a verb in one weak pattern behaves as though it belongs to a different weak verb pattern. This is consistent with other findings (SCHWARZWALD 1978a) of the interchange of patterns. Especially interesting is the fact that replacements were found to occur in other weak verb patterns as well.

These findings appeared in the speech of all subjects in the experiment, occurring with greater frequency in the disadvantaged population with its smaller vocabulary and its lack of exposure to the standardized formal learning of grammatical rules. The

phenomena observed here are general ones which also figure in the subjects' free conversation.

Summary

In our study, we found significant differences between the types of verbs (common and uncommon), types of population (advantaged and disadvantaged) and age-groups (from fifth graders to adults). The demographic variables of age and social status influence the speaker's linguistic performance, not directly but as a by-product of the learning factors. The younger speakers were less conscious of the formal grammatical rules of the language, which accounts for the higher incidence of errors in their speech. The adults are exposed to these rules, for which reason the incidence of errors in their case declines steadily. The differences between the advantaged and the disadvantaged populations are also attributable to the influence of schooling. The disadvantaged do not complete their formal linguistic education and have a high incidence of errors, while in the case of the advantaged population, which continues to study, the incidence of errors declines.

The types of errors appear to indicate that the most conspicuous linguistic tendency manifested here is that of simplification. The term simplification is used in its linguistic meaning. On the one hand, simplification is the process of changing grammatical, marked forms into unmarked and simple ones by means of analogy, e.g., morphological attraction, paradigmatic regularity or lexical resemblance (BLOOMFIELD 1933, de SAUSSURE 1959, KING 1969). On the other hand, simplification is referred to in connection with enonomy and generality: the more economy involved in the morphophonemic system, in its feature system, phonemic inventory or in the rules, the more simplification takes place. Furthermore, the more the speaker can generalize the rules of grammar, the simpler the grammar is (HARMS 1968, HYMAN 1975). All the linguistic phenomena described above prominently exhibit the realization of simplification in its two senses[9]. The speakers simplify the rather vague stimuli at their disposal and transform them into simpler ones from the standpoint of their phonetic and morphological structure and common ones from the standpoint of usage. This process is especially marked in the case of the disadvantaged population. Thus, we find that while the high incidence of errors in the disadvantaged group does point to deviations from the formal norm, it also indicates a tendency towards simplification in the processes - one which is inherent to language itself.

Footnotes

* This study was sponsored in part by the Israel Commission for Basic Research and the Research Department of Bar Ilan University. A preliminary version of the paper was read at the Seventh World Congress of Jewish Studies in Jerusalem, in August 1977.

1 Hebrew spelling is predominantly consonantal. There are several vowel letters but they do not provide all variations of the vowels. These are provided by the additional vocalization signs beneath and above the letters.

2 The rule which alters the stops k, b, p into fricatives x, v, f, respectively, under certain circumstances.

3 They were sampled at random.

4 The sum of 124+134 does not equal 240 because some verbs included both first and second radical <p, b, k>.

5 This is in fact a general psychological phenomenon related to the manner of grasping and responding to a vague stimulus. Rorschach inkblots are an example of this.

6 It is worth adding here that half of the experimenters responsible for recording the subjects were of Sephardic origin and half of Ashkenazi origin, and that all worked with subjects from both populations. The origin of the person in charge of the recording had no bearing on the subjects' scores.

7 In order to make the comparison possible, the number of average errors for each subject was devided by 9 for the uncommon verbs, since the number of the common and uncommon verbs was not equal.

8 Verbs with wy? or a historical initial ŋ are conjugated somewhat differently than the strong verbs, which retain all of the radical consonants in all parts of the conjugation.

9 Detailed discussion of the processes and their impact on linguistic performance will be included in my book, Form and Formation in the Hebrew Verb.

References

BALGUR, R.A., Basic Word List from Newspapers for the Cultural Disadvantaged. Megamot 20, 1974, 225-253 (in Hebrew).

BEREITER, C., Academic Instruction and Preschool Children. In: Language Programs for the Culturally Disadvantaged. Report of the NCTE Task Force on Teaching English to the Disadvantaged, Champaign III, 1965.

BERNSTEIN, B., Class, Codes and Control. Vol. 1. Paladin 1971.

BLOOMFIELD, L., Language. New York 1933.

CAIS, J., An Israeli Test of BERNSTEIN'S Theory of Verbal Behavior. PhD. Thesis, Jerusalem 1977 (in Hebrew).

CHOMSKY, C., The Acquisition of Syntax in Children from 5 to 10. M.I.T. 1969.

CHOMSKY, N., Aspects of the Theory of Syntax. M.I.T. 1965.

CHOMSKY, N., HALLE, M., The Sound Pattern of English. New York 1968.

DAVIS, I., The Language of the Disadvantaged. Iyunim Bechinuch 1976, 133-138 (in Hebrew).

DE SAUSSURE, F., Course in General Linguistics (Ed. E. BALLY, A. SECHEHAYE, Trans. W. BASKIN). New York 1959.

HARMS, R.T., Introduction to Phonological Theory. New Jersey 1968.

HASAN, R., Code, Register and Social Dialect. In: BERSTEIN, B. (Ed.), Class, Codes and Control, Vol. 2. Routledge & Kegan Paul, London-Boston 1973, 253-292.

HYMAN, L.M., Phonology: Theory and Analysis. New York 1975.

IRWIN, O.C., Infant Speech. Journal of Speech and Hearing Disorders 13, 1948.

KING, R.D., Historical Linguistics and Generative Grammar. New Jersey 1969.

LABOV, W., The Logic of Nonstandard English. In: W. FREDRIK (Ed.), Language and Poverty. Markdam 1970.

LAWTON, D., Social Class, Language and Education. London 1968.

MINKOWITZ, A., The Disadvantaged Student. Jerusalem 1969 (in Hebrew).

NIR, R., The Instruction of Hebrew as Mother Tongue at High-Schools. Tel-Aviv 1974 (in Hebrew).

SCHWARZWALD, O.R., The Perception of the Weak Verb. To appear in Balshanut Shimushit 2, 1978a (in Hebrew).

SCHWARZWALD, O.R., Normativism and Naturalness in the Application of a Morphophonemic Rule in Modern Hebrew. To appear in Bar-Ilan Annual 1978b, (in Hebrew).

SCHWARZWALD, O.R., Tenses and Conjugations - their Reflection in an Empirical Study. To appear in Zeev-Ben-Haiim Jubilee Volume 1978c (in Hebrew).

STAHL, A., Language and Thought of Culturally Deprived Children in Israel. Tel Aviv 1977 (in Hebrew).

TEMPLIN, M.C., Certain Language Skills in Children. Inst. of Child Welfare. Monograph No. 25, 1957.

BIBLIOGRAPHY OF QUANTITATIVE LINGUISTICS IN POLAND

Irena Kamińska, Wrocław

This bibliography, which extends up to 1977, includes work which has been published by Polish authors inside and outside Poland, as well as works of foreign authors which appeared in Poland. It contains only publications on Linguistics in which theory of probability or mathematical statistics were used. We have not included publications in which merely numerical data about isolated linguistic elements are found. An exception to this is a small number of works of an innovative nature.

List of abbreviations

BFON	-	Biuletyn Fonograficzny. Poznań.
BPTJ	-	Biuletyn Polskiego Towarzystwa Językoznawczego. Kraków.
FO	-	Folia Orientalia. Kraków.
GZHum	-	Gdańskie Zeszyty Humanistyczne. Gdańsk.
JOS	-	Języki Obce w Szkole. Warszawa.
JP	-	Język Polski. Kraków.
KNf	-	Kwartalnik Neofilologiczny. Warszawa.
LP	-	Lingua Posnaniensis. Poznań.
PHum	-	Przegląd Humanistyczny. Warszawa.
PF	-	Prace Filologiczne. Warszawa.
PJ	-	Poradnik Językowy. Warszawa.
PL	-	Pamiętnik Literacki. Warszawa.
Prace Śl	-	Prace Naukowe Uniwersytetu Śląskiego. Prace Językoznawcze. Katowice.
RSl	-	Rocznik Slawistyczny. Kraków.
SO	-	Slavia Occidentalis. Poznań.
Spr PAN	-	Sprawozdania z Prac Naukowych Wydziału Nauk Społecznych Polskiej Akademii Nauk.
StP	-	Studia Polonistyczne. Uniwersytet im. A.Mickiewicza w Poznaniu. Poznań.
ZM	-	Zastosowania Matematyki. Warszawa.
ZNUG	-	Zeszyty Naukowe Uniwersytetu Gdańskiego. Gdańsk.
ZNUŁ	-	Zeszyty Naukowe Uniwersytetu Łodzkiego. Łódź.
ZN WSP Op	-	Zeszyty Naukowe Wyższej Szkoly Pedagogicznej w Opolu. Opole.
ZP	-	Zeszyty Prasoznawcze. Kraków.

1. ABERNATHY,R., Mathematical linguistics and poetics. In: Poetics, Poetyka, Poėtika, Vol.I. Warszawa: Państwowe Wydawnictwo Naukowe, 1962,563-569.

2. ARAPOW,M.W. [=Arapov,M.V.], Struktura ilościowa tekstu skończonego [Quantitative structure of a closed text]. In: Semantyka tekstu i języka. Wrocław: Ossolineum,1976,144-163.

3. BAJOR,K., Zastosowanie metod statystyki językoznawczej do analizy zasobu słownikowego podręczników języka rosyjskiego [The use of methods of linguistic statistics for the analysis of vocabulary of textbooks for teaching Russian]. Łódź:Uniwersytet Łódzki, 1970, 24 p.

