GEORGIA INSTITUTE OF TECHNOLOGY
OFFICE OF CONTRACT ADMINISTRATION
SPONSORED PROJECT INITIATION

Date: August 24, 1979

Project Title: Syntatic Structure of Information and Information Processes

Project No: G-36-639

Project Director: Dr. Vladimir Slamecka

Sponsor: National Science Foundation

Agreement Period: From 8/1/79 Until 1/31/82 (Grant Period)

Type Agreement: Grant No. IST-7827002, dtd. August 1, 1979

Amount:
$137,790 NSF (G-36-639)
7,326 GIT (G-36-333)
$145,116 Total

Reports Required: Annual Progress Report(s); Final Project Report

Sponsor Contact Person (s):

Technical Matters

NSF Program Official
Edward C. Weiss
Program Director
Fundamental and Advanced Research Program
Division of Information Science and Technology
Directorate for Scientific, Technological, and International Affairs
National Science Foundation
Washington, D. C. 20550

202/632-5818

Contractual Matters
(thru OCA)

NSF Grants Official
Philip M. King
MPE/STIA Branch, Section II
Division of Grants and Contracts
Directorate for Administration
National Science Foundation
Washington, D. C. 20550

202/632-5965

Defense Priority Rating: n/a

Assigned to: ICS (School/Laboratory)

COPIES TO:

- Project Director
- Division Chief (EES)
- School/Laboratory Director
- Dean/Director—EES
- Accounting Office
- Procurement Office
- Security Coordinator (OCA)
- ✓ Reports Coordinator (OCA)
- Library, Technical Reports Section
- EES Information Office
- EES Reports & Procedures
- Project File (OCA)
- Project Code (GTRI)
- Other _____

GEORGIA INSTITUTE OF TECHNOLOGY OFFICE OF CONTRACT ADMINISTRATION

SPONSORED PROJECT TERMINATION SHEET

Date 6/15/82

Project Title: Syntatic Structure of Information and Information Processes

Project No: G-36-639

Project Director: Dr. Vladimir Slamecka

Sponsor: National Science Foundation

Effective Termination Date: 1/31/82

Clearance of Accounting Charges: 1/31/82

Grant/Contract Closeout Actions Remaining:

- [ ] Final Invoice and Closing Documents
- [x] Final Fiscal Report
- [ ] Final Report of Inventions
- [ ] Govt. Property Inventory & Related Certificate
- [ ] Classified Material Certificate
- [ ] Other _____

Assigned to: Information and Computer Sciences (School/~~Laboratory~~)

COPIES TO:

Administrative Coordinator	Research Security Services	EES Public Relations (2)
Research Property Management	Reports Coordinator (OCA) ✓	Computer Input
Accounting	Legal Services (OCA)	Project File
Procurement/EES Supply Services	Library	Other _____

FORM OCA 10:781

NATIONAL SCIENCE FOUNDATION
Washington, D.C. 20550

FINAL PROJECT REPORT
NSF FORM 98A

PLEASE READ INSTRUCTIONS ON REVERSE BEFORE COMPLETING

PART I–PROJECT IDENTIFICATION INFORMATION

1. Institution and Address	2. NSF Program	3. NSF Award Number
Georgia Institute of Technology 225 North Ave. NW Atlanta, GA 30332	Information Science & Technology	IST-7827002
	4. Award Period From 8/1/79 To 1/31/82	5. Cumulative Award Amount $137,790

6. Project Title

Syntactic Structure of Information and Information Processes

PART II–SUMMARY OF COMPLETED PROJECT *(FOR PUBLIC USE)*

The objects of this long-term research into the syntactic structure of information and information processes are the shape of natural language words, and the nature of the type-token relation for natural language words in contiguous text. In the present research phase, advanced instrumentation has been developed for eventual measurement of Shannon's redundancy curve as a function of sign shape. Based on 30,000 word measurements, an Mk VI Vernier eidometer has been tested to a precision of 3.8 bpm; the design studies show that the instrument will come close to the original design specification of 5.0 bpm. An electronic teescope was designed, built, and tested; and a new method was developed for timing the teescope for information measurement. The nature of information decay in short-term memory was studied in relation to measurement of interpretation errors, and a computer program was written to assist this measurement.

PART III–TECHNICAL INFORMATION *(FOR PROGRAM MANAGEMENT USES)*

1. ITEM *(Check appropriate blocks)*	NONE	ATTACHED	PREVIOUSLY FURNISHED	TO BE FURNISHED SEPARATELY TO PROGRAM	
				Check (✓)	Approx. Date
a. Abstracts of Theses					
b. Publication Citations			✓		
c. Data on Scientific Collaborators			✓		
d. Information on Inventions			✓		
e. Technical Description of Project and Results			✓		
f. Other *(specify)*					

2. Principal Investigator/Project Director Name *(Typed)*	3. Principal Investigator/Project Director Signature	4. Date
Vladimir Slamecka		5/20/82

Form Approved OMB No. 9[illegible]

GEORGIA INSTITUTE OF TECHNOLOGY
SCHOOL OF INFORMATION AND COMPUTER SCIENCE • ATLANTA, GEORGIA 30332 • (404) 894-3152

April 28, 1982

Dr. Edward C. Weiss
Division of Information Science
and Technology
National Science Foundation
Washington, DC 20550

Dear Dr. Weiss:

This letter and its enclosures serve as the final report on our two-year NSF grant, IST-7827002, "Syntactic Structure of Information and Information Processes." The grant expired on January 31, 1982. The attached technical reports (Appendices B-F) emanated from the project since the submission of our last report; that report contained papers written prior to its date. A complete list of 19 papers produced during the two-year grant period appears in Appendix A.

The objectives of our research were three-fold: advances in instrumentation for basic research in information science; explication of interpretation errors, a concept used in place of Miller-Bruner-Postman's original concept of placement error; and measurement of Shannon's Redundancy Curve as a function of shape. With respect to the first objective, the intention was to develop a more precise version of the eidometer, and to design, construct, and develop measurement methods for an electronic teescope. The greater part of the effort of the project was expended on making 30,000 word measurements for use in designing the Mk VI Vernier-scale eidometer. The eidometer achieves a precision of 3.8 bpm, compared to 2.8 bpm by the older Mk IV design. The word measurements are complete, and the Mk VI design is nearly complete, although several months of work will be required to finish it. The design studies show that the instrument will come close to the original design requirement of 5.0 bpm.

The design of the electronic teescope was completed and the bid for its construction was won by E.J. McGowan Associates of Chicago, who subsequently built and delivered the system. A method of timing the teescope for information measurements was developed; a paper on this technique, authored by Richard Lo, will be forthcoming.

Substantial progress has been made in studies to determine the nature of information decay in short-term memory, and to explicate the interpretation error concept. Four major effects on the cause of most of the

loss of information in short-term memory which must be accounted for in the measurement of interpretation errors are: transposition errors, displaced strings, confusion errors, and serial information decay. A computer program was developed to analyze teescope data for these effects, and to determine the relative influence each has on the measurement of interpretation errors.

We have not been able to complete the measurement of Shannon's Redundancy Curve as a function of shape, or to carry out the final analysis of information decay in immediate memory. In part this was due to a six-month delay in the delivery of the electronic teescope. Also, although the project was completed within the original budget, inflation during the three and a half years since the submission of our proposal reduced the value of the budget by approximately one-third, forcing us to decrease the level of man-effort and to sacrifice in the area of instrumentation.

I trust that you will consider the productivity of the project, as represented by 19 papers, to be above the ordinary. In addition, five other papers are in the process of being written since the expiration of the project.

On behalf of the School of ICS and my colleagues and students who participated in this project, may I take this opportunity to express to the National Science Foundation my sincere appreciation for providing support for the project. It gives me pleasure to thank you personally for your support of our work and for your dedicated guidance of basic research in information science.

Sincerely yours,

Vladimir Slamecka
Principal Investigator

cc: Director, ICS
Office of Contract Administration

Enclosures

APPENDIX A

LIST OF PAPERS PRODUCED BY
SYNTACTIC STRUCTURE OF INFORMATION PROJECT

1. Pearson, C. "The Problem of Communicating Results in Empirical Semiotics." Presented at the SIG/ES Workshop on Immediate Problems in Empirical Semiotics held at the Second International Semiotics Congress; Vienna, Austria; July 2-6, 1979. Submitted for inclusion in the published proceedings.

2. Pearson, C. "Information Decay in Immediate Memory: A Second Order Correction to the Law of Word Interpetation." Presented at the Experimental Semiotics Session of the Second International Semiotics Congress; Vienna, Austria; July 2-6, 1979. Submitted for inclusion in the published proceedings.

3. Pearson, C. "Empirical Methodology in Information Science and Semiotics." Presented at the Workshop on Fuzzy Formal Semiotics and Cognitive Processes held at the Second International Semiotics Congress; Vienna, Austria; July 2-6, 1979. Submitted for inclusion in the published proceedings.

4. Pearson, C. "Theses of Empirical Semiotics." Presented at the Theses Session of the Second International Semiotics Congress; Vienna, Austria; July 2-6, 1979. Submitted for inclusion in the published proceedings.

5. Pearson, C. "Semiotics and the Measurement of Shape." A Seminar presented to the Technische Universität Berlin; July, 1979. English version submitted to Progress in Information Science and Technology; German version submitted to *Zeitschrift für Semiotik*.

6. Pearson, C. "Performance Evaluation of the Mk V Eidometer." Presented to the Symposium on Empirical Semiotics sponsored by SIG/ES at the 1979 Annual Conference of the Semiotic Society of America; Bloomington, Indiana; October, 1979.

7. Howell, D.P. "A New Technique for Eidometer Construction." Presented to the Symposium on Empirical Semiotics sponsored by SIG/ES at the 1979 Annual Conference of the Semiotic Society of America; Bloomington, Indiana; October, 1979.

8. Pearson, C. "Attributes of Information." Presented to the session on Foundations of Information Science at the Annual Conference of the American Society for Information Science; Minneapolis, Minn., October, 1979.

9. Pearson, C. "The Echelon Counter: A New Instrument for Measuring the Vocabulary Growth Rate and the Type-Token Relationship." Presented to the session on Concepts and Measurement sponsored by SIG/FIS at the Annual Conference of the American Society for Information Science; Anaheim, California; October, 1980. Published in Communicating Information: Proceedings of the 43rd ASIS Annual Meeting, 17(1980), p367-369.

10. Pearson, C. "The Basic Concept of the Sign." Presented to the session on Concepts and Measurement sponsored by SIG/FIS at the Annual Conference of the American Society for Information Science; Anaheim, California; October, 1980. Published in Communicating Information: Proceedings of the 43rd ASIS Annual Meeting, 17(1980), p367-369.

11. Pearson, C. "Information Science: The Challenge of a Basic Science." Presented to the session on The Challenge of Information Science sponsored by SIG/FIS at the Annual Conference of the American Society for Information Science; Anaheim, California; October, 1980.

12. Lo, R.H. "Measurement of Information Transfer Rates." Presented to the Symposium on Empirical Semiotics sponsored by SIG/ES at the Annual Conference of the Semiotic Society of America; Lubbock, TX; October, 1980. To appear in the published proceedings of the conference.

13. Pearson, C. "The Mark VI: A New Eidometer Design Concept." Presented to the Symposium on Empirical Semiotics sponsored by SIG/ES at the Annual Conference of the Semiotic Society of America; Lubbock, TX; October, 1980. To appear in the published proceedings of the conference.

14. Pearson, C. "Scientific Paradigms for Semiotics and Information Science." An invited paper presented to the plenary session on Paradigms for Empirical Semiotics at the Annual Conference of the Semiotic Society of America; Lubbock, TX; October, 1980. Summary to appear in the published proceedings of the conference.

15. Phongphatar, T. "SemLab's Type-Token Program Now Available on the Cyber-70." Press release mailed to journals in information science and semiotics; August, 1980.

16. Pearson, C. "The Role of Scientific Paradigms in Empircal Semiotics." Invited address delivered to the plenary session on Paradigms of Empirical Semiotics at the Annual Conference of the Semiotic Society of America; Lubbock, TX; October, 1980. Summary to appear in the published proceedings of the conference.

17. Flowers, J., C. Pearson, T. Phonphatar. "A Method for Generating High Order Markov Words." Presented at the 19th Annual Southeast Regional ACM Conference; Atlanta, Georgia; March, 1981.

18. Pearson, C. "The Semiotic Paradigm." Presented to the session on Basic Approaches to Fundamental Research in Information Science sponsored by SIG/FIS at the 1981 Annual Conference of the ASIS; Washington, DC; October, 1981.

19. Pearson, C. "Application of the Finite-Difference Calculus to the Observation of Symbol Processes." Presented to the session on The Role of Mathematics in Semiotic Observations sponsored by SIG/ES at the Fourth Annual Symposium on Empirical Semiotics held in conjunction with the Sixth Annual Meeting of the Semiotic Society of America; Nashville, TN; October, 1981. To appear in the Proceedings of the SSA.

APPENDIX B

THE ROLE OF SCIENTIFIC PARADIGMS IN EMPIRICAL SEMIOTICS

Charls Pearson
School of Information and Computer Science
Georgia Institute of Technology
Atlanta, 30332, USA

September, 1980

A summary of an invited address delivered to the plenary session on "Paradigms of Empirical Semiotics" sponsored by SIG/ES as part of its third annual "Symposium on Empirical Semiotics" at the 1980 Annual Conference of the SSA in Lubbock, Texas; October 17, 1980. To appear in the published proceedings.

THE ROLE OF SCIENTIFIC PARADIGMS IN EMPIRICAL SEMIOTICS

Charls Pearson
School of Information and Computer Science
Georgia Institute of Technology
Atlanta, 30332, USA

ABSTRACT

The notion of a "scientific paradigm" was popularized by Thomas Kuhn in The Structure of Scientific Revolutions, first published in 1962. For Kuhn's purposes, it was not necessary to classify scientific paradigms into various categories. However, in order to analyze the paradigms of empirical semiotics and determine which paradigms in other empirical sciences have analogies which carry over to empirical semiotics and which do not, it is necessary to classify scientific paradigms into at least five categories. These are: 1) conceptual, philosophical, and linguistic paradigms; 2) theoretical paradigms; 3) mathematical paradigms; 4) experimental paradigms; and 5) applicational paradigms.

This paper summarizes the above classification system and describes and characterizes these five paradigm categories. It falls into the area of philosophical semiotics. It assumes an empirical approach to semiotic knowledge but is independent of any specific theoretical, experimental, or mathematical paradigms. Indeed, it sets the stage for any later discussion of such paradigms.

THE ROLE OF SCIENTIFIC PARADIGMS IN EMPIRICAL SEMIOTICS

Charls Pearson

School of Information and Computer Science
Georgia Institute of Technology
Atlanta, Georgia 30332

The notion of a "scientific paradigm" was popularized by Thomas Kuhn in The Structure of Scientific Revolutions, first published in 1962 [1]. For Kuhn's purposes, it was not necessary to classify scientific paradigms into various categories. However, in order to analyze the paradigms of empirical semiotics and determine which paradigms in other empirical sciences have analogies which carry over to empirical semiotics and which do not, it is necessary to classify scientific paradigms into at least five categories. These are: 1) conceptual, philosophical, and linguistic paradigms; 2) theoretical paradigms; 3) mathematical paradigms; 4) experimental paradigms; and 5) applicational paradigms.

The purpose of this paper is to motivate the above classification system and to describe and characterize these five paradigm categories. It falls into the area of philosophical semiotics. It assumes an empirical approach to semiotic knowledge but is independent of any specific theoretical, experimental, or mathematical paradigms. Indeed, it sets the stage for any later discussion of such paradigms.

INTRODUCTION

Despite its milleniums-long adumbration, semiotics has reached no agreed-upon paradigms, in Kuhn's sense of the word, and in fact, there is little agreement on what the competing paradigms are. The theoretical paradigms are vague and imprecise, the experimental paradigms unrecognized, and the mathematical paradigms often ignored. All this makes for exceeding difficulties in the communication of results within empirical semiotics.

Scientific communication -- the communication of precise and rigorous scientific results -- requires the existence of universally agreed-upon paradigms -- or at least universal agreement on what the disagreed-upon paradigms are -- in order to take place effectively. In the present state of empirical semiotics this situation does not exist. In fact, the negative status of the situation is self-reinforcing in that the inability to communicate effectively, engendered by the lack of agreed-upon paradigms, in turn hinders the development of agreement on satisfactorily evolved paradigms.