4. BAJOR,K., Frekwencja semantyczna najczęściej występujących rzeczowników rosyjskich [Semantic frequency of the most frequent Russian nouns]. ZNUŁ 89,1972,45-54.

5. BARTKOWIAKOWA,A., Zagadnienia matematycznej lingwistyki [Problems of mathematical linguistics]. Matematyka 14/4, 1961,224-225.

6. BARTKOWIAKOWA,A., O rozkładzie określeń w zdaniach opisowych Żeromskiego i Sienkiewicza [On the distribution of complements in descriptive sentences of Żeromski and Sienkiewicz]. ZM 6/3, 1962,287-303.

7. BARTKOWIAKOWA,A., O rozkładzie i kolejności zdań współrzędnych w utworach powieściowych Żeromskiego i Sienkie-

wicza [On the distribution and order of hypotactic and paratactic sentences in the novels of Żeromski and Sienkiewicz]. ZM 7/2,1964,133-154.

8. BARTKOWIAKOWA,A., Matematyka i dochodzenie autorstwa [Mathematics and the determination of authorship]. Matematyka 20/3, 1967,99-106.

9. BARTKOWIAKOWA,A., GLEICHGEWICHT,B., O długości sylabicznej wyrazów w tekstach autorów polskich [On the syllabic length of words in texts of Polish authors]. ZM 6/3,1962, 309-319.

10. BARTKOWIAKOWA,A., GLEICHGEWICHT,B., Zastosowania dwuparametrowych rozkładów Fucksa do opisu długości sylabicznej wyrazów w różnych utworach prozaicznych autorów polskich [Application of Fuchs' two-parameter distributions to the study of syllabic word-length in the works of various Polish novelists]. ZM 7/4,1964,345-352.

11. BARTKOWIAKOWA,A., GLEICHGEWICHT,B., O rozkładach długości sylabicznej wyrazów w różnych tekstach [On the distribution of the syllabic length of words in different texts]. In:Poetyka i matematyka. Warszawa;Państwowy Instytut Wydawniczy, 1965, 164-172.

12. BARTKOWIAKOWA,A., ZUBRZYCKA,L., O polskich badaniach języka przy użyciu metod matematycznych [On Polish linguistic studies using mathematical methods]. Wiadomości Matematyczne 7/1,1963,87-97.

13. BARTMIŃSKI,J., Oralność tekstu pieśni ludowej w świetle statystyki leksykalnej [The orality of folk-song texts in the light of lexical statistics]. In:Semantyka tekstu i języka. Wrocław:Ossolineum, 1976, 175-201.

14. BAUDOUIN de COURTENAY,J., Ilościowość w myśleniu językowym [Quantity as a dimension of thought about language]. In: Symbolae grammaticae in honorem J.Rozwadowski, Vol.I. Kraków:Drukarnia Uniwersytetu Jagielońskiego, 1927,3-18.

15. BUDZYK,K., Lingwistyka statystyczna a zadania wiedzy o stylu i wierszu [Statistical linguistics and problems of stylistics and verse]. PHum 1,1959,31-40.

16. CZAPLEJEWICZ,E., Matematická poetika a studium poezie [Mathematical poetics and the study of poetry]. Česká literatura 13,1965,419-427.

17. CZAPLEJEWICZ,E., O matematycznym modelowaniu w poetyce [On the creation of mathematical models in poetics]. Ruch Literacki 7/2,1966,65-80.

18. CZAPLEJEWICZ,E., Poetyka matematyczna a badanie poezji [Mathematical poetics and the study of poetry]. In:Wiersz i poezja. Wrocław:Ossolineum, 1966,81-94.

19. CZEKANOWSKI,J., Polska synteza slawistyczna w perspektywie ilościowej [The Polish slavistic synthesis in quantitative perspective]. Kraków 1947, 50 p.

20. ČERVENKA,M., O semantyce czeskiego aleksandrynu [On the semantics of the Czech Alexandrine]. In: Wiersz i poezja. Wrocław: Ossolineum, 1966,21-32.

21. ČISTJAKOV,V.F., Častotnosti glasnyx i soglasnyx v 50 jazykax raznogo grammatičeskogo stroja [Frequencies of vowels and consonants in 50 languages of different grammatical structure]. LP 16,1972,45-48.

22. DOBRUŠIN,R.L., Matematyczne metody w lingwistyce [Mathematical methods in linguistics]. Rocznik Polskiego Towarzystwa Matematycznego, Seria 26, 1962-63/2,217-242.

23. DOROSZEWSKI,W., Mowa mieszkańców wsi Staroźreby [The dialect of the inhabitants of the village of Staroźreba]. PF 16,1934,249-278.

24. DOROSZEWSKI,W.,Pour une représentation statistique des isoglosses. Bulletin de la Société de Linguistique de Paris 36,1935,28-42.

25. DOROSZEWSKI,W., O statystyczne przedstawienie izoglos [On the statistical representation of isoglosses]. In: Doroszewski,W., Studia i szkice językoznawcze. Warszawa: Państwowe Wydawnictwo Naukowe, 1962,380-392.

26. DUKIEWICZ,L., Polskie diady spółgłoskowe typu TS i ST [Polish consonant pairs of the types TS and ST]. Warszawa: Prace Instytutu Podstawowych Problemów Techniki Polskiej Akademii Nauk, 1969, 28 p.

27. DZIURNIKOWSKI,A., Cyfrowa analiza i synteza mowy [Numerical analysis and synthesis of speech]. Zeszyty Naukowe. Akademia Spraw Wewnętrznych 12,1976,285-287.

28. EHRLICH,J., Dispersion of potentially distinctive features in contemporary Polish. BFon 14,1973,3-17.

29. EKIERT,T., SEIDLER,J., Badanie struktury słów języka polskiego [The study of word-structure in Polish]. Warszawa: Prace Instytutu Podstawowych Problemów Techniki Polskiej Akademii Nauk, 1961, 25 p.

30. FISIAK,J., Wstęp do współczesnych teorii lingwistycznych [Introduction to modern linguistic theories]. Warszawa: Wydawnictwo Szkolne i Pedagogiczne, 1975, 143 p.

31. FONTAŃSKI,H., Ponjatija i metody matematičeskoj statistiki i teorii verojatnostej, primenjaemye pri issledovanii rozmerov predloženij [Concepts and methods of mathematical statistics and probability theory applied to the study of sentence length]. ZN WSP Op 1972/9, 41-51.

32. GUIRAUD,P., Zagadnienia i metody statystyki językowej [Translation of:Problèmes et méthodes de la statistique linguistique. Translator:H.Kniagininowa]. Warszawa: Państwowe Wydawnictwo Naukowe, 1966, 159 p.
Reviewed by: Morawska,L.:KNf 8,1961,448-451;
Sambor,J.:PJ 1966/10, 432-434.

33. HOCKETT,Ch.F., Kurs językoznawstwa współczesnego [Translation of : A course in modern linguistics. Translator: Z.Topolińska, M.Jurkowski]. Warszawa:Państwowe Wydawnictwo Naukowe, 1968, 703 p.
Reviewed by: Piskorowska,H.: JOS 1969/3,171-172;
Zgółka,T.:Studia Metodologiczne 1971/8, 137-144.

34. IVANOV,V.V., O zastosowaniu metod ścislych v literaturoznawstwie [On the application of exact methods to the study of literature](Translation from the Russian). PL 59/2,1968,339-351.

35. IVIĆ,M., Kierunki w lingwistyce [Translation of: Pravci u lingvistici (Trends in linguistics). Translator:A. Wierzbicka]. First ed. Wrocław: Ossolineum, 1966,263 p., Second ed. Wrocław: Ossolineum, 1975, 378 p.
Reviewed by: Gwóźdź,A.: Poglądy 1976/23,12.

36. JANČAK,P., Možnosti statistického zpracování kolísavých nárečních jevů na jazykových mapách [Possibilities of statistical processing of oscillating dialect phenomena on linguistic maps]. PF 18,1963,59-78.

37. JASSEM,W., Fonetyczno-akustyczne założenia automatycznego rozpoznawania fonemów [Phonetic-acoustical bases for the automatic recognition of phonemes]. Warszawa: Prace Instytutu Podstawowych Problemów Techniki Polskiej Akademii Nauk, 1970, 25 p.

38. JASSEM,W., Mowa a nauka o łączności [Speech and communications science]. Warszawa: Państwowe Wydawnictwo Naukowe, 1974, 259 p.

39. JASSEM,W., KUDEL-DOBROGOWSKA,K., Inwarianty w przebiegach parametru F_0 [Invariants of the F_0 parameter]. BPTJ 32, 1974,159-171.

40. JASSEM,W., SZYBISTA,D., Subiektywne prawdopodobieństwo wyrazów polskich [The subjective probability of Polish words]. Warszawa 1975 (mimeo).

41. JURKOWSKI,M., Częstotliwość form stopnia wyższego i najwyższego przymiotników i przysłówków w polskich tekstach publicystycznych [The frequency of comparative and superlative forms of adjectives and adverbs in Polish journalistic texts]. PJ 1973/8,465-470.

42. KEMPF,Z., Problem częstości językowej [The problem of frequency in language]. ZN WSP Op 4,1966,49-60.

43. KLONECKI,W., On identifiability of mixtures of composed Poisson distributions. ZM 12/3,1971,259-271.

44. KNIAGININOWA,M., Próba zastosowania metod statystycznych w badaniach stylistyczno-składniowych [An attempt to apply statistical methods in stylistic-syntactical analyses]. JP 42/2,1962,92-116.