Some way must be found to break this circle of infinite regress. Without agreement on what the other competing paradigms are and even without precise and explicit understanding of our own paradigms, we must begin to acknowledge and talk about these paradigms and the role they play in empirical analysis. At the SIG/ES Workshop on Immediate Problems in Empirical Semiotics held at the Second International Semiotics Congress in Vienna last July, Pearson proposed a way of attacking this problem [5].

As modified and finally adopted by the workshop, and later adopted last year by SIG/ES also, as a recommendation for all papers within empirical semiotics the proposal requires each of us in presenting results in empirical semiotics to state our own paradigms. In most cases this need not be elaborate or precise -- a few sentences should do. But we should be aware of our own, and each other's, methodology, procedures, and assumptions. Since most papers in empirical semiotics emphasize only one of the five paradigm types of empirical language, theory, experiment, mathematical analysis, or application, this proposal was specifically to mention, or state explicitly, the three or four paradigms other than the one being specifically discussed in the paper.

If this proposal is adopted for the presentation of papers in empirical semiotics generally, then we may expect that within only a few short years we may reach agreement on the broad outlines of what the competing paradigms are, and it will gradually become obvious to us all what needs to be done to make them more precise and to empirically assess the relative merits of one against the other. Indeed, the theme of today's symposium was adopted with this in mind.

It therefore behooves us to examine the concept of a scientific paradigm and to attempt to establish a classification into categories.

In the next section, I discuss five categories of scientific paradigms that I think will play an important role in the development of a scientific semiotics. The conclusions are summarized in section 3 and all references are listed alphabetically in section 5.

THE PARADIGMS OF SCIENCE

The development and progress of science has been shown to depend in an essential way on the process of scientific communication. Five different kinds of empirical paradigms have been recognized and all five are necessary for effective scientific communication. These are 1) philosophic, conceptual, or linguistic paradigms; 2) theoretical paradigms; 3) experimental paradigms; 4) mathematical paradigms; and 5) applicational paradigms.

Conceptual Paradigms and the Language of Science

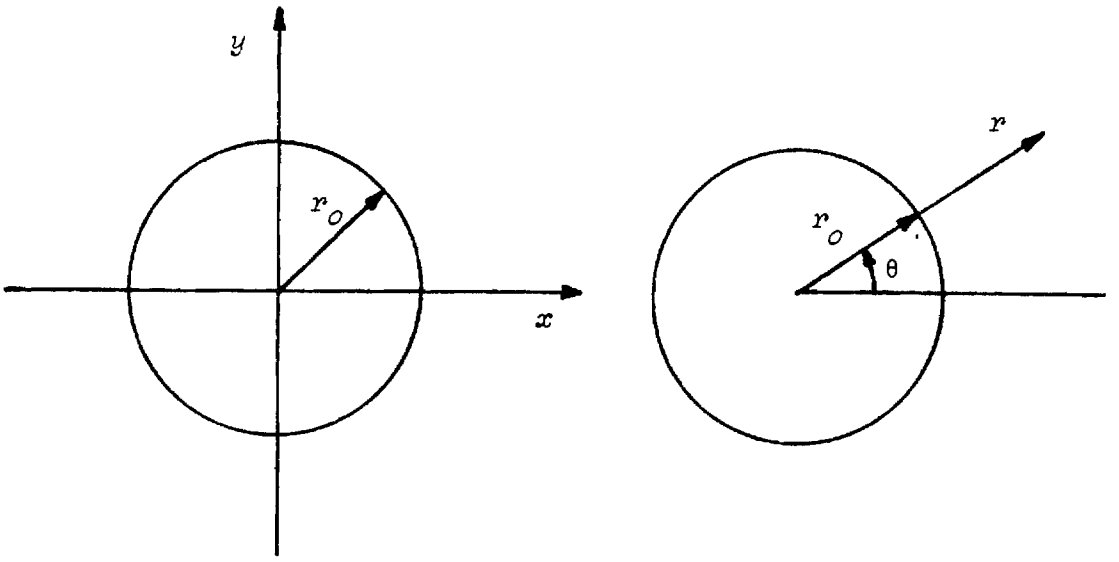
Philosophic, conceptual, or linguistic paradigms provide the very language in which the scientist carries out his thinking, frames his theories, designs his experiments, analyzes his results, etc. Linguistic paradigms embody basic metaphysical assumptions, either explicitly or implicitly, and provide a terminology, a grammar (phraseology), context, point-of-view, *Weltanschauung*, and a decision on what problems and phenomenas are of of interest and which are to be ignored. Examples of several major language paradigms are: 1) empirical language; 2) religious language; and 3) literary language.

Languages are to scientists as coordinate systems to mathematicians. There are no right or wrong ones, only better or worse ones for particular purposes. And a good one can work wonders for creativity while a bad one can block even the most powerful thinker. They are nonsubstantive in the sense that they are like mathematical coordinate systems. A circle may be described equally precisely in polar coordinates or rectangular coordinates; these are merely two distinct geometrical languages.

However, their effects may be drastically substantive

in that certain empirically substantive questions may be drastically easier to express in one language than another. This is illustrated in figure 1. Figure 1.a shows a circle as described by rectangular coordinates and gives the corresponding algebraic equation. Figure 1.b shows the circle as described by polar coordinates, and the much simpler algebraic equation associated with the polar description.

Solution procedures may be substantially easier to think out in some language different from the usual one, etc. As an example, it was drastically easier for Kepler to discover and state his laws of planetary motion using Copernicus's heliocentric language of astronomy than Ptolemy's geocentric



$x^2 + y^2 = r_o^2$ $\qquad$ $r = r_o$

a) $\qquad$ b)

Fig. 1. The circle described in both rectangular and polar coordinates, and both geometrically and algebraically relative to each.

language. Discovering the best language for a given branch of science is a trial and error process. It can only be determined aposteriori, and never apriori. Like other empirical paradigms, linguistic paradigms evolve as a result of our experience in using them and occasionally go thru Kuhnian revolutions.

Several example linguistic paradigms of semiotics are 1) Peirce's language of logical analysis; 2) continental, or French, structuralism; 3) Marxist, or Soviet, language of process and action; and 4) my own Language of Menetics which was explicitly designed for its use in the statement and solution of empirical problems in semiotics.

Many of our most important scientific results are expressed not in the form of quantitative laws, but only qualitatively in the adoption of a system, or language. There is no law of Copernicus, for example, only the Copernician system, or heliocentric language of astronomy and yet this one change in language has often been credited with enabling all of the results of modern astronomy. To come closer to home, I will give a linguistic example. We never talk of Boas's Law, for instance, we just use the language of phonemics and structural linguistics which Sapir was able to develop based on Boas's results. And the structuralist worldview and the DeSaussurian discussions out of which it arises are regarded by many as the beginning of modern "Scientific" linguistics.

In discussions of scientific methodology we are often instructed to choose an appropriate notation. But this is only an approximation to the true problem, that of choosing a good language. A system of notation is not a language -- it is a small, but important part of a language. A language includes a notation, as well as a terminology, a viewpoint, a selection of which observable phenomenas to be interested in, and an approch to integrating all of this. In fine, a language is nothing short of a complete *Weltanschauung*. Kuhn [1] indicates an understanding of both the nature and role of languages in science. In all cases of creativity, he says, one of the first steps is to use the imagination to construct, out of data supplied by memory and observation, a framework of ideas that will serve as a foundation for further work. This framework with its attendant terminology and notation is the language of the investigation.

As an example of the confusions that can arise in discussions of this topic, I have been asked how one could characterize Newton's laws of motion as a linguistic development. The answer, of course, is that one would not normally do so. Newton's work was a piece of pure science carried out primarily in the language of the Copernican system as modified by Kepler and Galileo. Newton did, however, modify the language he received by augmenting it with the terminology for "action at a distance" and adding a whole new notation system, that of the "fluxions". In order to see the development of language at work in physics, we must look about 150 years earlier to Copernicus's development of the heliocentric system.

The importance of the linguistic framework is beginning to be recognized even among the applied investigators of our own field. Newell and Simon in a discussion of the nature of computer science, for instance, say:

> All sciences characterize the essential nature of the systems they study. These characterizations are invariably qualitative in nature, for they set the terms within which more detailed knowledge can be developed. Their essence can often be captured in very short, very general statements. One might judge these general laws, due to their limited specificity, as making relatively little contribution to the sum of a science, were it not for the historical evidence that shows them to be results of the greatest importance [2,p115].

Theoretical Paradigms

Theoretical paradigms state the basic theoretical principles which are to be used in deriving explanations of the fundamental phenomenas of interest and the observational laws describing them, and provide the translation rules for interpreting theoretical concepts in terms of observational concepts. Examples of several theories of physics are: 1) Newton's Theory of Gravitation; 2) Einstein's Theory of Gravitation (General Relativity); 3) Maxwell's Electromagnetic Theory; etc. Theories compete empirically on the basis of their ability to explain known phenomenas, their simplicity and elegance, and their ability to motivate new empirically interesting questions and experimental procedures.

Examples of semiotic theories are: 1) Rossi-Landi's Theory of Economic Sign Structure; 2) Peirce's Theory of Sign Process; 3) Morris's Theory of Sign Structure; and 4) my own Universal Sign Structure Theory.

It is necessary to distinguish clearly between models, which are just mathematical functions or other mathematical structures, and whose discussion falls within the domain of applied mathematics, and theories, which contain models as one or more of their components but also contain theoretical interpretations in terms of semiotic principles and observational interpretations in terms of translation rules between theoretical concepts and observational concepts. It is this ability to interpret a semiotic theory in terms of experimentally controlled observations that gives it its status as an empirical theory.

Just as in any other empirical science, a scientific understanding of semiotic knowledge is gained only by the deliberate invention of explicitly testable and mathematically specified theories whose purpose is to explain how semiotic knowledge (the mathematically analyzed data from controlled experiments) fits together in a simple and unified way.

The invention of such theories occurs by abduction, or Peirce's third mode of reasoning. Since theories are the deliberate creation of the fallible human mind they must be validated by testing. This occurs by a combination of mathematical deduction within the theoretical realm, translation from the theoretical language to the observational language, and comparison to the results of induction on the experimental data.

The results of experimental observation are isolated facts, a collection of individual data, ontological singulars. Science is not interested in isolated facts *per se*. By induction, invariant regularities in the data are determined. These are called 'laws of semiotic nature' and have the status of ontological generals. It is this general knowledge which is the first goal of science. Laws provide us with semiotic knowledge, but they give us no scientific understanding. Laws do not explain their own existence, they just exist. They do not tell us why they are as they are nor explain

the relations between themselves. In order to obtain this second, or higher, goal of science, theories are required. Theories are ontological abstractions. They frame hypotheses in terms of nonobservable concepts such that if the theories were true* then this would explain why the laws are such as they are.

The results of mathematical deduction on the theories are called 'theorems'. Theorems are also ontological abstractions, but they are necessary in order to subject the theory to eventual observational test. This is done by translating certain theorems of the theory into the observational language. If the translated theorem matches a law, it is accorded evidence in favor of the theory. If the translated statement accords with no known law, experiments are designed, conducted, and the results analyzed in order to search for the predicted regularity. Most experiments in science arise as a result of this directed search process. If the regularity is found, this also is accorded evidence in favor of the theory. Evidence from previously unknown regularities is often accorded higher value than evidence from the known regularities which motivated invention of the theory. If the translated theorem is contradicted by the results of observation and the known laws, this is accorded evidence against the theory.

This has been a simplistic presentation, purely for the purpose of presenting the concept of scientific theories in empirical semiotics, and should not be interpreted as implying that theories are abandoned or adopted on the basis of an algebraical summing up of the evidence in favor of or against them.

Experimental Paradigms

Experimental paradigms provide the experimental methodologies, the measurement techniques, and the procedures to be used in designing and carrying out rigorously controlled experiments for submitting questions to nature for her to

*The word 'true' is used here in an abstract, or metaphorical, sense, since by definition theories are abstractions framed in non-observational terms and hence are neither true nor false in the positivistic sense.

answer. The Michelson-Morley and Davisson-Germer experiments are well-known paradigms of experimental physics. Word Recognition and Sentence Comprehension are well-known paradigms of experimental psychology. Closer to home, Zipf's Word Counting Procedure and my own eidometric techniques provide paradigm examples from experimental semiotics.

Experimental paradigms interact with technology in that precisely controlled experimental methodologies require the use of precise, objective, and reliable instruments for the control and measurement of the experimental phenomenas. In semiotics these instruments very often have to be invented in order to make an experiment possible. The validity, reliability, precision, and repeatability of scientific instruments must be assayed for the procedures in which they will be used. This gives science very much the aspect of metrology. Two examples of instruments designed specifically for semiotics experiments and having assayed performance are the eidometer and the echelon counter. The eidometer measures the eidontic deviance of word shapes and the Mk V design has an assayed precision of 3.79 bpm when used with the procedures required for the Word Interpretation Experiment The echelon counter measures word types and word tokens in text samples and has an assayed precision of $\pm$ $1/n$ wt where n wk is the size of the sample being measured.

Mathematical Paradigms

Mathematical paradigms provide tools for reasoning as a service to the theoretical, experimental, and applied paradigms. They provide the analytical methods and procedures for manipulating theoretical principles, solving equations, analyzing data, designing experiments, analyzing instrument error, and reducing statements in basic science to their practical applications. Three well-known mathematical paradigms in quantum mechanics are: 1) calculus of partial differential equations; 2) matrix calculus; and 3) operator calculus. Currently, the most useful mathematical paradigms in empirical semiotics stem from inferential statistics, discrete mathematics, and finite difference techniques.

Applicational Paradigms

Applicational paradigms, while not properly a part of basic science itself, sometimes help determine the goals of

theory building and the direction of development for the basic science in that they can help determine what feedback from practical applications to be sensitive to and which phenomenas to explain. For instance, even though thermodynamic laws are what they are because they describe objective and general regularities of nature, the way they were discovered and the order in which they were discovered was largely determined by the goal of explaining the practical phenomena of steam engineering. In semiotics today, information technology is playing much the same role as did steam engineering in 19th century physics. The field of computer science is also beginning to require explanations in terms of basic semiotic laws and theories for its many practical relationships, especially in the field of language design.

We should be aware of the possibility of "pure science", the development of basic science in isolation from any projected application. Peirce was especially sensitive to this possibility, calling it the method of the true scientist: one who sought intellectual understanding for the pure joy of learning and with no thought of practical benefit in mind.

CONCLUSIONS AND SUMMARY

There are five distinct kinds of paradigms required for the scientific development of semiotics. These paradigms will evolve empirically from our experience with working with and revising them. In working with and revising them, they will interact with each other. There is no such thing as a paradigm in isolation. The linguistic paradigm, theoretical paradigm, experimental paradigm, and mathematical paradigm fit together as a unit but each must be present. When one changes, so too does each of the others to a certain extent.

We are now very much as physics at the time of Archimedes in the stage of our very first paradigms. Our present ones are very crude but we must use them to gain empirical experience so as to improve the ones we have, test and evaluate them, compare between competing paradigms, and occasionally even go thru Kuhnian revolutions.

ACKNOWLEDGEMENTS

This work was partially supported by the National Science Foundation under grant No. IST-7827002. I would

also like to thank my colleague Vladimir Slamecka for helpful discussions and encouraging support.

REFERENCES

1. T.S. Kuhn, "The Structure of Scientific Revolutions," University of Chicago Press, Chicago (1962).

2. A. Newell and H.A. Simon, "Computer science as empirical inquiry: symbols and search," CACM, 19(3):113-126 (1976).

3. C.K. Ogden and I.A. Richards, "The Meaning of Meaning," Harcourt, Brace & World, New York (1923). (8th ed., Harvest Books, 1946)

4. C. Pearson, "Towards an Empirical Foundation of Meaning." Ph.D. Dissertation, Georgia Institute of Technology, Atlanta (1977). Available from University Microfilms, 300 North Zeeb Road, Ann Arbor, 48106, USA.

5. C. Pearson, "The Problem of Communicating Results in Empirical Semiotics." Presented at the SIG/ES Workshop on Immediate Problems in Empirical Semiotics held at the Second International Semiotics Congress in Vienna, Austria, July 1979. To appear in the published proceedings of the Congress.

6. C. Pearson, "Theses of Empirical Semiotics." Presented at the Theses Session held at the Second International Semiotics Congress in Vienna, Austria, July 1979; and submitted for inclusion in the published proceedings of the Congress.

7. C. Pearson, "Semiotics and the Measurement of Shape." Seminar presented at the Technische Universitaet Berlin, July 1979.

8. C. Pearson and V. Slamecka, "A theory of sign structure," Semiotic Scene, 1:1-22 (1977).

THE EPISTOMOLOGICAL STATUS OF SHANNON'S REDUNDANCY CURVE AND MARKOV'S LAW

By Charls Pearson
School of Information and Computer Science
Georgia Institute of Technology

March 1981

Presented at the 19th Annual Southeast Regional ACM Conference; Atlanta, Georgia; March 27-29, 1981.

THE EPISTOMOLOGICAL STATUS OF SHANNON'S REDUNDANCY CURVE AND MARKOV'S LAW

By Charls Pearson
School of Information and Computer Science
Georgia Institute of Technology

ABSTRACT

In 1951 Claude Shannon published an analysis of natural language information systems based on his 1948 calculus of modal statistics. This analysis contained a curve predicting a relationship between the redundancy of an information system approximating natural language and the order of the Markov-chain approximating the source of that language and was able to determine mathematical upper and lower bounds for this dependence. His work was motivated in part by an earlier result by A.A. Markov showing that as the order of a Markov-chain approximating the source of a natural language information system increases, the shape of the symbols generated by that system approach the normal shape of symbols in the natural language from which the frequencies for the Markov-chain were compiled.

Shannon's Curve and Markov's Law are useful empirical relations altho their present expression is flawed epistomologically by employing one empirical and one formal coordinate each. The concept of eidontic deviance allows us to correct this in part by providing a quantitative empirical measurement of the strangeness of a word shape relative to normal word shapes in a given natural language. This leads to the Law of Word Interpretation and this together with the Miller-Bruner-Postman Effect leads to the possibility of experimental verification and precise measurement for the first time of Shannon's curve.

1. INTRODUCTION:

The objective of our research in the Georgia Tech SemLab has been to investigate the nature of the sciences underlying computer technology and information technology. In investigations reaching back over five years or more, and sponsored primarily by two grants from the NSF Division of Information Science and Technology, there are many indications that this is the same science: *semiotics*, or the science of signs and sign processes. Thus computer technology and information technology are closely related by their common scientific underpinnings. For instance, computer technology is not usually regarded as the study of wiring, transistors, and the flow of electricity; this is a proper part of electrical engineering. Computer technology is more fittingly characterized as the study of how computers compute, how they process programs, how to design computational processes, and how to design languages and operating systems to facilitate such operations. Each of these processes involves signs in an integral way and requires a knowledge of semiotic structure for a fuller understanding of how to effectively and efficiently design them.

The evidentiary support for the above example is too long and detailed to give in the time and space limitations of this paper and is not its purpose anyway. The example was given only for background illustration. I'm sure that each of you has at least an intuitive feeling for the validity of this claim. Likewise the arguments showing that semiotics is the science underlying computer technology and information technology are many and diverse, each too long and detailed to present here. The main purpose of this paper is to sketch an outline of just one of these arguments, one concerning the epistomological status of Shannon's redundancy curve and Markov's Law. The background of these relations is given in the next section.

2. BACKGROUND:

In 1948 Claude Shannon published his calculus of modal statistics [10], which has most unfortunately received the appellation of *information theory*; and in 1951 he applied this to the analysis of printed American* [11]. While Shannon's 1948 publication is well-known among computational engineers, the 1951 publication, with its focus on natural language, has received much less attention in the computational literature. It contained, however, a curve, whose nature has been a source of puzzlement in the information science literature for the last 30 years. In [11] Shannon attempted a statistical analysis of printed American using the uncertainty function and his calculus of modal statistics. He defined the redundancy of an information system as one minus the ratio of the actual uncertainty associated with the alphabet (or vocabulary, in other situations) as used in a particular information system divided by the maximum uncertainty obtainable with the use of that same

*Which he incorrectly called "English".

alphabet in any information system, or

$$R = 1 - \frac{U}{U_{max}} \qquad (1)$$

Redundancy is therefore not associated with a particular sign, or message*, but with a particular information system and its alphabet (or vocabulary, as the case may be). Shannon next attempted to characterize an information system by a Markov approximation of its source, using Markov's Law, a statistical law of information systems discovered earlier by the Russian semioticist, A.A. Markov. Markov's Law states that the higher the order of the Markov-chain approximating the source of an information system, the more the shape of the resulting symbols (or messages) look like the shape of the symbols (or messages) of the language from which the relative frequencies of the Markov-chain were compiled. A Markov-Chain is a stochastic process yielding a sequence of statistical outcomes each of whose probability distributions depends on a finite number of the previous outcomes in the sequence. For instance the occurence of the letters in printed American have a certain frequency distribution, but this distribution is radically altered when the letter is known to follow a given previous letter: the frequency of the letter 'h' rises dramatically following the letter 't' because of the prevalence of the word 'the' in American. The order of a Markov-chain is the number of previous outcomes the probability distribution of the next outcome is dependent upon, plus one.

Using this analysis, the known relative frequencies of monograms (letters), diagrams (letter-pairs), and trigrams (letter-triples), and purely logical reasoning, Shannon was able to show that the redundancy of an information system depends on the order of the Markov-chain approximating the source of that language, and was able to determine mathematical upper and lower bounds for this dependence. These bounds are given by Shannon's curve [11], Fig. 1.

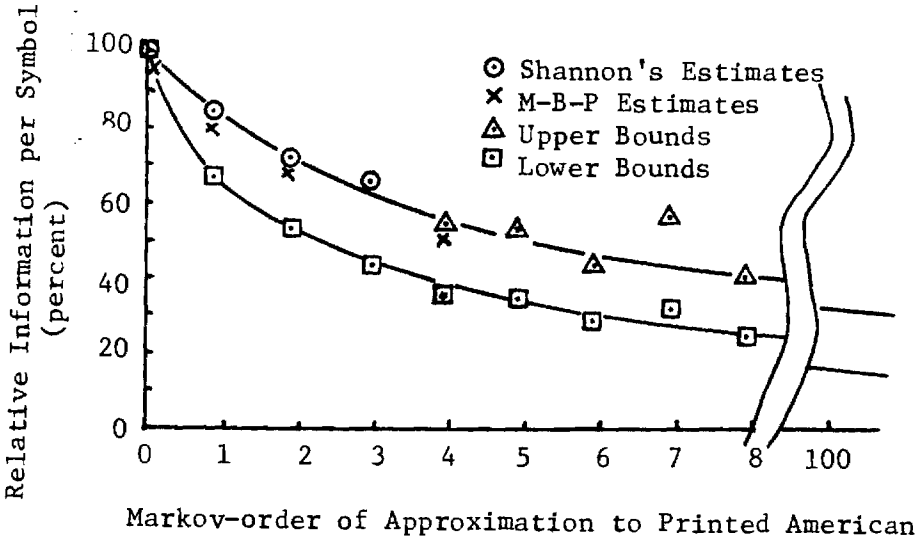


Fig. 1: Shannon's Redundancy Curve.

*A source of great confusion in the literature.

Altho Shannon established the upper and lower bounds of this curve mathematically, it has never been measured experimentally, and some reasons for this lack will be discussed later in the paper.

It should be mentioned at this point that despite the success of much statistical analysis in both computer technology and information technology, much emphasis in modern linguistics (a basic tool of both disciplines) has been concentrated on showing that natural language cannot be stochastic. This was first proved by Chomsky [1] by an argument that has since been repeated and improved several times. I have no quarrel with this proof as it has been refined (Chomsky's original proof is correct insofar as it is stated explicitly — he left a gap in the argument that was later filled in). The point to be made here is simply that altho natural language cannot be ultimately *merely* a stochastic process, nevertheless stochastic processes have many times proven to be excellent approximations of natural language based information systems.

It is time to turn now to some modern attempts to understand the nature of this curve and its measurement.

3. THE UNIVERSAL SIGN STRUCTURE THEORY:

Of the many extant theories of semiotics, all are nonquantitative, and extremely primitive. Only a few, in fact, are even relational in structure. Among these latter, the newest and least established is my Universal Sign Structure Theory, but it is this last one that has been most successful in its attempts to improve our understanding of the relations involved in computer technology and information technology.

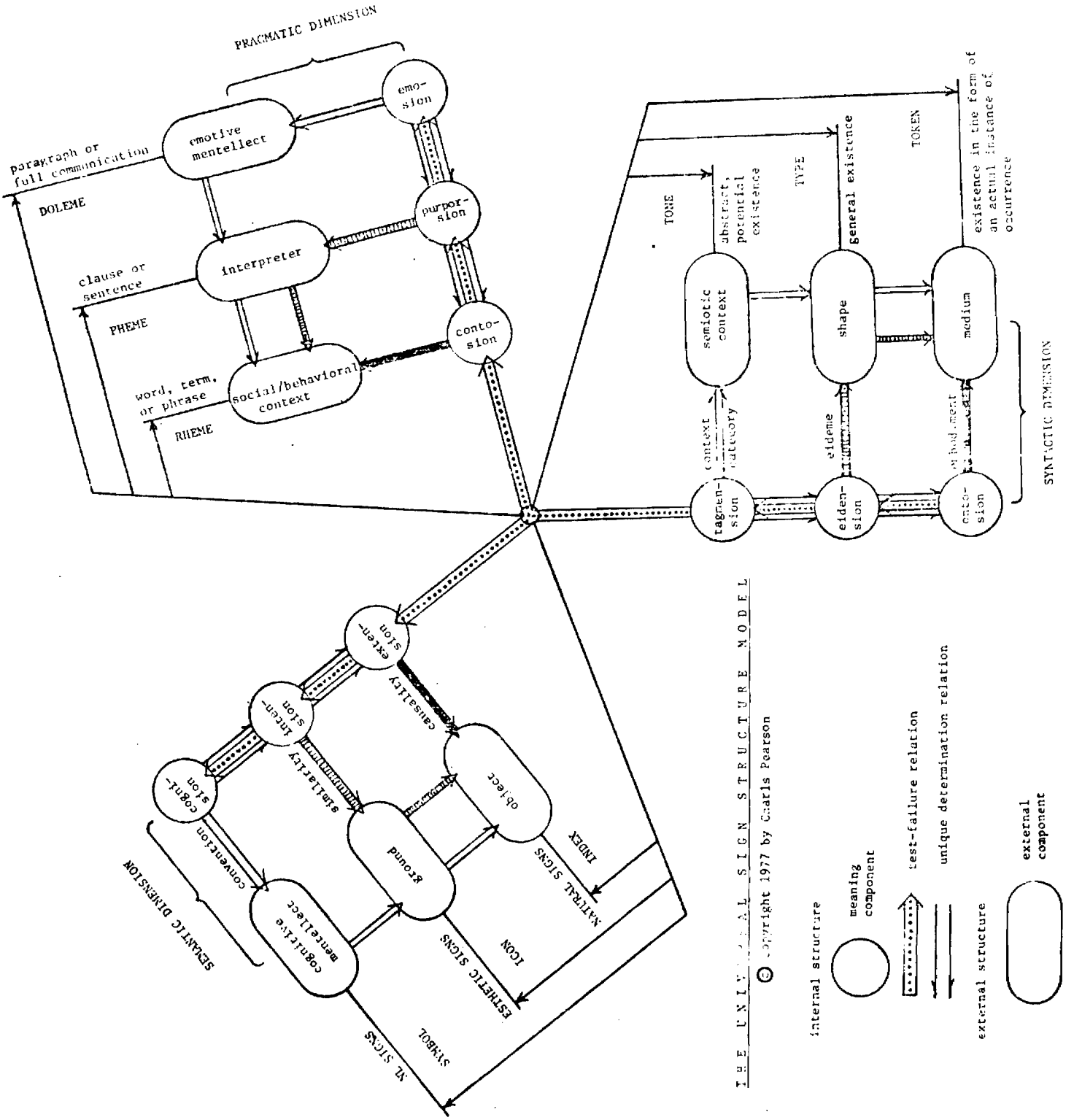
The theory involves a relational model, called the Universal Sign Structure Model, three theoretical principles, and a calculus. In [9] Pearson and Slamecka were able to prove nine representation theorems showing how Pearson's theory represneted the nine fundamental kinds of signs or messages. The work of [9] also led to the definition of 'meaning' as any one (or combination of more than one) of the nine internal components of the sign, and of 'information' as any empirically interesting observable aspect of any of the nine external components of the sign. This latter leads to the definition of 'information measure' as a mathematical model of any empirically consistent, aspect of information.

The Universal Sign Structure Model is shown in Fig. 2 and the three principles of the theory are:

1. The Trinarity Principle: *A sign must consist of a trinary relation.*

   To be consistent, therefore, the model has three parts, called the Syntactic Dimension, the Semantic Dimension, and the Pragmatic Dimension.

Fig. 2. The Universal Sign Structure Model



2. The Principle of Internal/External Balance: *The internal and the external structure of a sign must be balanced, consisting of exactly one internal component for each external component and vice versa.*

   The internal components are the components of meaning, while the external components are the generators of information.

3. The Principle of Additional Structure: *Whenever a sign has more than the minimum structure, the additional structure is built up from the center out (as per Figure 2), and for each dimension independently.*

Details of the theory can be found in [9].

Of most interest to us here is the component of 'shape'. The Universal Sign Structure Theory tells us that many present-day discussions of informaton, especially those dealing with communication systems and information storage systems, are dealing with observable aspects of shape.

The concept of shape in semiotics is a rigorously defined technical concept, but like all technical concepts, it is adumbrated within natural langauge by an intuitive concept. The technical concept of shape is an empirical explication of our ordinary intuitive concept of geometrical shape.

SHAPE is that by which two different sign types can be distinguished when both occur in the same syntactic context.

Thus O and X can be distinguished by their ordinary geometrical shape, while DOG and CAT are distinguished by their spelling so that in written information, orthographic shape refers to spelling. Thus the shape of messages inside a computer consists of the pattern of electrical pulses and no pulses on a wire*, or the pattern of magnetized spots on an oxide coated surface, etc. The shape of messages on card I/O is the configuration of holes and non-holes in the cards. An extreme, but elucidating, example of the technical concept of shape occurs in an information system consisting of marine flags. Here many of the signalling elements (the flags) have the same shape in our ordinary intuitive sense (geometrical shape) and the technical concept of shape refers to the color or color patterns of the flags.

With this detour into semiotics behind us we are now ready to analyze Shannon's redundancy curve.

4. SEMIOTIC ANALYSIS:

The Universal Sign Structure Theory explains that Shannon's curve describes a relation between the source of an information system and the syntactic context of that system. It says in effect that as the order of a Markov

---

*Often incorrectly called 'bits' and 'no-bits'.

information source increases, the relative comentropy* of that source decreases. This can be further interpreted by recalling that Markov's Law states that for Markov information sources based on natural language statistics, as the order of the source increases the shape of the symbols generated by that source look more and more like the shape of the symbols of the language on which the Markov frequencies are based. Thus Shannon's curve, together with Markov's Law, implies that for Markov information sources based on natural language statistics, as the shape of the symbols generated by that source look more and more like the shape of the symbols of the language on which the Markov frequencies are based, the relative comentropy of that source decreases. This conclusion relates the syntactic context of a Markov approximation to a natural language information system to the shape of the symbols in that system. How necessary are the assumption of Markov structure and natural langauge structure to this conclusion? In other words, is it possible that the same conclusion holds for all natural language information systems no matter what mathematical models we use to approximate them? or even for all information systems in general?

This question is an empirical problem in semiotics and its answer is beyond the scope of this short paper. However, a necessary first step in attempting to answer it requires a determination of the epistomological status of the two relations.

5. EPISTOMOLOGICAL ANALYSIS:

a) Introduction

We have seen that semiotic analysis shows that Markov's Law is a relation between the source of a natural language information system and the shape of the symbols generated in that system, and that Shannon's curve is a relation between the source of a natural language information system and the syntactic contect of that system. An epistomological analysis would ask further how we know these relations and what they mean. To ask how we know a subject is to ask how we gain knowledge about it and classically there are two broad, but extreme, answers to this question. We can gain knowledge empirically, and we can gain knowledge rationally**.

Each of these answers determines a different method of analyzing the meaning of a relationship. These two methods have no overlap so that one must proceed to determine the meaning either empirically or formally, but not by

---

*Also called 'entropy', 'negentropy', 'uncertainty', and most unfortunately, 'information'.