45. KNIAGININOWA,M., PISAREK,W., Język wiadomości prasowych [The language of newspaper reports]. Kraków 1966,170 p.

46. KONDRATOV,A., Czterostopowy jamb N.Zabołockiego i niektóre zagadnienia statystyki wiersza [The four-foot iambus of N.Zabołocki and some problems of verse statistics]. In: Poetyka i matematyka. Warszawa:Państwowy Instytut Wydawniczy, 1965, 97-111.

47. KOPCZYŃSKA,A., MAYENOWA,M.R., O niektórych cechach struktury głoskowej tekstów wierszowanych [On some features of the sound structure of versified texts]. In: Prace z poetyki. Wrocław: Ossolineum, 1968,9-17.

48. KOPCZYŃSKA,Z., PSZCZOŁOWSKA,L., Użyteczność metod statystycznych w leksykografii, stylistice i wersyfikacji [The applicability of statistical methods in lexicography, stylistics, and versification]. Spr PAN 3/1,1959,104-107.

49. KOPCZYŃSKA,Z., PSZCZOŁOWSKA,L., O dwóch typach ośmiozgłoskowca [On two types of octonary lines]. PL 58/4,1967, 459-462.

50. KOPCZYŃSKA,Z., PSZCZOŁOWSKA,L., Z zagadnień struktury językowej polskiego sylabowca [On problems of the linguistic structure of Polish syllabic versification]. PL 59/2,1968,183-193.

51. KOSIEL,U., Statystyczne badanie cech osobniczych głosu w średnim rozkładzie energii w widmie mowy polskiej [A statistical study of individual features of the voice in the mean energy distribution in the spectrum of Polish]. Archiwum Akustyki 5/2,1970,205-229; 7/2,1972, 85-104.

52. KRĄŻYŃSKA,Z., Opis statystyczny kategorii gramatycznej osoby w polskich tekstach drukowanych XVII wieku [A statistical description of the grammatical category of person in printed Polish seventeen-century texts]. StP 3, 1976,57-68.

53. KUDELA,K., Synteza mowy i jej zastosowanie językoznawcze [Speech synthesis and its application in linguistics]. BPTJ 27,1969,199-215.

54. KULKA,B., Możliwości zastosowania metod statystycznych w nauczaniu języka polskiego [Possibilities for the application of statistical methods to the teaching of Polish]. PJ 1977/4, 173-182.

55. KURASZKIEWICZ,W., Statystyczne badanie słownictwa polskich tekstów XVI w. [A statistical study of the vocabulary of Polish texts of the sixteenth century]. In: Z polskich studiów slawistycznych. Prace językoznawcze i etnograficzne, Vol.1. Warszawa: Państwowy Instytut Wydawniczy, 1958, 240-257.

56. KURASZKIEWICZ,W., Szkice o języku Mikołaja Reja [Notes on the language of Mikołaj Rej]. In: Odrodzenie v Polsce, Vol.III/1. Warszawa: Państwowy Instytut Wydawniczy, 1960,170-180.

57. KURASZKIEWICZ,W., Étude de la paternité des textes anonymes d'après la méthode de la statistique linguistique. In: Poetics, Poetyka, Poètika, Vol I. Warszawa: Państwowe Wydawnictwo Naukowe, 1962,625-633.

58. KURASZKIEWICZ,W., La richesse du vocabulaire dans quelques grands textes polonais en vers. Wrocław: Ossolineum, 1963, 22 p.

59. KURASZKIEWICZ,W., Opracowanie językowe "Postylli" Reja [A linguistic study of Rej's "Postylla"]. In: Rej,M., Postylla, Vol I. Wrocław: Ossolineum, 1965.

60. KURASZKIEWICZ,W., Uwagi o statystyce w "Słowniku" [Remarks on the statistics of the "Słownik"]. In: Słownik polszczyzny XVI w., Vol 1. Wrocław: Ossolineum, 1966,XIV-XXV.

61. KURASZKIEWICZ,W., Obfitość słownictwa w kilku dużych tekstach polskich [The richness of vocabulary in some long Polish texts]. StP 1,1973,45-63.

62. KURASZKIEWICZ,W., Częstotliwość wyrazów w "Panu Tadeuszu" Adama Mickiewicza i w "Wizerunku" Mikołaja Reja [Word frequency in Adam Mickiewicz's "Pan Tadeusz" and in Mikołaj Rej's "Wizerunek"]. StP 2,1975,71-95.

63. KURASZKIEWICZ,W., ŁUKASZEWICZ,J., Ilość różnych wyrazów

w zależnosci od długości tekstu [The frequency of different words as a function of text length]. PL 42/1,1951, 168-182.

64. KURCZ,I., Założenia ogólne frekwencyjnych badań nad słownictwem i cele, którym te badania służą[The general basis of frequency analysis of vocabulary and its aims]. Przegląd pedagogiczny 1973/1-2,65-75.

65. KURCZ,I., LEWICKI,A., SAMBOR,J., WORONCZAK,J., Słownictwo współczesnego języka polskiego. Listy frekwencyjne [The vocabulary of present-day Polish. Frequency lists].
Vol.1, Parts 1-2:Teksty popularnonaukowe [Popularizing scientific texts].
Vol.2,Parts 1-2:Drobne wiadomości prasowe [Short press reports].
Vol.3,Parts 1-2:Publicystyka [Journalistic prose].
Vol.4,Parts 1-3:Proza artystyczna [Novelistic prose].
Vol.5,Parts 1-2: Dramat artystyczny [Drama].
Warszawa: Polska Akademia Nauk, 1974-1977.

66. LEWICKA,H., Blaski i cienie językoznawstwa stosowanego [Successes and failures of applied linguistics]. KNf 12/2,1965,175-181.

67. LEWICKI,A., MASŁOWSKI,W., SAMBOR,J., WORONCZAK,J., Słownictwo współczesnej publicystiki polskiej [The vocabulary of present-day Polish journalistic prose]. Warszawa: Polska Akademia Nauk, 1971, 583 p.
Revied by: Kaspin,I.W.:JP 55/1,1975,49-51;
Kowalikówna,K.: JP 54/1,1974,53-60;
Mańczak,W.: RSl 35,1974,68-71;
Zagrodnikowa,A.: Zeszyty prasoznawcze 14/3, 1973,121-122;
Zarębina,M.: JP 53/4,1973,300-302.

68. LEWICKI,A., SAMBOR,J., Projekt słownika frekwencyjnego współczesnego języka polskiego [The present-day Polish frequency dictionary project]. Spr PAN 12/4,1969,90-103.

69. LEVÝ,J., W sprawie ścisłych metod analizy wiersza [Remarks on exact methods of verse analysis]. In: Poetyka i matematyka. Warszawa:Państwowy Instytut Wydawniczy,1965, 23-71.

70. LOTKO,E., Semantika a frekvence záporných lexikálních jednotek v současné češtině [The semantics and frequency of negative lexical units in modern Czech]. RSl 30,1969, 27-37.

71. LUBAŚ,W., Rym Jana Kochanowskiego. Próba lingwistycznej charakterystyki i oceny [The rhyme of Jan Kochanowski. An attempt at a linguistic characterization and evaluation]. Katowice 1975, 176 p.

72. LUTOSŁAWSKI,W., Sur une nouvelle méthode pour déterminer la chronologie des dialogues de Platon. Paris 1896.

73. ŁOBACZ,P., Entropia oraz parametry akustyczne jako kryteria interpretacji fonematycznej [Entropy and acoustical parameters as criteria for phonemic interpretation]. BPTJ 29,1971,77-93.

74. ŁOBACZ,P., JASSEM,W., Fonotaktyczna analiza mówionego tekstu polskiego [A phonotactic analysis of a spoken Polish text]. BPTJ 32,1974,179-197.

75. MALMBERG,B., Nowe drogi w językoznawstwie. Przgląd szkół i metod [Translation of:Nya vägar inom språkforskningen. En orientering i modern lingvistik (New directions in linguistics;a survey of schools and methods)]. Warszawa 1969, 383 p.
Reviewed by:Dołęga,J.M.: Studia Philosophiae Christianae 7/1,1971,253-255;
Lewicki,S.: Nowe Książki 14,1970,874;
Kloska,G.:Lud 54,1970,205-207;
Zgółka,T.:Studia Metodologiczne 1971/8,137--144;
Zgółka,T.:Nurt 1970/2,54-56.

76. MAŃCZAK,W., Pojęcie ilości w języku [The notion of quantity in language]. Studia Filozoficzne 1959/6,111-127.

77. MAŃCZAK,W., Fréquence d'emploi des occlusives labiales, dentales et vélaires. Bulletin de la Société de linguistique de Paris 54,1959,208-214.

78. MAŃCZAK,W., Polska fonetyka i morfologia historyczna [Polish historical phonetics and morphology]. Warszawa: Państwowy Instytut Wydawniczy, 1965, 170 p.

79. MAŃCZAK,W., Fréquence et analyse linguistique. Statistique et analyse linguistique (Colloque de Strasbourg 20-22 avril 1964). Paris 1966,99-103.

80. MAŃCZAK,W., La nature du supplétivisme. Linguistics 28,1966, 82-89.

81. MAŃCZAK,W., A propos de "l'analogie progressive". Linguistics 33,1967,82-86.

82. MAŃCZAK,W., Le developpement phonétique irrégulier dû à la fréquence en russe. Lingua 21,1968,287-293.

83. MAŃCZAK,W., 16-wieczne wyrażenia typu "pisał jest, pisali są" [Sixteenth century expressions of the type "pisał jest, pisali są"]. JP 48/4,1968,298-301.

84. MAŃCZAK,W., Le developpement phonétique des langues romanes et la fréquence. Kraków 1969, 98 p.

85. MAŃCZAK,W., Z zagadnień językoznawstwa ogólnego [Problems of general linguistics]. Wrocław: Ossolineum, 1970,321 p.

86. MAŃCZAK,W., Nieregularność w formach wyrazu "wszystek" [The irregularity of the forms of the word "wszystek"]. International Journal of Slavic Linguistics and Poetics 13,1970,143-154.

87. MAŃCZAK,W., Sur la théorie de catégories "marquées" et "non marquées" de Greenberg. Linguistics 59,1970,29-36.

88. MAŃCZAK,W., Słowiańska fonetyka historyczna a frekwencja [Slavonic historical phonetics and frequency]. Kraków 1977, 338 p.
Revied by: Kucała,M.: JP 57/5,377-380.

89. MAY,S., O pewnych właściwościach statystycznych języka polskiego [On certain statistical characteristics of the Polish language]. In: Jagłom,A.M., Jagłom,J.M., Prawdopodobieństwo i informacja. Warszawa: Książka i Wiedza, 1963,368-372.

90. MAYENOWA,M.R., Lingwistyka matematyczna w ZSRR [Mathematical linguistics in the Soviet Union]. Nowa Kultura 1959/24,3-7.

91. MAYENOWA,M.R., Możliwości i niebezpieczeństwa metod matematycznych w poetyce [Possibilities and dangers of mathematical metods in poetics].In: Poetyka i matematyka. Warszawa: Państwowy Instytut Wydawniczy,1965,5-20.

92. MAYENOWA,M.R., Granice matematyzacji (w opisie wiersza) [The limits of mathematization (in verse description)]. Kultura i Społeczeństwo 1965/24,229-233.

93. MAYENOWA,M.R., Accent pattern in Polish syllabic verse. In: Poetics, Poetyka, Poėtika, Vol.II. The Hague-Paris-Warszawa: Państwowe Wydawnictwo Naukowe, 1966,371-374.

94. MAYENOWA,M.R., O matematyzacji lingwistyki [On the mathematization of linguistics]. Podstawowe Problemy Współczesnej Techniki 12,1967,135-162.

95. MEINHOLD,G., Die Grundordnung sprachrhythmischer Bewegung. BFon 12,1971,57-78.

96. MIKOŁAJCZYK,S., Składnia "Quo vadis" H.Sienkiewicza w ujęciu statystycznym (dla potrzeb stylistiki) [The syntax of H.Sienkiewicz's"Quo vadis" from the statistical point of view (for stylistic purposes)]. JP 50/2,1970, 85-97.

97. MILEJKOWSKA,H., Analiza statystyczna spółgłosek w tekście rosyjskim i polskim [Statistical analysis of consonants in a Russian and a Polish text]. GZHum 13, Filologija Rosyjska 3,1970,85-98.

98. MILEWSKI,T., Zachodnia granica pomorskiego obszaru językowego w wiekach średnich [The western border of the Pomeranian language area in the Middle Ages]. SO 10,1931, 124-152.

99. MILEWSKI,T., Językoznawstwo matematyczne [Mathematical linguistics]. Ruch Filozoficzny 21/2,1962,187-188.

100. MILEWSKI,T., On the validity of glottochronology. Current Anthropology 3/2,1962,142.

101. MILEWSKI,T., Założenia językoznawstwa typologicznego [Fun-

damentals of linguistic typology]. BPTJ 21,1962,3-39.

102. MORAWSKA,L., Stylistyka statystyczna (Na marginesie prac P.Guirauda)[Statistical stylistics (from the perspective of the works of P.Guiraud)]. KNf 6/1,1959,51-59.

103. MOSZYŃSKI,L., Język Kodeksu Zografskiego, cz.1. [The language of the Codex Zographensis, Part 1]. Wrocław: Ossolineum, 1975, 284 p.

104. NAGÓRKO,A., Stan i perspektywy badań ilościowych w słowotwórstwie opisowym [The status and future of frequency studies in the description of word-formation]. PJ 1974, 1-13.

105. NAGÓRKO-KUFEL,A., Z badań nad częstościami elementów tekstowych języka polskiego dla potrzeb pisma niewidomych [Frequency studies of Polish text elements for use in Braille]. PJ 1975/1,7-13.

106. OSTROWSKA,E., O współczesnej poezji językoznawczo [Modern poetry from a linguistic point of view]. JP 51/4,1971, 244-248.

107. OŻDŻYŃSKI,J., Cechy kategorialne stylu prasowych wiadomości sportowych [Categorial features of the style of newspaper sport articles]. JP 57/5,1977,363-375.

108. PISAREK,W., Rzeczowniki w reportażach i wiadomościach prasowych [Nouns in press reports]. JP 49/1,1969,37-43.

109. PISAREK,W., Mierzenie stylistycznej i emotywnej wartości słownictwa [Measurement of the stylistic and emotive value of vocabulary]. ZP 12/2,1971,37-46.

110. PISAREK,W., Frekwencja wyrazów w prasie. Wiadomości - komentarze - reportaże [Word frequency in the Polish press. News - Commentaries - Reports]. Kraków 1972, 309 p.
Reviewed by: Kowalikówna,K.: JP 54/1,1974,53-60;
Poniatowski,Z.: Euhemer 17/2,1973,85-87;
Szymczak,M.: ZP 14/2,1973,95-97;
Trot,R.:Prasa Polska 26/10,1972,43-44;
Zerębina,M.:JP 53/4,1973,302-304.

111. PONIATOWSKI,Z., Z badań statystycznych nad Nowym Testamentem [On statistical studies of the New Testament]. Euhemer 1965/4,13-27.

112. PONIATOWSKI,Z., Analiza statystyczna ewangelicznego opisu Pasji Jezusa [A statistical analysis of the description of Christ's Passion in the Gospels]. Studia Religioznawcze 1970/3,71-79.

113. PONIATOWSKI,Z., Nowy Testament w świetle statystyki językowej [The New Testament in the light of statistical linguistics]. Wrocław: Ossolineum, 1971, 202 p.

114. PRZYBYCIN,A., Próba statystycznej analizy szyku zdań [An attempt at a statistical analysis of sentence order]. Prace Śl 1,1969,77-90.

115. PSZCZOŁOWSKA,L., Wiersz - styl - dialog. Wokół dwóch redakcji Warszawianki [Verse - style - dialogue. On two editions of the Warszawianka]. PL 56/4,1965,439-464.

116. PSZCZOŁOWSKA,L., Długość wiersza a budowa zdania [Verse length and sentence structure]. In: Poetyka i matematyka. Warszawa: Państwowy Instytut Wydawniczy, 1965, 79-96.

117. PSZCZOŁOWSKA,L. La divisibilité du metre et substrat linguistique. In: To Honor Roman Jakobson, Vol.II. The Hague:Mouton, 1967,1624-1633.

118. PSZCZOŁOWSKA,L., Pri prameňoch istej tendencie pravidelného verša [At the source of a certain tendency in regular verse]. Slovenská literatúra 1967/4,364-371.

119. PSZCZOŁOWSKA,L., Some problems of the structure of assonance in folk and art poetry. In: Teorie verše II. Brno: Universita J.E.Purkyně, 1968,83-88.

120. PSZCZOŁOWSKA,L., Z problematyki związków pomiędzy strukturą dźwiękową a strukturą klauzuli wierszowej [On the problem of the relation between sound structure and verse clause]. In: Prace z poetyki. Wrocław: Ossolineum 1968,18-25.

121. RADZIMOWSKA,K., Stylistyka w ujęciu statystycznym [Stylis-

tics from the statistical point of view]. PJ 1970/1, 29-35.

122. ROCŁAWSKI,B., Polski system fonostatystyczny [The Polish phonostatistical system]. GZHum 1969/4,99-104.

123. ROCŁAWSKI,B., Ze studiów fonostatystycznych nad kaszubszczyzną. Rozkład częstości występowania fonemów [Phonostatistical studies on the Kashubian language. Frequency distribution of phoneme occurrences]. Gdańskie Studia Językoznawcze 1975,107-130.

124. ROCŁAWSKI,B., Ze studiów fonostatystycznych nad kaszubszczyzną. Średnia długość fraz, wyrazów i sylab. Rozkład częstości wyrazów x-fonemowych [Phonostatistical studies on the Kashubian language. Average length of phrases, words, and syllables. Frequency distribution of words consisting of x phonemes]. ZNUG 1975/3,65-83.

125. ROCŁAWSKI,B., Zarys fonologii, fonetyki, fonotaktyki i fonostatystyki współczesnego języka polskiego [A compendium of the phonology, phonetics, phonotactics and phonostatistics of modern Polish]. Gdańsk 1976, 202 p.

126. ROZWADOWSKI,J., O pewnym prawie ilościowym rozwoju języka [On a certain frequency law of language development]. In: Rozwadowski,J., Wybór pism, Vol.3. Warszawa 1960, 96-105.

127. RYBICKA,H., Frekwencja form komparatiwu i superlatiwu przymiotników i przysłówków w polskiej prozie biblijnej XIV wieku [The frequency of comparatives and superlatives of adjectives and adverbs in Polish sixteenth century Bible prose]. In: Zagadnienia kategorii stopnia w językach słowiańskich. Warszawa:Wydawnictwo Uniwersytetu Warszawskiego, 1976,125-131.