**'empirically" with small e and 'rationally' with small r. These are independent of any schools of philosophy such as Empiricism or Rationalsim, but which would also have to include many others, such as Pragmatism, and Logical Positivism, etc.

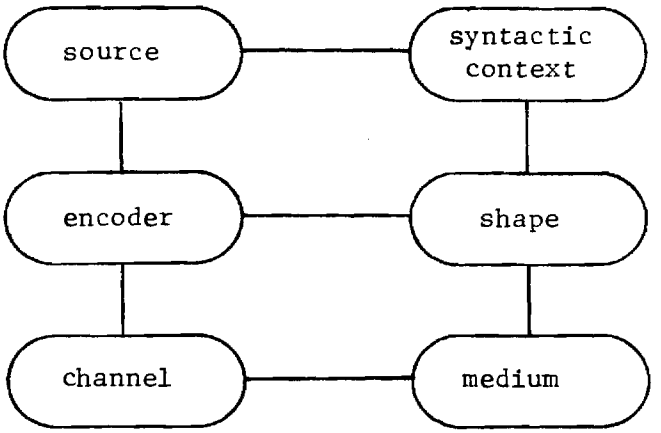


Fig. 3. Syntactic Structure of an Information System.

some compromise combination* of the two. The empirical method of analyzing the meaning of a relation always starts by asking how the variables entering into the relation are measured: "What are their operational definitions?". The formal method always determines meaning by asking for the formal definitions of the terms, the axioms which govern the system, and the theorems which relate the terms within the system of axioms.

Suppose we start our analysis by letting the variables tell us which method of gaining their knowledge must have been used. Each of the two relations involves two variables with either three or four variables being involved altogether depending on whether the Markov order variable in one relation is the same Markov order variable as the one in the other relation. We could begin by supposing that if both variables in one of the relations were formal variables then knowledge of the relation itself must have been gained formally and conversely if both variables in the relation had only empirical meaning then knowledge of the relation must have been gained empirically. However, if one of the variables in a relation is purely formal and the other is purely empirical then we must conclude either that knowledge of the relation can never be gained -- i.e., we can never know this relation because it cannot be grasped in a single unified act of knowing -- or that something is wrong in our expression of the relation.**

---

*This is not to claim that empirical procedures never contain any formal subprocedures as aids, or tools, they certainly do; but a given procedure itself is either completely empirical or completely formal.

**It is possible that what is wrong with our expression of the relation is that we have "borrowed" a mathematical concept for temporary use to control an important empirical variable for which we do not yet have a method of measurement. This "borrowing" occurs frequently in the method of semiotic reinterpretation and the paradigm inversion method.

b) Markov's Law

But the two variables in Markov's Law are the strangeness of the shape of the symbols generated in a natural language information system relative to a given natural language, and the order of a Markov approximation mathematically modeling the source of that information system. Now a natural language information system is an empirical thing. It exists objectively in the real world. It is one of nature's great laws that every local gathering of human beings in the world, past, present, or future, possesses a natural language independently of linguists, semioticsts, or information technicians. And the shape of the symbols generated in such a natural language (the phonemic patterning in spoken language) is an operationally observable variable. It was the great forte of American Structuralism to develop detailed manuals describing minutely the operational procedures for field recording the shape of the symbols observed in novel languages. Hence the shape of the symbols generated by natural language information systems can only have empirical meaning.

But a model of a source (as opposed to that source itself) is a mathematical entity, as is also a Markov-chain (a stochastic process) and the order of a Markov-chain (a pure integer -- not even a rational number). Hence the order of a Markov approximation mathematically modeling the source of an information system has only formal meaning -- it cannot be observed or measured, it can only enter into relationships with other formal entities. This forces us to conclude either that we can never gain knowledge of Markov's Law or that something is wrong in our expression of it.

But we do have knowledge of Markov's Law. It was discovered by A.A. Markov, described to us by him, and has been replicated in every attempt since then to verify this relation. Therefore, something must be wrong with our expression of Markov's Law because epistomologically speaking, in its present form, it is a bastard; it is the offspring of two parent parameters which cannot be legally united.

c) Shannon's Curve

The two variables in Shannon's Curve are the relative comentropy of a natural language information system and the order of a Markov approximation mathematically modeling the source of that information system. Again we have a natural language information system which is an empirical entity and the relative comentropy of the syntactic context of that system is an operationally observable variable, all we have to do is count symbol types and symbol tokens and construct frequencies. It was Estoup who first noticed the importance of doing this and carried out the first experimental observations. Of course, G.K. Zipf is best known for his methods of counting types and tokens, and it was Mandelbrot who first systematically examined the relationship between comentropy and type-token counting. Hence the relative comentropy of the syntactic context of a natural language information system can only have empirical meaning.

But as we have already seen, the order of a Markov-chain can have only formal meaning. Again we must conclude either that we can never gain knowledge of Shannon's Curve or that something is wrong in our expression of it. And also again, we do have knowledge of Shannon's Curve. It was discovered by him and published in [11]. It has since been verified independently by several other investigators, altho it has never been empirically measured as yet (but this is a different problem which will be discussed further a little later). It has evern been applied successfully to produce additonal discovereies in human information processing, as for instance by Miller-Bruner-Postman in [4]. Therefore as with Markov's Law there must be something wrong with our expression of Shannon's Curve.

d) Measurement

We saw that Markov's Law describes an empirical relationship. Knowledge of it was gained experientially -- Markov counted letters, digrams, and trigrams in *Eugene Onêgin*. The meaning of the law must therefore be determined by asking how we measure the strangeness of shape relative to a given natural language information system and how we measure the source of a natural language information system. Also Shannon's Curve describes an empirical relationship. Knowledge of it was gained by adding mathematical analysis to observed counts of letters, digrams, and trigrams in written American. The meaning of Shannon's Curve must be determined by asking how we measure the relative comentropy of the syntactic context of a natural language information source and also how we measure the source of that system.

The procedures for measuring the relative comentropy of a natural language information system are well known thruout the discipline. They are described in every text on "information theory" and several very sophisticated instruments are available to aid in these measurements [5,8].

Methods of measuring the source of a natural language information system are less well understood. As observed by the transformationalists, natural langauge cannot ultimately be *merely* a stochastic system, and perhaps what is apprximated by the Markov order of a source is the empirical degree to which a source can be approximated stochastically. There has so far as I know been no analysis yet made of this concept. However, if we notice that the two formal concepts used in the two relations as approximations are the same as far as their approximate nature allows us to determine we can assume an identify at the conceptual level and cancel them out in combining the two relations together in reaching a very useful result. The two approximate concepts both concern the order of a Markov-chain that approximates the source of a natural language information system. Our combined conclusion would then be that as the order of a Markov chain approximation to the source of a natural language information system varies so as to make the symbols generated look more and more like that of natural language, the relative comentropy of the syntactic context of that information system will decrease.*

*My Law of Word Interpretation, first announced in [7] indicates that this law holds generally and we do not need to restrict it to systems with Markov approximations.

Now the only problem left is to determine how to measure the strangeness of the shapes of words in a natural language information system relative to a given natural language. The solution to this problem was adumbrated by Miller-Bruner-Postman [4] in the paper previously cited and not surprisingly it concerns the use of Markov information sources. Miller-Bruner-Postman observe that they could not measure the shape of their words, but by using Markov generators, they had for the first time experimental controls on the degree of strangeness. The full solution to this problem was ultimately achieved with the concept of eidontic deviance* which measures on an instrument called an eidometer the strangeness of word shapes relative to a given language [6]. Using this instrument I was able to measure the Miller-Bruner-Postman Effect and thereby discover the Law of Word Interpretation which relates ease of experimental interpretation to degree of strangeness:

$$E_p = a + bS \tag{2}$$

where E_p is placement error in letters per word, S is eidontic deviance measured in °ED, and a and b are constants.

Eidontic deviance, the eidometer, and the Law of Word Interpretation are tools that are now available to reanalyze Shannon's Curve into completely empirical form and to measure it experimentally for the first time. This is now being done by my student Mr. Richard Lo [2].

6. SUMMARY:

We have seen that Shannon's Curve, and Markov's Law as well, are useful empirical relations altho their present expression is flawed epistomologically by employing one empirical and one formal coordinate each. The concept of eidontic deviance allows us to correct this in part by providing a quantitative empirical measurement of the strangeness of a word shape relative to normal word shapes in a given natural language. THis leads to the Law of Word Interpretation and this together with the Miller-Bruner-Postman Effect leads to the possibility of experimental verification and precise measurement for the first time of Shannon's curve.

7. ACKNOWLEDGEMENTS:

The research reported herein was supported in part by grant #IST-7827002 from the National Science Foundation, Division of Information Science and Technology. I would also like to acknowledge the help and support of my colleague and co-principal investigator, Professor Vladimir Slamecka, and my two research assistants, Richard Lo and Thanarak "Rak" Phongphatar.

---

*Literally "how much does the shape of a word deviate from the normal shape for a given language".

8. REFERENCES:

[1] Chomsky, N. "Three Models for the Description of Language". IRE Trans. on Inf. Theory, IT2(1956), p113-124.

[2] Lo, R. "The Measurement of Comentropy Transfer Rates". Proceedings of the Fifth Annual Conference of the Semiotic Society of America; Lubbock, Tex.; October, 1980.

[3] Markov, A.A. *"Essai D'une Recherche Statistique sur le Texte du Roman "Eugene Onegin",".* *Bull. de l'Academie Imperiale des Sciences de St. Petersbourg,* 7(1913).

[4] Miller, G.A.; Bruner, J.S.; and Postman, L. "Familiarity of Letter Sequences and Tachistoscopic Identification". Jour. Gen. Psychology, 50(1954), p129-139.

[5] Pearson, C. "Quantitative Investigations into the Type-Token Relation for Symbolic Rhemes". Proceedings of the Semiotic Society of America, 1(1976), p312-328.

[6] Pearson, C. "An Objective Concept of Word Shape". Paper presented at the Second Annual Conference of the Semiotic Society of America; Denver, October 15-16, 1977.

[7] Pearson, C. "A New Law of Information: an Empirical Regularity Between Word Shapes and Their Interpretation". Paper presented to the forty-first annual conference of the American Society of Information Science, New York; November, 1978.

[8] Pearson, C. "The Echelon Counter: A New Instrument for Measuring the Vocabulary Growth Rate and the Type-Token Relationship". Proceedings of the ASIS Annual Meeting, 17(1980), p364-366.

[9] Pearson, C.; and Slamecka, V. "A Theory of Sign Structure", Semiotic Scene, 1(1977), #2, p1-22.

[10] Shannon, C.E. "A Mathematical Theory of Communication". Bell System Tech. Jour., 27(1948), p379-423, and 623-656.

[11] Shannon, C.E. "Prediction and Entropy of Printed English". Bell System Tech. Jour., 30(1951), p50-64.

A METHOD FOR GENERATING HIGH ORDER MARKOV WORDS

By
Jeff Flowers
Charls Pearson
Thanarak 'Rak' Phongphatar

School of Information and Computer Science
Georgia Institute of Technology
Atlanta, 30332, USA

March 1981

Presented at the 19th Annual Southeast Regional ACM Conference; Atlanta, Georgia; March 27-29, 1981.

A METHOD FOR GENERATING HIGH ORDER MARKOV WORDS

By
Jeff Flowers
Charls Pearson
Thanarak 'Rak' Phongphatar

School of Information and Computer Science
Georgia Institute of Technology

ABSTRACT

The source of a natural language information system can be approximated by an nth order Markov information source in which words are generated by a stochastic process that selects letters according to a probability distribution that is determined by the actual outcome of the n-1 preceding letters. To approximate a real natural language, the probabilities used in the Markov information source must be based on the observed relative frequencies of that language.

Due to the size of the frequency tables that would be required for the observations, no computer based Markov information source for words of order greater than 3 has ever been built, altho generators for order 0 thru 3 are fairly common. A method has now been designed that sidetracks this problem. It will produce artificial words for all orders greater than 3, altho for order 7 or greater, all words observed to date have been real words. The first version of this method has been implemented on the CYBER 70 for written American words.

The method is based on an application of the Law of Zipf and Estoup which involves the rank-frequency distribution of natural language words. The computer algorithm, as implemented, dynamically builds all probability tables required while restricting the length of these tables using a rank-frequency table from a previous type-token analysis of natural language text. The structure of the algorithm will be discussed and examples of actual output words of various orders will be given.

A METHOD FOR GENERATING HIGH ORDER MARKOV WORDS

By
Jeff Flowers
Charls Pearson
Thanarak 'Rak' Phongphatar

School of Information and Computer Science
Georgia Institute of Technology

1. INTRODUCTION:

A stochastic process is the mathematical abstraction of an empirical process whose development is governed by probabilistic laws. A stochastic process* is defined to be a sequence of random variables $< X_i >$. Further a zero memory stochastic process is a stochastic process in which the probability distribution of a random variable X_i does not depend on the outcome of any of the previous random variables in the sequence [1].

When the Russian semioticist A.A. Markov** applied the concept of zero memory stochastic processes to his study of vowel/consonant alternation in Russian [2] he found that the model did not quite fit the facts, but yet the process of vowel/consonant alternation was more highly structured than required by simple stochastic processes. He found that the probability of occurrence of a vowel or consonant depended primarily only on whether the immediately preceding letter was a vowel or consonant. This observation did not hold for general letter processes, but Markov was still able to develop a simple mathematical concept that generalized the concept of zero memory stochastic processes and adequately modeled the results of vowel/consonant alternation studies as well as general letter sequence studies. Today this mathematical concept is called a Markov-chain and may represent the first concept of mathematics that was motivated by a requirement of the semiotic sciences.***

A Markov-chain of order m is a stochastic process in which the probability distribution for a random variable X_i depends only on the actual outcome of the m-1 previous random variables, but not on any outcomes previous to these. A zero-order Markov-chain by convention is a sequence of random variables each of whose outcomes is equally likely. By the above definitions, the probability distribution for the outcome of a random variable in a first-order Markov-chain, while not necessarily uniform or equally likely, does not depend on any previous actual outcomes, (depends on the outcome of the 0 previous random variables). Thus zero-memory

---

*Strictly speaking, a discrete parameter stochastic process.

**Better known in the United States as a mathematician

*** The golden age of mathematics of the 17th, 18th, and 19th centuries was stimulated mainly by the requirement for models by the physical sciences.

stochastic processes consist of zero- and first-order Markov-chains; while vowel/consonant alternation in natural language (not just Russian) can be modeled mathematically by second-order Markov-chains.

Markov information sources are information sources whose statistical properties can be modeled by finite Markov-chains. They thus consist of either a finite alphabet, or a finite vocabulary, plus a transition matrix which determines the transition probability from each state consisting of the m-1 previous outcomes to each possible new state consisting of the m-2 previous outcomes plus the occurrence of the next letter, or symbol.

We have already seen that Markov information sources are useful for modeling vowel/consonant alternation phenomena. It is general knowledge that Markov information sources are quite useful for approximating general information systems. It is a significant result that all natural language based information systems can be approximated by Markov information sources. In fact, Markov's Law states that the higher the order of a Markov information source modeling a natural language, the better the shape of the symbols generated by that source approximates the normal shpe of the symbols in the language from which the statistics for the Markov-chain were compiled.

2. MARKOV WORD GENERATORS:

A Markov word generator is a natural-language based Markov information source that generates artificial words that approximate a particular natural language. In order to approximate a real natural language, the probabilities used in the Markov information source must be based on the observed relative frequencies of that language. Due to the size of the frequency tables that would be required for the observations, no computer based Markov information source for words of order greater than 3 has ever been produced by this method, altho generators for order 0 thru 3 are fairly common. For example written American requires 29 letters: 'A' - 'Z', hyphen, apostrophe, and blank, and in order to produce mth order approximations, 29^m frequencies are required. Altho many of these frequencies are zero (e.g., no QQQ's occur in written American), no memory structure has ever been designed to take advantage of this sparseness to reduce the size of memory required for high order word generation.

It is desirable, however, for many purposes involving the design of information systems, and especially for experiments in the Georgia Tech SemLab investigating the syntactic structure of information, to produce words of high Markov-order. A method has now been designed that avoids the above problem. It will produce artificial words for all orders greater than 3, altho for order 7 or greater, all words observed to date have been real words. The first version of this method has been implemented for written American words, using a Pascal algorithm for the CYBER 70.