128. SADALSKA,G., Vergleich der initialen Konsonantenphoneme im Schwedischen und Polnischen. BFon 16,1975,63-78.

129. SAFAREWICZ,J., Krytyka metody ilościowej stosowanej w ocenie pokrewieństwa językowego [A critique of the quantitative method with regard to its application to the determination of language relationship].BPTJ 8, 1948,30-40.

130. SALONI,Z., Review of:Gladkij,A.V., Mel'čuk,I.A., Elementy matematičeskoj lingvistiki. Moskva 1969. BPTJ 29, 1971,197-202.

131. SAMBOR,J., Review of: Frumkina,R.M., Statističeskie metody izučenija leksiki. Moskva 1964. PL 56/4,1965,598-603.

132. SAMBOR,J., O statystycznej strukturze słownictwa tekstów języków naturalnych (na materiale "Pana Tadeusza" A. Mickiewicza i Nowego Testamentu) [On the statistical structure of the vocabulary of texts in natural languages (based on A.Mickiewicz's "Pan Tadeusz" and the New Testament)]. Spr PAN 9/3,1966,38-47.

133. SAMBOR,J., Badania statystyczne nad słownictwem. Na materiale "Pana Tadeusza" [Statistical investigations of vocabulary, based on "Pan Tadeusz"]. Wrocław: Ossolineum, 1969, 163 p.
Reviewed by:Pisarkowa,K.: JP 51/2,1971,147-150.

134. SAMBOR,J., Analiza stosunku "Type-Token" czyli objętości słownictwa /W/ i długości tekstu /N/. Na przykładzie "Pana Tadeusza" [Analysis of the "type-token" relation i.e. the relation between the size of the vocabulary and the length of the text, exemplified on "Pan Tadeusz"]. PF 20,1970,65-70.

135. SAMBOR,J., The distribution of low frequency words in texts of natural languages. BPTJ 28,1971,125-133.

136. SAMBOR,J., Z zagadnień gramatyki w słowniku frekwencyjnym współczesnego języka polskiego [On the problems of grammar in the frequency dictionary of modern Polish]. BPTJ 29,1971,117-129.

137. SAMBOR,J., Słowa i liczby. Zgadnienia językoznawstwa statystycznego [Words and numbers. Problems of statistical linguistics]. Wrocław: Ossolineum, 1972, 305 p.
Reviewed by: Lehfeldt,W.: Linguistics 132,1974,105-109;
Maćkowiak-Witkowska,M.: BFon 16,1975,128-131;
Mańczak,W.: LP 18,1974,123-126;
Nagórko,A.: PJ 1973/3,157-159.

138. SAMBOR,J., Z zagadnień gramatyki słownictwa rzadkiego [On the problems of a grammar of rare vocabulary]. BPTJ 31,1973,47-58.

139. SAMBOR,J., Słownictwo bardzo częste w pięciu stylach współczesnej polszczyzny pisanej [Very frequent vocabulary in five styles of modern written Polish]. PJ 1974/9,466-475;1974/10,533-537.

140. SAMBOR,J., Z zagadnień semantyki derywatów rzadkich [Problems of semantics of low frequency derivations]. In: Tekst i język. Problemy semantyczne. Wrocław: Ossolineum, 1974,271-298.

141. SAMBOR,J., Review of: Altmann,G., Lehfeldt,W., Allegemeine Sprachtypologie. Prinzipien und Messverfahren.München 1973. BPTJ 33,1975,197-203.

142. SAMBOR,J., O słownictwie statystycznie rzadkim. Na materiale derywatów we współczesnej publicystyce polskiej [On statistically rare vocabulary. Based on derivations in modern Polish journalistic prose]. Warszawa: Państwowe Wydawnictwo Naukowe, 1975, 113 p.

143. SATKIEWICZ,H., Kryterium ilościowe jako wskaźnik produktywności formantów słowotwórczych [The quantitative criterion as an index of the productivity of word-building formants]. PF 19,1969,109-118.

144. SIGURD,B., Struktura języka. Zgadnienia i metody językoznawstwa współczesnego (Translation of: Språkstruktur. Den moderna språkforskningens metoder och problemställningar. Translator: Z.Wawrzyniak). Warszawa: Państwowe Wydawnictwo Naukowe, 1975, 179 p.

145. SKALMOWSKI,W., Statystyczny model komunikacji językowej i kilka parametrów języka nowoperskiego [A statistical model of speech communication and some parameters of New Persian]. Sprawozdania Kom. Orient. PAN (Oddział w Krakowie) 1960,278-280.

146. SKALMOWSKI,W., Ein Beitrag zur Statistik der arabischen Lehnwörter im Neupersischen. FO 3,1961,171-175.

147. SKALMOWSKI,W., Polskie przekłady Hafiza w świetle prawa

Zipfa-Mandelbrota [Polish translations of Hafiz in the light of the Zipf-Mandelbrot law]. Sprawozdania Kom. Orient. PAN (Oddział w Krakowie) 1961,125-127.

148. SKALMOWSKI,W., Uber einige statistisch erfassbare Züge der persischen Sprachentwicklung. FO 4,1962,47-80.

149. SKALMOWSKI,W., Review of: Axmanova,O.S., Mel'čuk,I.A., Frumkina,R.M., Padučeva,E.V., O točnyx metodax issledovanija jazyka. Moskva 1961. BPTJ 21,1962,193-195.

150. SKALMOWSKI,W., A note on distribution of Arabic verbal roots. FO 6,1964,97-100.

151. SKALMOWSKI,W., Probleme der Gesetzmässigkeit in der mathematischen Linguistik. BPTJ 23,1965,23-33.

152. SŁAWIŃSKI,F.F., Obliczenia wyrazów zawartych w trzech słownikach: 1) Lindego, 2) w Wileńskim i 3) Rykaczewskiego [Word-counts in three dictionaries: 1) Linde, 2) Wileński, and 3) Rykaczewski]. Warszawa: Drukarnia Józefa Sikorskiego, 1873, 36 p.

153. SMÓŁKOWA,T., Słownictwo i fleksja "Lalki" Bolesława Prusa [Vocabulary and inflection in B. Prus' "Lalka"]. Wrocław: Ossolineum, 1974, 191 p.
Reviewed by: Safarewicz,J.: JP 55/2,1975,138-140.

154. SMÓŁKOWA,T., Wyrazy najczęstsze w prasie współczesnej i w Lalce B.Prusa [The most frequent words in modern journalistic prose and in B.Prus' "Lalka"]. PF 25, 1974,257-268.

155. SOBIERAJSKI,Z., Identyfikacja fonemów w percepcji mowy [Identification of phonemes in speech perception]. BFon 8,1967,105-123.

156. Sortowanie tekstów przy pomocy maszyny cyfrowej "Cyber 72" dla celów badań statystycznych, słownikowych i fleksyjnych. Oprac. L.Niemiec i in. [Text sorting with the computer "Cyber 72" for the purposes of the statistical study of vocabulary and inflection]. Rocznik Naukowo-Dydaktyczny Wyższej Szkoły Pedagogicznej w Krakowie. Prace Językoznawcze 1976/3,197-209.

157. STEFFEN,M., Częstość występowania głosek polskich [The frequency of occurrence of the sound of Polish]. BPTJ 16,1957,145-164.

158. STEFFEN-BATOGOWA,M., The problem of automatic phonemic transcription of written Polish. BFon 14,1973,75-86.

159. STEFFEN-BATOGOWA,M., Automatyzacja transkripcji fonetycznej tekstów polskich [The automatization of the phonetic transcription of Polish texts]. Warszawa: Państwowe Wydawnictwo Naukowe, 1975, 139 p.

160. SZCZODROWSKI,M., Językoznawstwo stosowane a lingwostatystyka [Applied linguistics and statistical linguistics]. JOS 1976/5,259-263.

161. ŚWIECZKOWSKI,W., Metoda ilościowa w badaniach syntaktycznych. Studium szyku słow oparte na materiale średnioangielskim [The quantitative method in syntactic investigations. A study of word-order based on Middle English material]. PL 51/3,1960,109-122.

162. TARNACKI,J., Studia porównawcze nad geografią wyrazów (Polesie - Mazowsze) [Comparative studies of word geography]. Warszawa 1939, 101 p.

163. TEŠITELOVÁ,M., O badaniu statystycznym zasobu słownikowego w języku czeskim [On the statistical investigation of Czech vocabulary]. PJ 1967/9, 401-408.

164. TOKARSKI,J., Dynamizm procesów językowych i metody jego badania [The dynamics of language processes and the methods of their investigation]. PJ 1952/8,1-16.

165. TOKARSKI,J., Metody ilościowe w językoznawstwie wobec nowych perspektyw [New perspectives in quantitative methods in linguistics]. PJ 1961/6,241-253.

166. TUORA,M.I., La fréquence et l'étymologie dans le vocabulaire roumain. KNf 20,1973,172-181.

167. WIERZBICKA,A., Hipotaksa i konstrukcje nominalne w rozwoju polszczyzny [Hypotaxis and nominal constructions in the development of Polish]. PL 53/1,1962,195-216.

168. WIERZBICKA,A., Z zagadnień renesansowej sztuki prozy.

Budowa zakończeń zdań [Problems of Renaissance prose. The structure of sentence endings]. In: Poetyka i matematyka. Warszawa: Państwowy Instytut Wydawniczy,1965, 115-144.

169. WIERZBICKA,A., K voprosu o porjadke slov v polskom i russkomkom stixe [On the problem of word-order in Polish and Russian verse]. In: Poetics, Poetyka, Poètika, Vol. II. The Hague-Paris-Warszawa: Państwowe Wydawnictwo Naukowe, 1966,345-369.

170. WIERZBICKA,A., System składniowo-stylistyczny prozy polskiego renesansu [The syntactic-stylistic system of Polish Renaissance prose]. Warszawa: Państwowy Instytut Wydawniczy, 1966, 278 p.

171. WŁODARCZYK,H., Obiektywna klasyfikacja fonemów w operciu o parametry skrajnych pasm widma mowy [Objective classification of phonemes based on the extrene bands of the speech spectrum]. Archiwum Akustyki 4,1969,263-279.

172. WŁODARCZYK,H., Podstawowe parametry fizyczne mowy w zastosowaniu do automaticznego rozpoznawania głosek i fonemów [The basic physical parameters of speech and their application to the automatic recognition of sounds and phonemes]. Warszawa 1971, 124 p.

173. WORONCZAK,J., Z badań nad wierszem Biernata z Lublina [Studies on the verse of Biernat of Lublin]. PL 49/3, 1958,97-118.

174. WORONCZAK,J., Statistische Methoden in der Verslehre. In: Poetics, Poetyka, Poètika, Vol.I. Warszawa: Państwowe Wydawnictwo Naukowe, 1962, 604-624.

175. WORONCZAK,J., Rytmika akcentowa sylabowca [The accentual rhythm of syllabic versification]. In: Poetyka i matematyka. Warszawa: Państwowy Instytut Wydawniczy,1965, 72-78.

176. WORONCZAK,J., Metody obliczania wskaźników bogactwa słownikowego tekstów [Methdos of calculating indices of the lexical richness of texts]. In: Poetyka i matematyka. Warszawa:Państwowy Instytut Wydawniczy,1965,145-165.

177. WORONCZAK,J., On an attempt to generalize Mandelbrot's distribution. In: To Honor Roman Jakobson, Vol.II. The Hague: Mouton, 1967,2254-2268.

178. WORONCZAK,J., O statystycznym określeniu spójności tekstu [On the statistical determination of text coherence]. In: Semantyka tekstu i języka. Wrocław: Ossolineum, 1976,165-173.

179. ZAGRODNIKOWA,A., Występowanie w prasie wyrazów autor - twórca - sprawca [The occurrence of the words autor - twórca - sprawca in newspaper prose]. ZP 12/1,1971, 56-60.

180. ZARĘBINA,M., Rola wyrazów w słowniku i w tekście [The rôle of words in the lexicon and in texts]. JP 50/1, 1970,33-46.

181. ZARĘBINA,M., Najczęstsze wyrazy polszczyzny mówionej [The most frequent words of spoken Polish]. JP 51/5,1971, 336-347.

182. ZARĘBINA,M., Statystyczna struktura wyrazowa tekstu gwarowego (na przykładzie wsi Ząb w powiecie nowotarskim) [Tha statistical structure of words in a vernacular text (based on the language of the village of Ząb/Nowotarsk)]. JP 53/1,1973,5-16.

183. ZARĘBINA,M., A statistical lexicalsemantic characterization of the main varieties of present-day Polish. LP 17,1973,37-48.

184. ZARĘBINA,M., Niektóre odmiany polszczyzny w ujęciu statystycznym [Certain Polish inflectional changes from the point of view of statistics]. Sprawozdania z Posiedzeń Komisji Oddziału Krakowskiego Polskiej Akademii Nauk 16/1,1973,43-44.

185. ZDANIUKIEWICZ,A.A., Procesy integracji językowej na Ziemiach Zachodnich i Północnych w świetle analizy ilościowej [Processes of linguistic integration in the Western and Northern regions in the light of quantitative analysis]. PF 23,1972,281-288.

186. ZEMBATY-MICHALAKOWA,M., Problem badania stylu metodą statystyczną na przykładzie Marii A.Malczewskiego i Wacława J.Słowackiego [The problem of style investigation by means of statistical methods based on A.Malczewski's "Maria" and J.Słowacki's "Wacław"]. Prace Literackie XIV.Acta Universitatis Vratislaviensis 177, Wrocław 1972,177-202.

187. ZEMBATY-MICHALAKOWA,M., Poezja Juliana Przybosia w świetle badań statystyczno-językowych [The poetry of J.P. in the light of quantitative linguistic studies]. Prace Literackie XVII. Acta Universitatis Vratislaviensis 298, Wrocław 1975, 155-183.

188. ZUBRZYCKA,L., O dostosowaniu klawiatury maszyny do pisania do struktury języka [On the adaptation of the typewriter keyboard to language structure]. ZM 6/4,1963, 419-439.

189. ZUBRZYCKA,L., O klawiaturze maszyny do pisania dostosowanej do języka polskiego [On the typewriter keyboard adapted to Polish]. ZM 8/2,1965,103-116.

190. ZUBRZYCKI,S., Concerning Yule's characteristic of style. ZM 4/4,1959,328-331.

191. ZUBRZYCKI,S., O niektórych pracach seminarium z zastosowań matematyki [On some seminar papers on the application of mathematics]. ZM 8/3,1966,265-281.

CURRENT BIBLIOGRAPHY

ABBREVIATIONS

NTI Naučno-texničeskaja informacija. Serija 2: Informacionnye processy i sistemy

PSML Prague Studies in Mathematical Linguistics

GENERAL

1. ARAPOV, M.V.: Dve modeli rangovogo raspredelenija [Two models of the rank distribution]. Voprosy informacionnoj teorii i praktiki 4, 1977, 3-42.

2. ARAPOV, M.V., ŠREJDER, Ju.A.: Klassifikacii i rangovye raspredelenija [Classifications and rank distributions]. NTI 1977/11-12, 15-21.

3. DAVID, J., MARTIN, R. (Eds.): Etudes de statistique linguistique. Paris: Klincksieck, 1977, 119 pp.

4. DOLPHIN, C.: Evaluation probabiliste des co-occurrences. In [3], 21-34.

5. KÖNIGOVÁ, M.: The scaling technique applied to text description. PSML 5, 1977, 211-221.

6. KRÁLÍK, J.: An application of exponential distribution law in quantitative linguistics. PSML 5, 1977, 223-235.

7. MIZUTANI, Sh.: Numerical tables for a non-parametric test of trends [In Japanese]. Mathematical Linguistics 11, 1977, 80-84.

8. MOSKOVICH, W.: Perspective paper: Quantitative linguistics. Walker, D., Karlgren, H., Kay, M. (Eds.): Natural Language Information Science. Stockholm: Scriptor, 1977, 57-74.

9. MULLER, Ch.: Observation, prévision et modèles statistiques. In [3], 9-19.

10. PIOTROVSKIJ, R.G.: Inženernaja lingvistika: teorija - èksperiment - realizacija [Engineering linguistics: theory - experiment - implementation]. IzvAN 37, 1978, 10-19.

11. PIOTROVSKIJ, R.G., BEKTAEV, K.B., PIOTROVSKAJA, A.A.: Matematičeskaja linguistika [Mathematical Linguistics]. Moskva: Vysšaja skola, 1977, 383 pp.

BIBLIOGRAPHY

12. CHARPENTIER, C.-M.: La statistique linguistique pratiquée dans les pays anglo-saxons. Eléments de bibliographie (1970-1976). In [3], 89-119.

PHONOLOGY

13. ALMEIDA, A., FLEISCHMANN, G., HEIKE, G., THÜRMANN, E.: Short time statistics of the fundamental tone in verbal utterances under psychic stress. BIPK 8, 1977, 67-77.

14. BOY, J.: Dechiffrierungsalgorithmen zur phonetischen Identifikation von Buchstaben. Bochum: Brockmeyer, 1977, 148 pp.

15. GRASSEGGER, H.: Merkmalsredundanz und Sprachverständlichkeit. Hamburg: Buske, 1977, 167 pp.

16. HAARMANN, H.: Prinzipielle Probleme des multilateralen Sprachvergleichs. Tübingen: Narr, 1977, 128 pp.

17. HAGA, J.: Estimating the rate of occurrence of Chinese characters in a book [In Japanese]. Mathematical Linguistics 11, 1977, 85-87.

18. KEMPGEN, S.: Syntagmatische Phonemtypologie. Phonetica 34, 1977, 108-131.

19. LEKOMCEVA, M.I.: K tipologičeskoj xarakteristike fonologičeskix sistem jazykov Balkanskogo poluostrova i Sredizemnomor'ja [On the typological characterization of the phonological systems of the Balkan and the Mediterrenean languages]. Balkanskij lingvističeskij sbornik. Moskva: Nauka, 1977, 243-274.

20. MÜLLER, B.S.: Buchstabenreichtum. Vorschläge für einige Maße und die Ergebnisse von deren Anwendung bei der Analyse einer Sammlung von Schreibfehlern. Entwicklungsunterlage TUKAN 14, 1978.

21. NOMURA, M.: Statistical analysis of writing-form variation of Japanese words [In Japanese]. Mathematical Linguistics 11, 1977, 3-19.

22. STELLMACHER, I.: Auswertung der Fehlerstatistik. Entwick-

lungsunterlage TUKAN 2, 1977.

23. STELLMACHER, I.: Auswertung einer mittelgroßen Fehlersammlung. Entwicklungsunterlage TUKAN 9, 1978.

24. TULDAVA, Ju.: Kvantitativnoe issledovanie struktury odnosložnogo slova v ėstonskom jazyke [A quantitative investigation of the structure of the Estonian one-syllable word]. Linguistica 10, 1978, 115-135.

25. YOKOYAMA, Sh., ITABASHI, Sh.: On the distance of Japanese words based on distinctive features and the second-order model [In Japanese]. Mathematical Linguistics 11, 1977, 165-177.