3. CLASSICAL METHOD:

Typical Markov word generators use polygram frequency tables as input. One such program is WORDGEN used by the Georgia Tech SemLab. WORDGEN has a

structure represented by Fig. 1. WORDGEN uses a static polygram table that has previously been built in a readable format by another program, GRAMS.

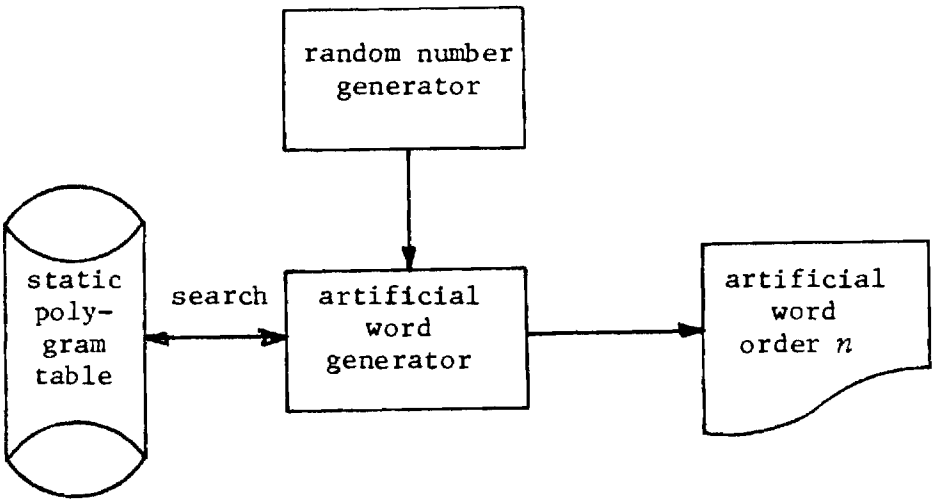


Fig. 1. Typical Artificial Word Generator of Order n.

WORDGEN obtains a random number from the random number generator, selects the next letter from the polygram table based on the m-1 preceding letters already generated, and adds this letter to the output string. It continues until a blank is generated which terminates the process and the word is output for printing.

WORDGEN uses $29^3 + 29^2 + 29$ frequencies or slightly less than 28,000 entries in the static polygram table, and generates words of orders 0, 1, 2, and 3 relative to written American. If WORDGEN were to generate words of order 6, it would require more than 700 million entries in the polygram table.

4. ALTERNATE METHODS:

Shannon [3] proposed an alternative to the polygram method of generating Markov approximations. He suggested that the required frequencies are self-contained in the text and that one could simply load a sufficiently large sample of text into memory, generate a random number, and find the first m-1 letter string following the location given by the random number. One would then search the text for the next occurrence of the last m-1 letters of this string and append the next letter. This process continues until a blank is obtained.

This method also has problems for computer implementation. For a piece of text like the BROWN Corpus which contains a million words, approximately

2 million words of memory are required. Altho this is smaller than the 700 million words required for the polygram example given earlier, it is still large as core sizes go. In addition, this method is very slow, since on the average, most of the text is searched each time a letter is generated.

Our new method adopts Shannon's basic suggestion but modifies it to reduce the storage requirements and increase the processing speed.

The method is based on an application of the Law of Zipf and Estoup which involves the rank-frequency distribution of natural language words and holophrases. The algorithm for this method, as implemented, dynamically builds all polygram tables (but only those portions actually required at the time) using a rank-frequency table from a previous type-token analysis of natural language text.

The rank-frequency table consists of

1. Rank of the word (in rank order) according to the frequency with which that word occurs in the text
2. Occurrence frequency of the word in the text
3. Word

as shown in Fig. 2.

Rank	Freq.	Word
1	6387	THE
2	2861	OF
3	2191	AND
⋮	⋮	⋮
n	1	DISHEARTENING

Fig. 2. Rank Frequency Table.

The space required for this algorithm consists of this rank-frequency table plus two vectors each of length n associated with the table. Since n was 7,000 in our first implementation, this gave us a memory requirement of 4*7,000 = 28,000 words, or approximately the same size as the polygram method.

The key to the effectiveness of the method lies in the nature of the Law of Zipf and Estoup which says that in natural language text a few words

occur a large number of times while most words occur very few times. The relation between the rank and frequency is given fairly accurately* by

$$F = \frac{C}{R} \tag{1}$$

The algorithm essentially replaces all the occurrences of a word by a single instance and uses this to compute all the required frequencies. In one version of WORDGEN3, the frequency vector itself is eliminated and Eq. 1 used to approximate the frequency.

The structure of WORDGEN3 is shown in fig. 3. It is similar to that of WORDGEN except that the static polygram table is replaced by three components: 1) a rank-frequency table; 2) a probability calculation; and 3) a dynamic polygram table.

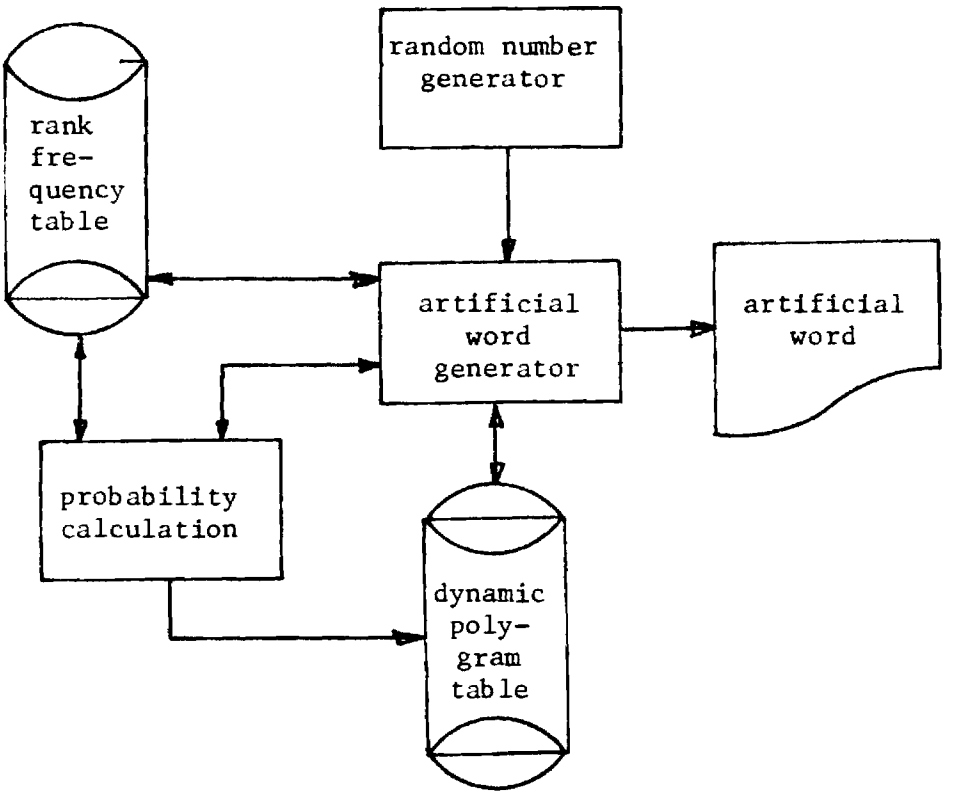


Fig. 3. Structure of WORDGEN3.

*More accurate forms of this equation are known and useful in certain applications, but this form is sufficient for our algorithm.

5. WORDGEN3 ALGORITHM:

The algorithm is given below in Pascal.

Algorithm 1. Build the probability table.

```
(*  We will use the following vocabulary:

    word(i)       :  ith word in the Rank Frequency Table
    freq(i)       :  frequency of word(i) from the Rank Frequency Table
    length(i)     :  length of word(i)
    occurrence(i) :  number of times the n-1 substring of preceding generated
                       letters occurred in word(i).                    *)
```

The algorithm for building the probability table depends on the Markov order m.

Case 1. order m = 1

```
prob[1] := length[1]*freq[1]   ; and
prob[i] := prob[i-1] + length[i]*freq[i]
```

Case 2. order m > 1

Case 2.1. Start by choosing the first n-1 letters.

```
prob[1] := freq[1] ; and
prob[i]  := prob[i-1] + freq[i]
```

Case 2.2. prob based on n-1 preceding letters

```
prob[1]  := freq[1]*occurrence[1] ; and
prob[i]  := prob[i-1] + freq[i]*occurrence[i]
```

Algorithm 2. Artificial Word Generator.

```
(*  no : temporary
    art_word : generated word
    order : Markov order
    loc : location in the generated word
    maxlength : maximum length of word to be generated
    no_word : number of words to be generated
    required_word : required words to generate
    index : index of the word to be selected
    random : random number from random generator                    *)

    begin
         no_word := 0 ;
         repeat
              loc := order-1;
              if loc = 0 then loc := 1;
              repeat
                   build_prob;  (* according to the Markov order *)
                   get-random (random);
```

```
            (*   this random rounded up to
                      1 ≤ random ≤ prob (length-of-prob)      *)
                 search_prob(index);
                 if (loc < order-1) and (order <> 1) then
                 (*   move first m-1 letter to art_word   *)
                 move (word[index], art_word,1,1,order-1)
                 else
                      begin
                           if index = 1 then
                                no := mod (random, freq[1])
                           else
                                no := mod ((random - prob[index-1],freq[index]);
                           if order = 1 then
                                move (word[index],art_word,no,loc,1)
                           else
                                begin
                                     no := search(word[index],art_word,loc,
                                                  order-1,no);
                                     move (word[index],art_word,no,loc,1);
                                end;
                      end;
                 loc := loc + 1;
            until (ichar (art_word,loc-1) = blank) or loc > max length;
            print (art_word);
            no_word := no_word +1;
       until no_word > required_word;
       end.
```

- procedure move (string1; string; var string2: string; loc_in_string1, loc_in_string2, number-of-letter-move: integer);

```
(*   move letters at location loc-in-string 1st in string1 to
     string2, at location loc-in-string 2nd,
     number-of-letter-mover letters is moved.                    *)
```

- procedure build_prob;

```
(*   probability tables are built according to algorithm 1.      *)
```

- procedure search_prob (var index: integer);

```
(*   search probability table until
     prob [index-1]   random ≤ prob [index]
```

- procedure get_random (var random: integer);

```
(*   generate random number
          1 ≤ random ≤ prob [length-of-prob]                     *)

     function search (string1, string2: string; loc, order, no: integer):
     integer;
```

```
(*  search substring of string2 from loc-order to loc-1 until m occurrences
    are found and return the next location in string1.                  *)

    function ichar (string1: string; loc: integer): char;

(*  return the letter at location loc^th of string1)                    *)
```

From the algorithm described above, the probability table is rebuilt every time a letter is selected. The first m-1 letters ($m > 1$) are selected according to the frequency of the word in the text, and the next letter will be selected randomly based on the probability of that letter occurring following the preceding m-1 letters in the text (as can be seen from algorithm 1, case 2.2).

For order $m = 1$ the letter will be randomly selected from the word according to the frequency with which the letter occurs in the language. (algorithm 1, case 1).

6. DISCUSSION:

So far as we are aware, this is the world's first implementation of a high order Markov word generator. It is now available and can be requested from the Georgia Tech SemLab.* This version is written in CDC Fortran Extension IV and takes advantage of the CYBER 70 architecture. For instance, the CYBER 70 has a word length of 60 binits and uses 6 binits to present each letter; therefore 10 letters can be stored in each word of core. The initial test run used the first 7,000 word-types (wt.) from the output of TTKANAL which performed a rank-frequency analysis of the entire A file of the BROWN Corpus, consisting of the first 44 samples A01 thru A44. The A file contains 11,876 wt. and 88,000 wk. Run time for the initial test run was approximately 3 words per minute; however, this has been speeded up and is now about 12 words per minute.

All words produced by WORDGEN3 so far (about 4 weeks of experience) of order 7 or higher have been real words of written American. This may be due to the small size of the text (88,000 wk.), but also says something about the statistical structure of language, altho it is difficult to interpret the meaning of this. It reinforces Markov's Law.

The process has several disadvantages. It cannot generate order 0 words at all, and for orders 1 thru 3, it is more expensive than the classical polygram methods of word generation.

7. EXAMPLES:

The following examples from orders 0 thru 3 are from WORDGEN, a classical Markov generator, while the examples for orders 4 thru 6 are from WORDGEN3, using our new method. Markov's Law is evident in these examples. As the order of approximation increases, the words look progressively more and more like written American.

---

*Address inquiries to Dr. Charls Pearson.

Markov Order 0

DUSMN
SNVOUUGCFU'GE'SKRZ
CXWBIP'YDTPHWYIZO-I'ZBVPLFEIFKRQK'QMQ'CVVXSOW
BWRRW
WDNBS
W
DWXPJTNJV
LITSQVXLACRO
ASFHUJAUWSIJGWSYJXNVVGNKICMOMWJMRVUGXHTSVAJLNRBDZX-TLV'QM'SD
QXGMKHCJDGDCWZEUWZW-DY-QY'JVHAPMOXQ
'KQFLZCDJBFYIJK'KABNCS'

Markov Order 1

ERMUNSIAT
E
HNTASEMCTDM
GTE
ELLANLEACOIIETLFND
SEIAIEOA
REDGA
ECPO
DAI
TL
ASHRUSHLLIEYS
SULVQ

Markov Order 2

LIRD
MBLD
TOTHANE
NIGHINAT
INESPORE
CHANS
HATHEMIC
REST
INDAT
KNIAY
HENS
HECAREND

Markov Order 3

WHING
YOUGHT
GROT
AMENOT
INDATENTEAT
RACTED
LOWN
DIUMPLENT
TIONCER'S
WHADERS

Markov Order 4

THROUGHTER
PEDESTRIBUTIONS
MATTENDENT
PROGRAPER
SCHOOLESE
PRESEARCH
DESPITAL
OCTOR
VICKETS
GALLET

Markov Order 5

LEVELYN
SINGULATE
WILLIAC
LEGISTER
COLLECTURE
DETONAL
RESTROUS
MUSIALED

Markov Order 6

PURCHANDIZE
PERHAPSODY
GOVERNIGHT
MATTERDAY
NORMANCE
AUTHORIZON
PERFORMALIZE
GROUNDERNEATH

8. SUMMARY:

Our new method makes it possible for the first time to generate high order Markov words by computer. Altho the longest word in the BROWN Corpus is 31 letters which indicates an upper bound of Markov-order 30, above which all words generated will be real words actually used in the BROWN Corpus, practical experience to date indicates all words of order 7 or greater are real words.

9. ACKNOWLEDGEMENTS:

This work was supported in part by grant #IST-7827002 from the National Science Foundation, Division of Information Science and Technology. We would also like to thank our colleague Vladimir Slamecka for his continuing support and encouragement.

10. REFERENCES

[1] Allen, A.O. Probability, Statistics, and Queueing Theory with Computer Science Applications. Academic Press, 1978, p113-145.

[2] Markov, A.A. *"Essai D'une Recherche Statistique sur le Text du Roman "Eugene Onegin",". Bull. de l'Academie Imperiale des Sciences de St. Petersbourg,* 7(1913).

[3] Shannon, C.E. "A Mathematical Theory of Communication". Bell System Tech. Jour., 27(1948), p379-423, and 623-656.

THE SEMIOTIC PARADIGM

By Charls Pearson
School of Information and Computer Science
Georgia Institute of Technology
Atlanta, 30332, USA

June 1981

THE SEMIOTIC PARADIGM

By Charls Pearson

ABSTRACT

Thomas Kuhn defines a paradigm as the language, models, theories, methodology, decisions as to important problems, in short the total *Weltanschauung*, by which a science carries on its daily activities. All sciences start out in a preparadigm, groping stage in which no useful paradigm has as yet been identified. Information Science has often been described as still in this preparadigm, prescientific stage. However, several investigators, including the author, have adopted a very powerful point-of-view, called the Semiotic Paradigm which is nothing less than a total scientific paradigm. Its background, structure, use, and applications are examined in this paper.

The Semiotic Paradigm contains a language, theory, experimental methodology, point-of-view, models, decisions on important problems, etc. and is nothing short of a *Weltanschauung*. As such it would be impossible to completely describe it in minute detail in the confines of this short paper. Hence the paper is concise, often to the point of being cryptic, and many references are made to the author's previous papers. However, since most of you have heard or read many of these previous papers covering many details of the system, this paper has as its main purpose, the establishment of the big picture in a broad perspective and to establish an overview that allows later connection to the very many details involved.