G R A M M A R

26. ANDRUKOVIČ, P.F., KOROLEV, Ė.I.: O statističeskix i leksiko-grammatičeskix svojstvax slov [Statistical and lexico-grammatical properties of words]. NTI 1977/4, 1-9.

27. DUŠKOVÁ, L.: On some differences in the use of the perfect and the preterite in British and American English. PSML 5, 1977, 53-68.

28. GINZBURG, E.L.: K razrabotke kriteriev napravlenija proizvodnosti [On the problem of defining criteria of the direction of derivation]. Grammatika i norma. Moskva: Nauka, 1977, 83-91.

29. LEJOSNE, J.C.: Contribution à l'étude de l'opposition "verbes forts - verbes faibles" dans quelques dialectes germaniques. In 3 , 55-87.

30. PIOTROVSKAJA, A.A.: Avtomatičeskoe privedenie imennyx slovoupotreblenij k kanoničeskoj forme [Automatic reduction of nouns in text to a canonical form]. NTI 1977/1, 32-36.

S E M A N T I C S

31. ARSENT'EVA, N.G.: Ot semantičeskix svojstv slova k poverxnostno-sintaksičeskim: postanovka zadači i rezul'taty mašinnogo ėksperimenta [From semantic features of a word to its surface-syntactic features: Problem statement and computer experiment results]. NTI 1976/9, 25-33.

32. BAKOV, A.A., BUXALEVA, Ė.I., ZDOROV, I.P.: Ob odnom sposobe postroenija semantičeskoj klassifikacii [A method of semantic classification design]. NTI 1977/10, 11-14.

33. KRYLOV, Ju.K., JAKUBOVSKAJA, M.D.: Statističeskij analiz polisemii kak jazykovoj universalii i problema semantičeskogo toždestva slova [Statistical analysis of multiple meaning as a linguistic universal and the problem of the semantic identity of word]. NTI 1977/3, 1-6.

34. MARX, W.: Die Kontextabhängigkeit der Assoziativen Bedeutung. Zeitschrift für experimentelle und angewandte Psychologie 24, 1977, 455-462.

35. MARX, W.: Die Messung der Assoziativen Bedeutung von Sätzen. Zeitschrift für experimentelle und angewandte Psychologie 25, 1978, 113-119.

36. OSHIRO, Y.: A study of psychosemantics: On a perceiving of word meanings and personality [In Japanese]. Mathematical Linguistics 11, 1977, 65-80.

37. XAZAGEROV, T.G.: Nekotorye voprosy semantiki v svete teorii informacii [Some questions of semantics in the light of information theory]. IzvAN 36, 1977, 70-77.

L E X I C O L O G Y

38. BEČKA, J.V.: Changes in quantitative relations due to a growing corpus. PSML 5, 1977, 29-36.

39. GORODECKIJ, B.Ju.: Leksiko-statističeskaja inventarizacija kompleksa pod"jazykov [On making a lexico-statistical inventory of a complex of sublanguages]. Problemy teoretičeskoj i ėksperimental'noj lingvistiki. Moskva 1977, 21-43.

40. JIRÁKOVÁ, I.: Zavisimost' količestvennogo sostava grammatičeskix kategorij porjadka častej reči ot ob"ema častotnyx slovarej russkogo jazyka [The dependence of the quantitative structure of the grammatical category of the word-classes upon the volume of Russian frequency dictionaries]. PSML 5, 1977, 37-52.

41. KAASIK, Ü., TULDAVA, J., VILLUP, A., KÄREMAA, K.: Frequency dictionary of Estonian fiction (Words in non-conversational material) [In Estonian]. Acta et Commentationes Universitatis Tartuensis 413: Keele-Statistika 2, 1977, 5-139.

42. MARTIN, W.: Preliminaire opmerkingen over kwantitatieve lexicologie [Preliminary thoughts on quantitative lexicology]. Sterkenburg, P.v. (Ed.): Lexicologie. Groningen 1977, 189-196.

43. NADARAJŠVILI, I.Š., ORLOV, Ju.K.: Method polnoj fiksacii teksta pri lingvostatističeskom analize [A method for a complete fixation of a text in linguistic statistical analysis]. Linguistica 10, 1978, 65-84.

44. OVSIENKO, Ju.G.: Častotnyj slovar'-spravočnik russkoj razgovornoj reči [The word-count dictionary of spoken Russian]. Aktual'nye problemy učebnoj leksikografii. Moskva: Russkij jazyk, 1977, 94-101.

45. TULDAVA, J.: The frequency dictionary as an object of lexico-statistical analysis [In Estonian]. Acta et Commentationes Universitatis Tartuensis 413: Keele-Statistika 2, 1977, 141-169.

46. VANNIKOV, Ju.V.: Edinaja spravočnaja lingvostatističeskaja sistema kak baza učebnoj leksikografii [A unified linguistic statistical reference system as the base of lexicography]. Aktual'nyje problemy učebnoj leksikografii. Moskva: Russkij jazyk, 1977, 71-93.

DIALECTOLOGY

47. GERRITSEN, M., JANSEN, F.: The interplay between linguistics and dialectology: Some refinements of Trudgill's formula for dialect borrowing. Referat zum 3. Intern. Kongress über Historische Linguistik, Hamburg 22.-26. 8. 1977.

48. PŠENIČNOVA, N.N.: Primenenie metoda taksonomičeskogo analiza k klassifikacii govorov [The application of the method of taxonomical analysis to the classification of subdialects]. Dialektologičeskie issledovanija po russkomu jazyku. Moskva: Nauka, 1977, 3-14.

LANGUAGE VARIATION

49. ALEKSEEV, P.M.: Kvantitativnye aspekty rečevoj dejatel'nosti [Quantitative aspects of speech activity]. In [64], 43-58.

50. BEKTAEV, K.B., BELOCERKOVSKAJA, R.G., PIOTROVSKIJ, R.G.: Norma - situacija - tekst i lingvostatističeskie priemy ix issledovanija [Norm - situation - text and linguostatistical methods of their investigation]. In [64], 5-42.

51. BOGUSLAVSKIJ, V.M.: Prilagatel'nye vysokoj položitel'noj ocenki [Adjectives of highly positive evaluation]. In [64], 76-94.

52. BORČENKO, E.D.: Sintaksičeskij standart v gazete [The syntactical standard in newspapers]. In [64], 102-113.

53. BORUNOVA, S.N.: O razvitii norm proiznošenija (assimiljativnaja mjagkost' soglasnyx sceničeskoj reči) [On the evolution of standards of pronounciation (assimilative palatalization of consonants in the language of the theatre)]. In [64], 244-265.

54. ERMOLENKO, G.V.: Rol' rečevyx parallelej pri attribucii xudožestvennyx tekstov [The role of parallels in speech in the determination of authorship of literary texts]. In [64], 94-101.

55. GIMPELEVIČ, V.S.: Razvitie modelej oformlenija roda inojazyčnyx suščestvitel'nyx v russkom jazyke (po dannym statističeskogo analiza) [Evolution of models of the formation of gender of loan-words in Russian (based upon a statistical study)]. In [64], 189-207.

56. GRAUDINA, L.K.: Statističeskij kriterij grammatičeskoj normy [A statistical criterion for grammatical norms]. In [64], 135-173.

57. GRAUDINA, L.K.: Slovoizmenitel'nye varianty: Ves peremennyx èlementov v grammatike [Word inflection variants: The weight of variable elements in grammar]. In [64], 144-170.

58. JAKUBAJTIS, T.A., RUBINA, A.V.: Varianty slova i statistika [Variants of word and statistics]. In [64], 114-126.

59. KATLINSKAJA, L.P.: Princip êkonomii i grammatičeskie varianty (ob odnom aspekte normalizacii) [The principle of economy and grammatical variants (on one aspect of normalization)]. In [64], 173-188.

60. KOBRIN, R.Ju., PEKARSKAJA, L.A.: Lingvostatističeskij analiz upotreblenija terminov normativnyx slovarej i GOSTov v real'nyx naučno-texničeskix tekstax [Linguostatistical analysis of the use of terms of normative dictionaries and GOSTs in real scientific and technical texts]. In [64], 265-279.

61. KUZNECOVA, L.N.: O norme proiznošenija akterov na scene i v žizni (orfoêpičeskie varianty vozvratnyx častic) [On norms of the pronounciation of actors on the stage and in life (orphoepic variants of reflective particles)]. In [64], 228-244.

62. MIS'KEVIČ, G.I.: K voprosu o prinjatii normoj novyx slov [On the question of the acceptance of the standard forms of new words]. In [64], 208-218.

63. MOROZOVA, M.I.: Variantnye formy prilagatel'nyx (formy sravnitel'noj stepeni) [Alternative adjectival forms (forms of the comparative)]. In 64 , 59-75.

64. PIOTROVSKIJ, R.G., GRAUDINA, L.K., ICKOVIČ, V.A. (Eds.): Jazykovaja norma i statistika [Linguistic norm and statistics]. Moskva: Nauka, 1977, 301 pp.

65. ŠVARCKOPF, B.S.: Tipy sootnošenija variantov i statističeskie issledovanija normy [Types of correlation between variants and the statistical investigation of norms]. In [64], 126-134.

66. VORONCOVA, V.L.: Normy udarenija i statistika [Norms of stress and statistics]. In [64], 219-228.

67. VORONOV, A.L.: K voprosu areal'nogo i social'nogo var'irovanija slov v nacional'nom jazyke (na materiale sovremennogo nemeckogo jazyka) [On the question of geographical and social variation of words in national languages (based upon material from contemporary German)]. In [64], 280-299.