The Semiotic Paradigm does not stand alone as a candidate for the scientific paradigm of information science. Several other partial paradigms have also recently been suggested for this role. However, among them, only the Semiotic Paradigm stands as a complete, total, scientific paradigm, and the Semiotic Paradigm stands by itself where scientific paradigms really count -- in the productivity of further knowledge, theories, experiments and, yes, even applications.

THE SEMIOTIC PARADIGM

By Charls Pearson
School of Information and Computer Science
Georgia Institute of Technology
Atlanta, 30332, USA

1. ABSTRACT:

Information Science has often been described as being in the preparadigm, prescientific stage. However, several investigators, including the author, have adopted a very powerful point-of-view, called the Semiotic Paradigm which is nothing less than a total scientific paradigm. Its background, structure, use, and applications are examined in this paper.

The Semiotic Paradigm contains a language, theory, experimental methodology, point-of-view, models, decisions on important problems, etc. and is a complete *Weltanschauung*. As such it would be impossible to completely describe it in minute detail in the confines of this short paper. Hence the paper is concise, often to the point of being cryptic, and many references are made to the author's previous papers.

Among several other partial paradigms that have recently been suggested as candidate paradigms for information science, only the Semiotic Paradigm stands as a complete, total, scientific paradigm, and it stands by itself where scientific paradigms really count -- in its productivity of further knowledge, theories, experiments and, yes, even applications.

2. BACKGROUND:

In studies leading back more than ten years it had become evident by the early seventies that sign phenomenas constitute the single, all pervasive phenomenas of information science, and in an investigation carried out for the National Science Foundation in the mid-seventies it was determined more specifically that every problem in information science involves the investigation of some aspect of the structure of signs [17;18;20]. In my lecture before the 1979 Annual ASIS Conference in Minneapolis -- which unfortunately did not appear in the published proceedings of that meeting -- I outlined the various attributes of information and how each one could be interpreted as an observable attribute of signs. Last year in one of my lectures before this body [14], I demonstrated that the sign concept is the fundamental concept of all information science by showing how each concept of information science can be defined in terms of the sign and its ancillary concepts.

So even if it was with a great deal of naive enthusiasm when I started my doctoral research in the early seventies [5] that it seemed obvious to me that semiotics constituted the fundamental paradigm of information science, it is now with a great deal of hindsight that I say that it *is* obvious.

3, INTRODUCTION:

The notion of scientific paradigms was discussed by Kuhn in [1] where he used the concept to analyze scientific revolutions. For his purposes it was not necessary to further classify paradigms into subcategories. However, in order to analyze the foundations of information science and to direct progress in fundamental research as effectively as possible, it is necessary to gain a clearer understanding of the kinds of paradigm components involved in scientific paradigms, and in [15] I introduced a classification into five categories, four of which are always present in any scientific paradigm and a fifth which is often present. These categories are: 1) linguistic, conceptual, philosophical paradigms; 2) theoretical paradigms; 3) experimental paradigms; 4) mathematical paradigms; and hte optional one 5) applicational paradigms. All five categories were motivated, explicated, and exemplified in [15] which the interested reader may refer to for details. The Semiotic Paradigm has component paradigms from each category and these will be examined briefly in the following sections altho I will not treat them equally. I have given more space to those areas -- theory, measurement, and experiment -- that I feel are more important for the purposes of today's presentation. The linguistic, conceptual, and philosophical components of the Semiotic Paradigm are examined in section 4. THE LANGUAGE OF MENETICS. Section 5, THE UNIVERSAL SIGN STRUCTURE THEORY, deals with the theory component of the Semiotic Paradigm. Section 6 examines THE NATURE OF EXPERIMENTS IN INFORMATION SCIENCE/SEMIOTICS, and THE MEASUREMENT OF INFORMATION is discussed in section 7. Section 8 expands on sections 6 and 7 by presenting some LAWS OF INFORMATION SCIENCE/SEMIOTICS and section 9 expands on section 4 and prepares the way for section 10 by resolving a terminological ambiguity. It contains the AMBIGUITY OF "INFORMATION" MEASURES. Section 10 is a short section which mentions MATHEMATICAL METHODS FOR INFORMATION SCIENCE. The APPLICATIONS OF INFORMATION SCIENCE/SEMIOTICS is discussed in section 11 under the three categories of Refinements to Scientific Language and Theory; Semiotic Engineering; and Industrial Management. Section 12, OMISSIONS, is an apology for topics not covered, mostly due to lack of space, and section 13 contains some suggestions for the FUTURE OF INFORMATION SCIENCE/SEMIOTICS RESEARCH.

On objection to the Semiotic Paradigm can be adequately defended in this introduction. It concerns the present status of inquiry in both semiotics and information science, neither of which is very basic or has anything to do with science. Current semiotic writing is predominantly speculative and humanistic with most efforts concerned with possible applications to such areas as literary criticism, classics, architecture, etc. While the current information science literature is more quantitative and less humanistic, it is also speculative and dwells on the applications in the areas of information engineering and industrial management. For this reason, I have tried to capture the fundamental relations involved in the basic science with the aphorism $(IS)^3$ which stands for:

Information Science IS Instrumentation + Semiotics

It should be clear that what I am trying to do with this aphorism is to separate off the speculative and humanistic aspects from semiotics and the speculative and applied aspects from information science and identify the

resulting basic sciences. It is this basic science to which the Semiotic Paradigm applies, not the speculative, humanistic, or applied aspects, altho I think that even in these areas too the Semiotic Paradigm will prove of use.

4. THE LANGUAGE OF MENETICS:

The linguistic, conceptual, philosophical component of a scientific paradigm establishes the framework and point-of-view in which the scientist does his thinking and carries out his analysis. For further details concerning these paradigms, see [15]. The linguistic, conceptual, philosophic component of the Semiotic Paradigm is the Language of Menetics. It is founded on a metaphysics involving an atomistic approach to the carriers of information. In this metaphysics of structural atomism information is regarded as being carried by messages which are systems of one or more signs. Inasmuch as semiotics is the paradigm trinary science and the sign is the fundamental relation of semiotics, it is not surprising that semiotics provides the paradigm for information science. Details of the metaphysical presuppositions of the Language of Menetics are spelled out in [14].

The Language of Menetics was created in a deliberate attempt to design a language that was adequate in all three of Chomsky's senses for talking about information and meaning in empirically testable ways. It recognizes the fact that most talk of information and meaning is non-empirical and non-testable even tho such talk often grapples with important, substantive aspects of this topic, and that such partial languages as do exist for these topics (psychology, logic, linguistics, etc.) are highly restricted in that only a few aspects of meaning and information can be discussed in any one language and that few translation rules exist for translating between languages in the few cases where the same topic can be discussed in more than one language.

The Language of Menetics is universal in the sense that it is able to talk empirically about any aspect of information and meaning that can be discussed in any other language, and such discussion is always empirically testable even tho the language itself is always deliberately vague where empirical explication does not yet exist. The language has been shown to meet all three of Chomsky's levels of adequacy. That is, it is observationally adequate, descriptively adequate, and explanatorily adequate. By being able to discuss data, laws, and theories from different phenomenas and different disciplines empirically within one unified, integrated, and systematic language for the first time, the relationships between various aspects of meaning and information phenomenas can be seen. Not surprisingly, some of the phenomenas turn out to be the same as other phenomenas, but discussed in a different one of the many narrow languages used previously.

The Language of Menetics includes a technical, but systematic and transparent terminology, a grammar, a semiotic point-of-view, a metaphysics of structural atomism, a decision as to what kinds of problems are important for the study of meaning and information, what kinds of phenomenas are important for understanding these problems, and what kinds of methods are useful for analyzing these phenomenas for the purpose of solving the problems of choice. Each of these aspects is discussed in detail and the total language is presented systematically in [5].

5. THE UNIVERSAL SIGN STRUCTURE THEORY:

The theoretical component of a scientific paradigm embodies the abstract principles by which all scientific explanation takes place in a unified and disciplined way. In the popular mind the theoretical component of a scientific paradigm is often equated with the scientific paradigm itself, altho this is never the case technically. For further details concerning theoretical paradigms, see [15].

One of the most powerful and useful theories developed within the Language of Menetics has been the Universal Sign Structure Theory, the theoretical component of the Semiotic Paradigm. It is a very natural development within that language following in very obvious ways just as the Geocentric Theory followed at once in an obvious way from Ptolmy's Language of Orbits. The Universal Sign Structure Theory is a relational theory rather than a quantitative theory as are many theories of the physical sciences, and yet it is very powerful. It predicts or explains all of the relationships observed among meaning or information phenomenas and suggests ways they can be empirically explicated to upgrade their structure to that of scientific laws. This was the means by which I was motivated to discover my Law of Word Interpretation.

The theory contains three abstract principles, a relational model (equivalent to a digraph), a means for manipulating the principles, and rules for interpreting the resulting theorems in terms of observational concepts. The relations between various kinds of signs can be captured in terms of nine representation theorems. The relations between various aspects of information and meaning can be read directly from the relational model. Information is defined as any observable aspect of the external structure of a sign and meaning is the theoretical aspect of the sign. This explains the close relationship between meaning and information.

There is not room enough here to present any of the details of the theory or to show its usefulness other than to suggest a few examples such as its use to explain the syntactic structure of Shannon's Information Theory, or its use to explain the pragmatic structure of sentential information. However, the theory has been given in parts in various papers along with examples of its application. The interested investigator may refer to [4;5;6;7;8;9;10;11;12;13;14;15;16;17;18;19;20] for the details.

6. THE NATURE OF EXPERIMENTS IN INFORMATION SCIENCE/SEMIOTICS

The experimental component of a scientific paradigm provides the methodology for observing that part of nature that falls within the purview of a scientific discipline. It is made up of various task paradigms, measurement methodology, methods of data analysis, etc.

The experimental component of the Semiotic Paradigm is derived from all of the semiotic sciences -- but especially experimental psychology -- via the Paradigm Inversion Principle [7;11;13;15].

The Principle of Paradigm Inversion allows us to take advantage of any experimental paradigm in any of the semiotic sciences, some of which are much more developed as experimental sciences than information science/ semiotics, and use this to design information science experiments and make information measurements. The Principle of Semiotic Reinterpretation then shows us the empirical foundations of such measurements and shows us how to quantify them when they prove useful for further empirical research.

We do not lose the distinction between information science and the semiotic sciences when we do this. In fact, we can use the Principle of Paradigm Inversion to clarify the distinction between them. For instance, psychology uses the known structure of information to examine the structure of behavior while semiotics uses the known structure of behavior to examine the structure of information. We can also use the Paradigm Inversion Principle to determine how to measure the ephemeral semiotic attributes of information in terms of concrete properties of behavior evidenced in some one of the semiotic sciences such as psychology. The principle indicates the existence of a relationship between semiotics and each of the information sciences. It may be used to examine the details of this relationship but this has not as yet been done.

The Paradigm Inversion Principle was motivated in [6;7;11;13;15]. It postulates an open experimental structure which is the same for all empirical sciences, such as physics, psychology, and information science. The differences lie in how this open structure is completed and the use that is made of it. These details are developed in [11].

The Paradigm Inversion Principle may be criticized for its use of a standard set of interpreters or behavioral responders, viz, human beings. This criticism states that humans are too subjective or too variable, to use as measurement standards. This statement just does not hold up. The use of a fixed set of observers as a measurement standard is no more variable than the use of a bar of only one length as a length standard and is just as complete. We get entirely different concepts of temperature (not just different units) if we change our standard thermometer from alcohol to water, and mercury gives yet another. It took almost two centuries of experimentation to arrive at the theory which incorporated the abstract conception of termperature involving ideal gasses. It will also make many experiments before we arrive at the proper idealizations to replace concrete interpreters in information science but in the meantime experimentation and measurement are both possible and necessary.

7. THE MEASUREMENT OF INFORMATION:

Since the same experimental paradigms generate experiments in both information science and the semiotic sciences depending on the type of measurement involved, a question is raised concerning the nature of semiotic measurement. This question has two parts. The first is, what is the nature of individual information measures? how can their empirical foundation be determined? and how can their usefulness or lack of usefulness be explained?

The second is, what is the nature of information measurements in general? How are they similar to physical, psychological, and other scientific measurements? and what distinguishes them from these other kinds of measurement?

The first question was answered in [11] using the Principle of Semiotic Reinterpretation in conjunction with classical measurement theory to explicate the nature of individual semiotic measures. The second question is answered in this paper for the first time and therefore represents a definite advance in the development of the Semiotic Paradigm. It relies on the Paradigm Inversion Principle and a refinement to the concept of experiment presented in the last section to incorporate the concepts of instrument and observer. Inasmuch as this is a first examination of this question, details are not yet completely worked out.

The principal thesis of Semiotic Reinterpretation is that all empirically useful information measures can be reinterpreted not simply as a measurement of some external semiotic property, but as an empirical generalization, a natural law, stating an observable regularity in the way nature behaves. The natural law relates two measures together, one of which is the information measure to be reinterpreted. The second measure may be a psychological measure, a physiological measure, a physical measure, or any other scientifically important kind of measure, including even another semiotic measure.

Explicitly, the Principle of Semiotic Reinterpretation has two theses: 1) al definitions of information measures can be interpreted as measures (in the measurement theoretical sense) of external properties of signs, systems of signs, or sign processes; and 2) the definitions of all *useful* information measures can be *reinterpreted* as a natural law describing a regularity between a semiotic measure and some other measure. It is then this natural law which gives the empirical foundation for the usefulness of such measures.

Once such a concept has proven empirically useful, the Principle of Semiotic Reinterpretation shows us how to quantify it in appropriately useful ways using classical scientific methods. The details of these developments are presented in [11].

In order to consider the general nature of information measurements let us look at the structure of a typical experiment in any science, be it physical, semiotic, or information science. It will look like that shown in fig. 1.

For instance, a physical scientist would call S the specimen to be measured, a psychologist (a semiotic scientist) would call S the stimulus, a semioticist would call S a sign, and an information scientist would call S a signal, or message (something containing information. P is the experimental design. In the physical sciences it is the design of the apparatus of the experiment; in the semiotic sciences it is the design of the task. I is an instrument. In the physical sciences it would be a physical instrument, such as a galvanometer, in psychology it may be either a physical instrument such as a stop watch or a semiotic instrument such as a word

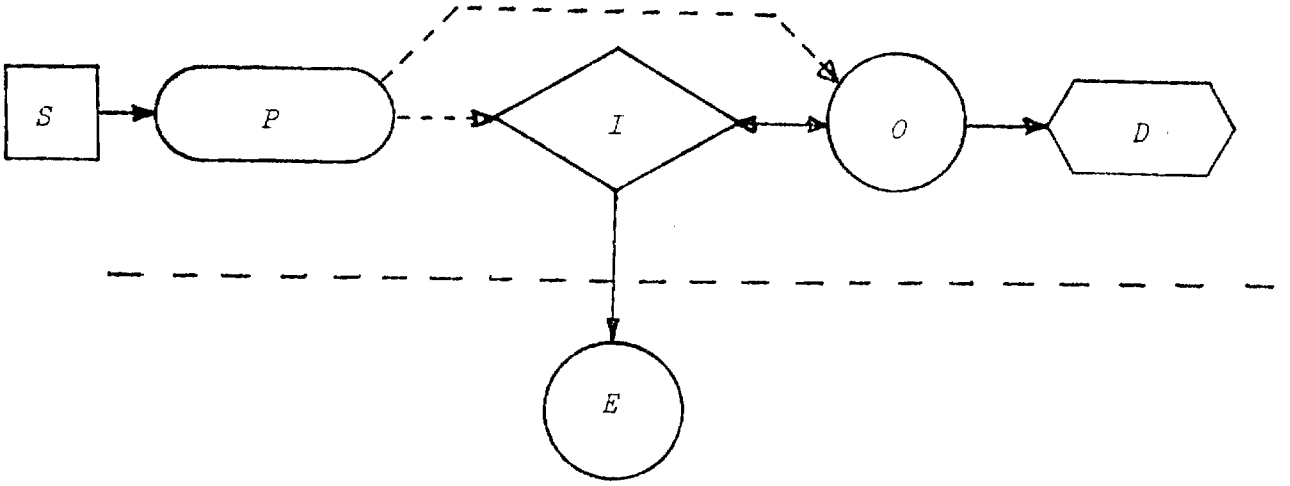


Fig. 1. Intuitive Structure of a Scientific Experiment.

recognition score, in information science it would be a semiotic instrument such as an eidometer. *O* is an observer. He is an integral part of the experimental paradigm despite the fact that the physical sciences very often completely overlook this aspect of the experiment. Quantum mechanics is today paying the penalty for this carelessness. However, in ordinary circumstances, the physical sciences are justified in ignoring this aspect of their experiments because the experimenter *E* usually acts as his own observer and has access to all of the data of the experiment in the same form as *O* has it. In information science and all the semiotic sciences this is not the case. There is a sharp distinction between the data available to *O* and that available to *E* and so *O* is a necessary part of the semiotic experiment. *D* is the data of the experiment. It is read from the instrument *I* and the experimental setup *P* and written down in symbolic form. This last step has a big effect in the interpretation of an information science experiment as will be seen shortly. *E* is the experimenter. He designs the experiment, administers *S*, and monitors *I*, *O*, and *D*. In the physical sciences he very often is the same individual as *O*, thus engendering the confusion mentioned above. *E* thinks of himself as both experimenter and observer and then mistakenly identifies the two concepts.