68. WILDGEN, W.: Differentielle Linguistik. Entwurf eines Modells zur Beschreibung und Messung semantischer und pragmatischer Variation. Tübingen: Niemeyer, 1977, 310 pp.

TEXT ANALYSIS

69. BRAINERD, B.: Graphs, topology and text. Poetics 6, 1977, 1-14.

70. CHISHOLM, D.: Generative prosody and English verse. Poetics 6, 1977, 111-153.

71. COMFORTIOVÁ, H.: On the problem of verbs in specialized texts. PSML 5, 1977, 69-89.

72. DEBIEVRE, M.: Essai d'application des méthodes de la statistique linguistique au problème posé par l'attribution du texte de la version français du Roman de Tristan. In [3], 35-54.

73. DE KOCK, J., GEENS, O.: Pour une stylistique linguistique et quantitative avec l'aide de l'ordinateur. Linguistique automatique et langues romanes (Document de linguistique quantitative 29). Paris 1977, 198-215.

74. DROMMEL, R.H.: Schätztests im Rahmen einer experimentellen Textlinguistik. BIPK 7, 1977, 11-61.

75. HELER, J.R.: Enjambement as a metrical force in romantic conversation poems. Poetics 6, 1977, 15-26.

76. KAREL'SKAJA, I.M.: O stileobrazujuščej aktivnosti roditel'nogo padeža (Iz nabljudenij s pomošč'ju statističeskoj metodiki) [The genetive as a generator of style (from observations made with the help of the statistical method)]. Leksika, Terminologija, Stili 6, Gor'kij 1977, 45-49.

77. KLIMEŠ, L.: Measurement of difficulty of Beckovský's and Palacký's historical texts for a modern reader. PSML 5, 1977, 149-161.

78. KRAUS, J.: Subjectivity of the characters in the structure of the discourse. PSML 5, 1977, 105-118.

79. LUDVÍKOVÁ, M.: On some statistical differences in two spoken

texts. PSML 5, 1977, 91-104.

80. MELIKJAN, N.: Opisanie statističeskoj struktury telegrafnyx tekstov [A description of the statistical structure of telegraph texts]. Vestnik obščestvennyx nauk Akademii Nauk Armjanskoj SSR 1977/9, 51-61.

81. MISTRÍK, J.: On modelling verbal genres. PSML 5, 1977, 191-198.

82. MIZUTANI, Sh.: Myoozyoo School and Negisi School: A statistical analysis of waka-poem vocabularies [In Japanese]. Mathematical Linguistics 11, 1977, 30-37.

83. POPOV, K.G.: Opyt primenenija statističeskogo metoda k xarakteristike ėlementov xudožestvennogo stilja pisatelja [An attempt at applying the statistical method to the characterization of the elements of the artistic style of a writer]. Československá rusistika 23, 1978, 4-10.

84. ROSS, D., Jr.: The use of word-class distribution data for stylistics: Keats' sonnets and chicken soup. Poetics 6, 1977, 169-195.

85. ŠAJKEVIČ, A.Ja.: Pozicionnyj analiz poetičeskoj strofy [Positional analysis of the poetic stanza]. Linguistica 10, 1978, 96-113.

86. TEŠITELOVÁ, M.: On the frequency of function words. PSML 5, 1977, 9-28.

87. TEŠITELOVÁ, M., NEBESKÁ, I., KRÁLÍK, J.: On the quantitative characteristics of the Czech texts of disputed authorship - RKZ. PSML 5, 1977, 119-147.

88. TULDAVA, Ju.: O kvantitativnyx xarakteristikax bogatstva leksičeskogo sostava xudožestvennyx tekstov [On the quantitative characteristics of the lexical richness of literary texts]. Linguistica 9, 1977, 159-175.

89. TULDAVA, Ju.: K probleme sopostavlenija sub"ektivnyx i ob"-ektivnyx xarakteristik stilja [On the problem of comparing the subjective and the objective characteristics of style]. Acta et Commentationes Universitatis Tartuensis 420: Studia Metrica et Poetica 2, 1977, 82-93.

90. VAŠÁK, P., MAZÁČOVÁ, S.: Rhyme, stanza and rhythmic types. PSML 5, 1977, 163-189.

91. VAŠÁK, P.: Metody určovaní autorství [Methods of determining authorship]. Praha: Academia, 1978, 264 pp.

92. ZUBOV, A.V.: Obrabotka na ESĖVM tekstov estestvennyx jazykov [On the handling of texts of natural languages with the help of the computer]. Minsk: Vyšejšaja škola, 1977, 171 pp.

PSYCHOLINGUISTICS

93. BROWN, A.L., SMILEY, S.S., DAY, J.J., TOWNSEND, M.A.R., LAWTON, S.C.: Intrusion of a thematic idea in children's comprehension and retention of stories. Child Development 48, 1977, 1454-1466.

94. BUCCI, W.: The interpretation of universal affirmative propositions. Cognition 6, 1978, 55-77.

95. DEVINE, B., LUNDBERG, U.: The influence of emotional conditions on similarity estimations of emotional terms. Scandinavian Journal of Psychology 18, 1977, 317-326.

96. DOOLING, D.J., CHRISTIAANSEN, R.E.: Episodic and semantic aspects of memory for prose. Journal of Experimental Psychology, Human Learning and Memory, 3, 1977, 428-436.

97. DOOLING, D.J., CHRISTIAANSEN, R.E.: Levels of encoding and retention of prose. Bower, G.H. (Ed.): The Psychology of Learning nad Motivation 11, 1977, 1-39.

98. ENGELKAMP, J., KRUMNACKER, H.: The effect of cleft sentence structures on attention. Psychological Research 40, 1978, 27-36.

99. FORD, M., HOLMES, V.M.: Planning units and syntax in sentence production. Cognition 6, 1978, 35-53.

100. FREDERIKSEN, C.H.: Semantic processing units in understanding text. Freedle, R.O. (Ed.): Discourse Production and Comprehension. Norwood New Jersey 1977, 57-87.

101. GÜNTHER, U., GROEBEN, N.: Abstraktheitssuffix-Verfahren: Vorschlag einer objektiven ökonomischen Messung der Ab-

straktheit/Konkretheit von Texten. Zeitschrift für experimentelle und angewandte Psychologie 25, 1978, 55-74.

102. JOHNSON, R.E., SCHEIDT, B.J.: Organizational encodings in the serial learning of prose. Journal of Verbal Learning and Verbal Behavior 16, 1977, 575-588.

103. KATZ, S., HOLLAND, M.F.: The role of a guessing strategy in the BRANSFORD-FRANKS paradigma. Acta Psychologica 42, 1978, 103-120.

104. KEENAN, J.M., MacWHINNEY, B., MAYHEW, D.: Pragmatics in memory: A study of natural conversation. Journal of Verbal Learning and Verbal Behavior 16, 1977, 549-560.

105. KIERAS, D.E.: Good and bad structure in simple paragraphs: Effects on apparent theme, reading time, and recall. Journal of Verbal Learning and Verbal Behavior 17, 1978, 13-28.

106. KINTSCH, W.: On comprehending stories. Just, M.A., Carpenter, P.A. (Eds.): Cognitive Processes in Comprehension. New Jersey: Hillsdale, 1977, 33-62.

107. MARSLEN-WILSON, W.D., WELSH, A.: Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology 10, 1978, 29-63.

108. MARX, W.: Die Messung der assoziativen Bedeutung von Sätzen. Zeitschrift für experimentelle und angewandte Psychologie 25, 1978, 113-119.

109. MORRIS, P.E.: Frequency and imagery in word recognition: Further evidence for an attribute model. The British Journal of Psychology 69, 1978, 69-75.

110. SCHMIDT, R.: Frequency estimation in verbal learning. Acta Psychologica 42, 1978, 39-58.

111. SHERMAN, J.L., KULHAVY, R.W.: Abstractness and prose comprehension. Acta Psychologica 42, 1978, 59-65.

112. SILVERMAN, G.: The effect of topic and repetition on type-token ratios of spoken monologues. Language and Speech 20, 1977, 232-239.

113. TTREISMAN, M.: Space or lexicon? The word frequency effect and the error response frequency effect. Journal of Verbal Learning and Verbal Behavior 17, 1978, 37-59.

114. WAERN, Y.: On the relationship between knowledge of the world and comprehension of texts. Assimilation & accomodation effects related to belief structure. Scandinavian Journal of Psychology 18, 1977, 130-139.

115. WAERN, Y.: Comprehension and belief structure. Scandinavian Journal of Psychology 18, 1977, 266-274.

116. WALTER, D.A.: Discrete events in word encoding. The American Journal of Psychology 90, 1977, 655-662.

LANGUAGE LEARNING

117. NESTLER, K.: Zur Ermittlung von Voraussetzungen für die Formulierung von Normen für die angemessene sprachliche Gestaltung von Lehrtexten. Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 30, 1977, 75-79.

SOCIOLINGUISTICS

118. HAYNES, L.M.: The sociology of local names of plants in Guyana, South America. Linguistics 193, 1977, 87-101.

DOCUMENTATION

119. DIETZE, J.: Ein frequenzstatistisches Verfahren zum automatischen Indexieren. Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 30, 1977, 58-64.

120. HELBICH, J.: The opinion of research workers on the selective power of single words. PSML 5, 1977, 58-64.

All contributions to the CURRENT BIBLIOGRAPHY should be sent to Prof. Dr. Werner Lehfeldt, Universität Konstanz, Fachbereich Sprachwissenschaft, P.O.Box 733, D-7750 Konstanz