During the measurement process of a physical experiment, an indexical relation exists between the experimental paradigm *P* and the instrument *I*; this causes a displacement of the instrument. An iconic relation between the displaced pointer position and the undisplaced pointer position establishes the size of the effect and, since a suitable numeric scale is coordinated to the pointer displacements, a symbolic relation allows the pointer position to be read and recorded as data. Because of this, the observer *O* is not needed, as the physical experimenter is quite able to observe the instrument directly by reading the pointer position for himself and record the data in symbolic form.

In the semiotic experiment no indexical relation exists between the experiment and the instrument. An indexical relation is established between the experiment and the observer *O*. Because of this *O* is an essential part of the semiotic experiment. The instrument *I* then serves as an aid in the translation of this indexical relation into an iconic relation directly between *P* and *O*. A symbolic scale on *I* now allows *O* to translate the iconic relation into a symbolic relation and hence to record it as data. *E* may observe *O*'s production of the symbol and may actually record the data himself, but because *O*'s translation processes are cognitive processes easily biased by knowledge of the true nature of *P*, *E* may not carry out the interpretation and translation processes himself. *O* and *E* are necessarily distinct.

In both kinds of experiments, *S*, *P*, *I*, and *E* are necessary. But in physical experiments, *O* is not necessarily distinct from *E*, and in semiotic experiments *O* is necessary and distinct. In semiotic experiments *O* is necessary to produce, in the absence of the indexical relation produced by physical effects, the iconic effect necessary for conversion to the symbolic data required for precise recording. It is the behavior of *O* that introduces the trinary relation into the experiment and it is this that makes it a semiotic experiment. By including *O* in the experiment, *E* may observe *S*, *P*, and *O* and thereby observe semiotic phenomena rather than just the physical phenomena he could only observe if he employed only *S* and *P*, and left *O* out of the setup.

We can now see that the observer in the semiotic experiment plays the same role as the pointer in the physical instrument; that is, he establishes the indexical relation, translates it into an iconic relation, and finally translates it into symbolic form for recording as data by himself or *E*. The semiotic instrument plays the role of the scale portion of the physical instrument; that is, it aids in the conversion from indexical to iconic and again in the conversion from iconic to symbolic. Therefore it would make much sense to reconceptualize the physical instrument as a combination of a transducer plus a scale and conceptually lump the transducer and the observer together, which we may symbolize as *T*. The semiotic instrument is a scale only, which we may symbolize as *C*. We then have the revised diagram of Fig. 2.

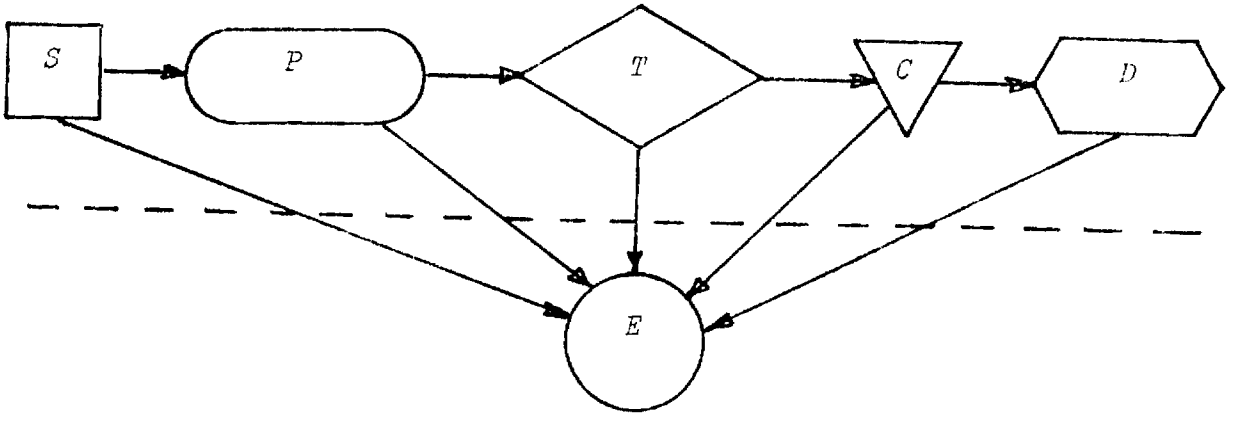


Fig. 2. Structure of a Scientific Experiment Reanalyzed.

We see that the structural analysis has not only been unified, it is also somewhat simplified by this reanalysis. Under this conceptualization the primary difference between a physics experiment and a semiotic experiment is that the physical transducer is purely physical, while the semiotic transducer necessarily includes a living interpreter, which is called the observer *O*. It is the presence of the interpreter in the transducer component that establishes the trinary relation and makes of this a semiotic experiment.

This is a sketch of the way I intend to incorporate the concept of measurement into the Semiotic Paradigm. The details have not as yet been worked out and the broad outlines are therefore still subject to revision.

8. LAWS OF INFORMATION SCIENCE/SEMIOTICS:

The semiotic nature of information science experiments carries implications about the structure of the laws of information science. These are determined by the Paradigm Inversion Principle. However, before examining such structure, it is necessary to delimit exactly what it is we are talking about when we say "the laws of information science". In modern philosophy laws are distinguished both from observation statements and from theoretical principles. Laws, or empirical generalizations, are general statements of invariant regularities among the observable concepts of nature, whereas observation statements are singular statements about observables. This simplistic way of making the distinction is all we can do in this limited space but will serve for the purposes of this discussion. Also by the term information science we should distinguish sharply from three somewhat related but obviously more popular fields of endeavor, information engineering, information technology, and management of information oriented activities. We are concerned here only with the basic science of information science, or what we previously called "semiotics".

In an experiment of basic information science, a fixed, finite, standard set of observers is part of the overall experimental design, and the set of stimulus messages is infinitely open. Two or more semiotic variables (called information attributes) are measured, let us call them *A* and *B* for example. Attributes *A* and *B* are each measured for each individual sign or message observed in the stimulus set. The data is collected and the measurements on *A* and *B* for the same individual sign or message are identified as the two components of a vector and a search is made for regularities (this is, of course, grossly simplified). If a regularity results, we can set up *A* and *B* as two coordinates of a graph and examine the nature of this regularity. Since both *A* and *B* are semiotic variables and the individuals are signs, and the interpreters are standardized, this is a law of information science/semiotics. The measurements are made by observing the interpretive behavior of the standard set of interpreters, perhaps by averaging some response over all the interpreters in the standard set. An example of a law of information science is the Law of Redundancy for Natural Language [2;3;16], shown in fig. 3. The differences between this law and Shannon's pseduo-relation are discussed in [18]. More detailed discussions of the nature of semiotic laws are contained in [11;15;16].

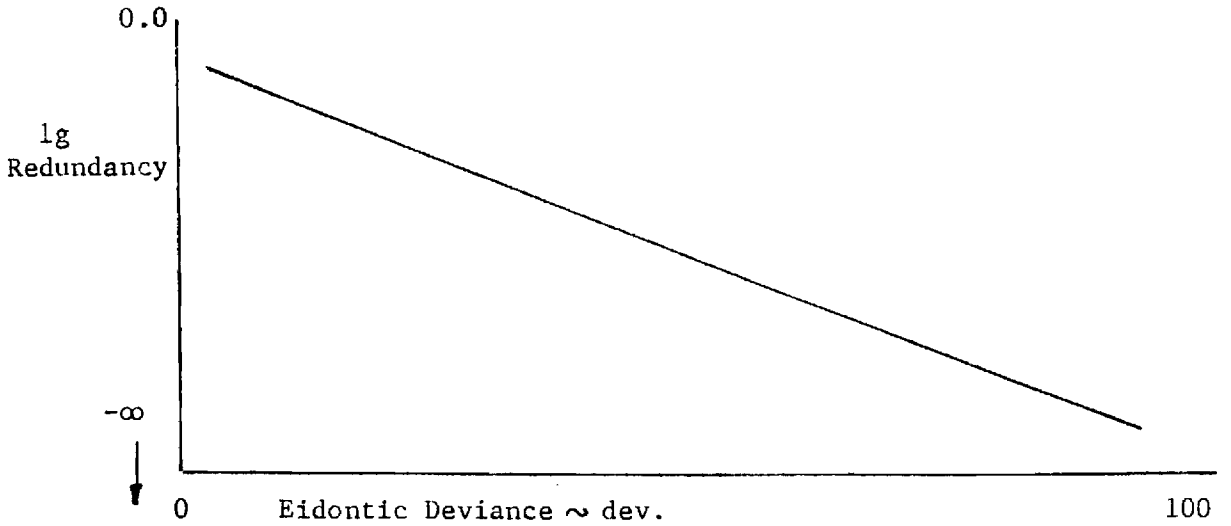


Fig. 3. Law of Redundancy for Natural Language.

9. AMBIGUITY OF "INFORMATION MEASURES":

The term 'information measure' is basic to information science; however it is often used ambiguously and the sense in which this term is used within the Semiotic Paradigm is not the same as the predominant current usage within the discipline. There are at least four senses of this term that are currently used often enough to warrant mention: 1) a measurement of some observable aspect of a sign; 2) an observable aspect of a sign which is capable of being measured/interesting enough to warrant measurement; 3) a mathematical model of an observable aspect of a sign; and 4) any mathematical function which could potentially be used for modeling such observable aspects. This fourth sense is perhaps the predominant usage of the term today. The Semiotic Paradigm employs only the second sense for this term.

By calling a mathematical function an "information measure" a great disservice is done to the discipline by confusing the distinction between an empirical aspect of nature and a mathematical model of that aspect. Because of this confusion, the discipline of information science itself has often been confused with a branch of mathematics (viz, NSF). The sine function is often used to model vibrations but the distinction between vibrations and the sine function is not thereby obliterated. We do not call the ratio of the side of a triangle opposite an angle to the hypotenuse of the triangle, the "vibration" of that angle. Such mathematical entities already have a useful name and there is nothing wrong in using it. It is 'function'. If there is some mathematical characterization of all and only those functions that may

legitimately model information measures, then this is an interesting study in its own right and the class of functions thereby determined deserves both a mathematical characterization, and a mathematical name. However, such a study can only come after a thoro empirical investigation of information measures. It is suggested that the current interest in information functions is due to a misnomer on the part of Shannon between "information theory" and the "theory of modal statistics". In the usage determined by the Semiotic Paradigm, an information measure is an observable aspect of semiotic nature that is capable of being measured and in which there is enough empirical interest to do so. In other words it is an empirically interesting aspect of an external sign component.

10. MATHEMATICAL METHODS FOR INFORMATION SCIENCE:

The golden age of mathematics was the seventeenth, eighteenth, and nineteenth centuries. Most of this development, even in pure mathematics, was stimulated by the needs of research in the physical sciences. For this reason, the mathematics of binary relations is highly developed, almost to the exclusion of any mathematics of trinary relations.

As the needs of research in the information/semiotic sciences begin to stimulate the development of new branches of mathematics the Semiotic Paradigm makes it easy to see that many of these branches will concern the mathematics of trinary relations. Thus the mathematical methods for information science will ultimately look radically different than any of today's mathematics.

Of the classical methods of mathematics, perhaps three are currently of most use in information science research, especially that conducted within the Semiotic Paradigm. These are: 1) inferential statistics; 2) calculus of finite differences; and 3) set theory and symbolic logic. Already two of the newer branches of mathematics have been stimulated as much, or perhaps more so, by research in the semiotic sciences. The theory of Markov chains was founded to satisfy needs arising out of the study of linguistic processes (A. A. Markov was a Russian semioticist) and has been successfully applied to the study of information sources, state descriptions, and many other problems of information science. Graph theory is currently of fundamental importance in the study of finite automata, formal languages and language recognizing machines, and the transformational grammars of natural languages. However, we can guess that the nature of linguistic transformations will ultimately produce a new "transformational calculus" with radically new mathematical characteristics.

I believe that ultimately the mathematics of mathematical semiotics will look radically different from the mathematics of mathematical physics; but it is too early to tell yet the form this mathematics will take. One thing is certain tho. This new mathematics will be developed by semioticists, and it will make much more use of the logic of trinary relations than current branches of mathematics do.

11. APPLICATIONS OF INFORMATION SCIENCE/SEMIOTICS:

The applicational components of a scientific paradigm, while not properly a part of basic science itself, sometimes help determine the goals of theory building and the direction of development for the basic science in that they can help determine what feedback from practical applications to be sensitive to and which phenomenas to explain. For instance, even tho thermodynamic laws are what they are because they describe objective and general regularities of nature, the way they were discovered and the order in which they were discovered was largely determined by the goal of explaining the practical phenomena of steam engineering.

In semiotics today, information engineering, information technology, and information management are playing much the same role as did steam engine technology in 19th century physics. The field of computer science is also beginning to require explanations in terms of basic semiotic laws and theories for its many practical relationships, especially in the field of language design.

We should also be aware of the possibility of "pure science", the development of basic science in isolation from any projected application. Charles Peirce, the father of information science, was especially sensitive to this possibility, calling it the method of the true scientist: one who seeks intellectual understanding for the pure joy of learning and with no thought of practical benefit in mind.

At this stage in the development of information science, one of the most important applications of our scientific investigations is further refinement of the scientific language, theory, experimental methodology, and mathematical methodology. The Semiotic Paradigm is an initial paradigm much like Ptolemy's orbital paradigm of astronomy, developed for the purpose of being made obsolete. Already it has undergone radical revision based on the outcome of empirical research. For instance, an investigation into the difference between lexical information and sentential information predicted the pragmatic nature of sentential mood (predominantly a syntactic phenomena) and led to a radical refinement in our understanding of the pragmatic structure of the sign. These results have currently been communicated only in private correspondence and therefore I cannot give you any reference to published literature for the details. Similarly, research in progress into the coding of information for the cognitive memories, suggests that refinements to our theories of semantic structure may soon be possible.

If we define engineering as the systematic and rational application of technology and scientific knowledge to human goals and purposes, then it becomes evident that about 98% of all effort that is called "information science" today is actually semiotic engineering. There is actually very little interest in science by most practitioners of "information science". Just to name some of these branches of semiotic engineering, we have: computer "science"; documentation engineering; library engineering/management; information engineering; etc. The advantage of the Semiotic Paradigm

to these applications is the development of a rational base for the development of these engineering disciplines. For instance, a semiotic understanding of the syntactic structure of the Type-Token Constellation, including the Rank-Frequency Law of Zipf and Estoup, could lead to the ability to derive Lotka's Law and Bradford's Law in rigorous form, including the controversial "droop" effect, from independently motivated relations concerning the structure of information. Lotka's Law and Bradford's Law are the cornerstone of library engineering.

In the one deliberate attempt to apply the Semiotic Paradigm to information engineering to date, I was able to use the independently known syntactic structure of the sign to motivate and explain many of the assumptions and methods used in "information theory", the calculus used in the engineering of many information systems [17;18;20].

The current state of information technology is completely "unscientific". It is one in which theory and experiment are completely divorced from technology. This can be diagrammed as fig. 4.

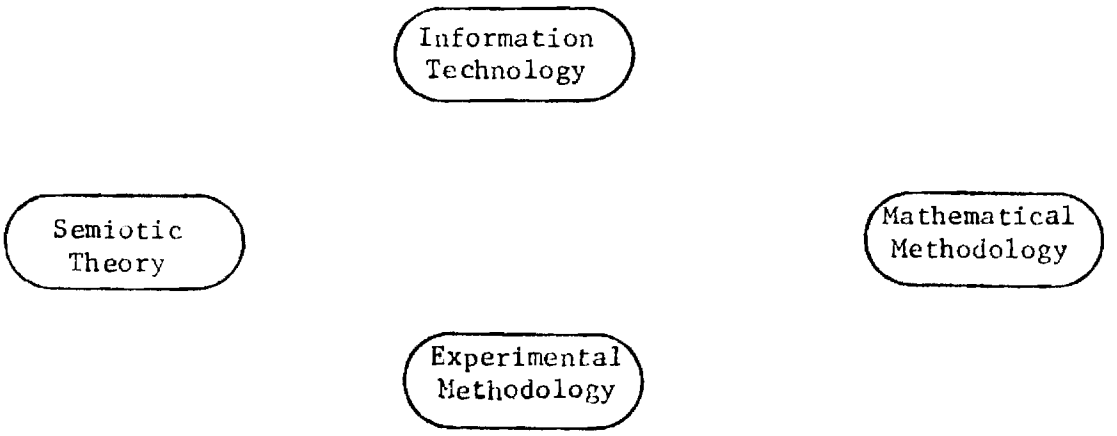


Fig. 4. The Current State of Information Science and Technology.

The goal of the Semiotic Paradigm is to provide a useful and rational base for linking these together much as fig. 5 with a resulting increase in the power and capabilities of information technology.

Finally, among the application areas of information science, we have information management, or the management of the information industry. This has so many aspects in common with other areas of management concern such as management of the automotive industry, management of the textile industry, etc., and is so different from the concern of science that there should actually be no effort within the area of information science to develop these applications. Accordingly the Semiotic Paradigm has no applicational paradigms in this area, altho I am sure that others more talented than I will be able to apply the Semiotic Paradigm even in these remote applicational areas. Actually much of what is called "information science" today is this information economics, information management applications stuff.

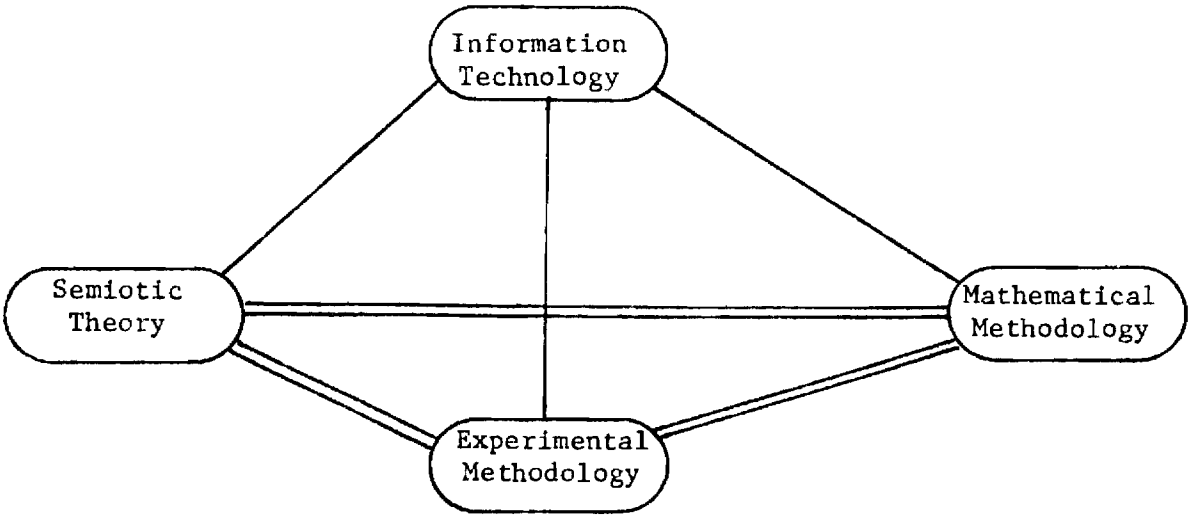


Fig. 5. The Hoped-for State of Information Science and Technology

12. OMISSIONS:

This discussion of the Semiotic Paradigm has been very superficial due to the confines of space and time. No details have been given except for the discussion of information measurement and even that discussion presented a sketch only. A fuller discussion would have presented the full glossary and grammar of the Language of Menetics; examples of its use; the principles, model, definitions, and theorems of the Universal Sign Structure Theory; examples of its application; a discussion of semiotic instruments, and the design of experimental tasks; a discussion of experimental methodology; data collection, reduction, and analysis; several examples of mathematical developments; and a full discussion of the applications. Unfortunately, this was not possible, but references were given wherever possible so that the interested investigator may follow up these details for himself.

13. FUTURE OF INFORMATION SCIENCE/SEMIOTICS RESEARCH:

There remain many details of the Semiotic Paradigm to be worked out, and now that the paradigm enables serious scientific research to take place, many details of the paradigm will continue to be revised. Indeed one hopes that it will enable progress to take place at such a rate that we here will all of us live to see one of Kuhn's scientific revolutions take place in which the Semiotic Paradigm is overthrown by some newer, more powerful paradigm. I suspect that when it happens, the new paradigm will also be semiotic in nature.

The Semiotic Paradigm contains some specific implications as to the direction of future research and where financial emphasis should be placed to encourage as much progress as fast and effectively as possible. Research should be directed towards the investigation of sign structure, and in order to conduct rigorous research in which all variables known to affect the outcome are either held constant or accounted for in the analysis, this means that research will involve one or two components of individual signs. Such research carries no direct ties to applications. In fact this is the epitome of scientific, laboratory, idealization to non-practical situations. We, as information scientists, NSF as research sponsors, and ASIS as our scientific peers and research evaluators, must learn this lesson and apply it. Research will become more experimentally and quantitatively oriented. This means a new sophistication for information scientists, for the NSF evaluators, and for our ASIS peers. No longer will a background in library management qualify one as an information scientist. It also means research will become more expensive. Lab equipment costs money. Not the billions of dollars that good physics experiments are currently costing, but a few hundred thousand to a million dollars per experiment, which by the older information science standards is very expensive. But by paying such amounts for legitimate experiments guided by theoretical questions and rigorous analysis of previous research, genuine progress is guaranteed. Not for each individual experiment of course, but for a much higher proportion than has been customary in the past.

14. ACKNOWLEDGEMENTS:

This work was supported in part by grant #IST-7827002 from the National Science Foundation, Division of Information Science and Technology. I would also like to thank my colleagues Pranas Zunde, for his intellectual stimulation and challenge, and Vladimir Slamecka, for his continuing support and encouragement. Many of the ideas contained herein were tried out and discussed with various members of Georgia Tech's Information Science Colloquium, including James Gough, Jr., Bill Underwood, Al Badre, and P.J. Siegmann, in addition to my two colleagues previously mentioned, whose stimulating criticism I sincerely appreciate. The ideas, opinions, and conclusions expressed in this paper are, of course, my own.

15. REFERENCES:

[1] Kunh, T.S. The Structure of Scientific Revolutions. U. of Chicago Press, 1962.

[2] Lo, R.H. "The Measurement of Comentropy Transfer Rates". Presented to the Third Annual Symposium on Empirical Semiotics held in conjunction with the 1980 Annual Meeting of the Semiotic Society of America; Lubbock, Texas; October, 1980. To appear in the proceedings.

[3] Lo, R.H. "Teescope Tuning Procedures for Comentropy Measurements". Presented at the First Annual Symposium on Foundations of Information Science held in conjunction with the 1981 Annual Meeting of the American Society for Information Science; Washington; October 28-29, 1981. To appear in the proceedings.

[4] Pearson, C. "Quantitative Investiations into the Type-Token Relation for Symbolic Rhemes". Proceedings of the Semiotic Society of America, 1(1976), p312-328.

[5] Pearson, C. "Towards an Empirical Foundation of Meaning". Ph.D. Thesis, Georgia Institute of Technology, 1977. University Microfilms; Ann Arbor, 48106, USA.

[6] Pearson, C. "The Cognitive Sciences: A Semiotic Paradigm". Presented at the National Conference on Mind, Brain, and Machine, Sponsored by the Society for the Interdisciplinary Study of the Mind; Gainesville, Fla.; April, 1978. To appear in the proceedings, in press.

[7] Pearson, C. "Empirical Study of Representation as a Unifying Methodology for Semiotics". Invited paper presented at the First Annual Symposium on Empirical Semiotics, held in conjunction with the Third Annual Conference of the Semiotic Society of America; Providence; October, 1978.

[8] Pearson, C. "A New Law of Information: An Empirical Regularity Between Word Shapes and their Interpretation". Presented to the Forty-first Annual Conference of the American Society for Information Science; New York City; November, 1978.

[9] Pearson, C. "The Problem of Communication in Empirical Semiotics". Invited paper presented at the Workshop on Empirical Semiotics held at the Second International Semiotics Congress; Vienna, Austria; July, 1979. To appear in the Proceedings.

[10] Pearson, C. "Decay of Information: A Second Order Correction to the Law of Word Interpretation". Presented to the Second International Semiotics Congress; Vienna, AUstria; July, 1979. To appear in the Proceedings.

[11] Pearson, C. "Empirical Implications for Semiotic Methodology". Invited paper presented at the Workshop on Cognitive Processes held at the Second International Semiotics Congress; Vienna, Austria; July, 1979. To appear in the Proceedings.

[12] Pearson, C. "The Theses of Empirical Semiotics". A "theses" presentation to the Second International Semiotics Congress; Vienna, Austria; July, 1979. Synopsis to appear in the proceedings.

[13] Pearson, C. "Semiotics and the Measurement of Shape". Seminar presentation at the Technische Universitaet Berlin; July, 1979. German version to appear in *Zeitschrift fuer Semiotik*, American version to appear in Progress in Information Science and Technology.

[14] Pearson, C. "The Basic Concept of the Sign". Proceedings of the ASIS Annual Meeting, 17(1980), p367-369.

[15] Pearson, C. "The Role of Scientific Paradigms in Empirical Semiotics". Proceedings of the Semiotic Society of America, 1980. (In press).

[16] Pearson, C. "The Epistomological Status of Shannon's Redundancy Curve and Markov's Law". Proceedings of the Nineteenth Annual Southeast Regional ACM Conference. Atlanta; March 27-29, 1981. (In press).

[17] Pearson, C; and Slamecka, V. Semiotic Foundations of Information Science. Final report for NSF grant #GN-40952. January, 1977. School of Information and Computer Science; Georgia Institute of Technology; Atlanta, Georgia.

[18] Pearson, C; and Slamecka, V. "A Theory of Sign Structure". Bull. Semiotic Soc. Amer., 1(1977), #2, p1-22.

[19] Pearson, C; and Zunde, P. "The Kolmogorov Potential as a Measure of Algorithmic Information". Toronto Semiotic Circle Monograph, 1980, #4. (In press).

[20] Slamecka, V; and Pearson, C. "The Portent of Signs and Symbols". Chapter 5 in The Many Faces of Information Science, (Weiss, E.C., Ed.). Westview Press, 1977, p105-128.

APPENDIX F

APPLICATION OF THE FINITE-DIFFERENCE CALCULUS TO THE OBSERVATION OF SYMBOL PROCESSES

By Charls Pearson

School of Information and Computer Science
Georgia Institute of Technology
Atlanta, 30332, USA

October 1981

Presented to the SIG/ES session on "The Role of Mathematics in Semiotic Observations" at the Fourth Annual Symposium on Empirical Semiotics held in conjunction with the Sixth Annual Meeting of the Semiotic Society of America, in Nashville; October 3, 1981. To appear in the Proceedings of the SSA.

APPLICATION OF THE FINITE-DIFFERENCE CALCULUS TO THE OBSERVATION OF SYMBOL PROCESSES

By Charls Pearson

School of Information and Computer Science
Georgia Institute of Technology
Atlanta, 30332, USA

ABSTRACT

Many of the methods of the finite-difference calculus have familiar analogs in the classical differential calculus altho the methods themselves are grounded on drastically different underlying theories and usually have drastically different results. For instance, the basic concepts of the differential calculus are 'real-valued measurements', 'continuum of the system of real numbers', and the 'limit operation'. The corresponding basic concepts of the finite-difference calculus are 'counting processes', 'discreteness of integer numbers', and 'summation operations'. Also corresponding to the integral of a continuous polynomial function $\int x^n dx$ we have the summation of a discrete factorial function $\sum n^{(n)}$, with

$$\int x^n dx = \frac{x^{n+1}}{n+1} + C$$

where C is a constant, usually determined by boundary conditions, and

$$\sum x^{(n)} = \frac{x^{(n+1)}}{n+1} + C(x)$$

where $C(x)$ is a periodic constant, again determined by the boundary conditions of the problem. The trick, therefore, in taking advantage of this particular analogy is to be able to transform back and forth between polynomial functions x^n and factorial functions $x^{(n)}$ and to be able to translate between periodic constants $C(x)$ and ordinary constants C. Similar tricks abound for utilizing the various other analogies between the two subjects.

The calculus of finite differences is therefore useful for modeling and describing discrete phenomena and discontinuous processes. Such for example are symbol relationships and the observation of symbol processes. This paper demonstrates the usefulness of the calculus of finite differences to model and describe observations of symbol processes. The clearer insight into symbol processes thus gained enables further refinements to the methods of observation which sharpen the observations themselves.

The Vocabulary Growth Rate curve for natural language has never been observed because of the low precision of classical counting procedures and the large measurement noise introduced by these procedures. These faults are due to the observation methods themselves and do not depend on whether the observations are carried out manually or by computer.

By using the finite-difference calculus to model the Vocabulary Growth Rate independently of its observability, and to show the relation between the Vocabulary Growth Rate and Type-Token Relation, insight is obtained that was used to redesign the classical counting methods. The resulting invention, the Echelon Counter, enabled measurement of the Vocabulary Growth Rate for the first time, as well as improved measurements of the Type-Token function.

While this paper demonstrates a point in mathematical semiotics, it uses an example from the experimental paradigm of Type-Token measurement. It is set within the linguistic-conceptual paradigm of Pearson's Language of Menetics and the theoretical paradigm of Pearson's Universal Sign Structure Theory.

A mathematical relation is obtained for the Type-Token Relation which satisfies all of the known boundary conditions exactly and describes the measured values approximately.

APPLICATION OF THE FINITE-DIFFERENCE CALCULUS TO THE OBSERVATION OF SYMBOL PROCESSES

By Charls Pearson
School of Information and Computer Science
Georgia Institute of Technology
Atlanta, 30332, USA

1. BACKGROUND:

Much of the present development of mathematics was originally motivated by its applications to problems in the physical sciences; in fact, the golden age of mathematics was the eighteenth century while the revolution in the physical sciences was still taking place. The physical scientists very early learned to appreciate the interrelation between mathematical methods and observational methods. This may be said to have originated with Gallileo who developed averaging methods in order to produce precise measurements with the crude instruments available to him.

Another golden age of mathematics is coming. It will be motivated by the application to problems in the semiotic sciences. But first, semiotics must make a science of itself with precise theories and rigorous experimental methods. The present paper precedes any of this development with a simple example of how a presently existing branch of mathematics may be used to model observations in experimental semiotics.

This paper demonstrates that the calculus of finite differences may be used to model and describe observations of symbol processes. The clearer insight into symbol processes thus gained enables further refinements to the methods of observation which sharpen the observations themselves. This is the first known attempt to apply the calculus of finite differences to experimental semiotics.

The Vocabulary Growth Rate curve for natural language has never been observed because of the low precision of classical counting procedures and the large measurment noise introduced by these procedures. These faults are due to the observation methods themselves and do not depend on whether the observations are carried out manually or by computer. However, Type-Token curves, Rank-Frequency Curves, and Number-Frequency curves obtained with classical counting instruments are displayed in figs. 1, 2, 3, 4, and 5. Even here the measurement noise and lack of precision are evident.

By using the finite-difference calculus to model the Vocabulary Growth Rate independently of its observability, and to show the relation between the Vocabulary Growth Rate and Type-Token Relation, insight is obtained that was used to redesign the classical counting methods. The resulting invention, the Echelon Counter, enabled measurement of the Vocabulary Growth Rate for the first time, as well as improved measurements of the Type-Token function.

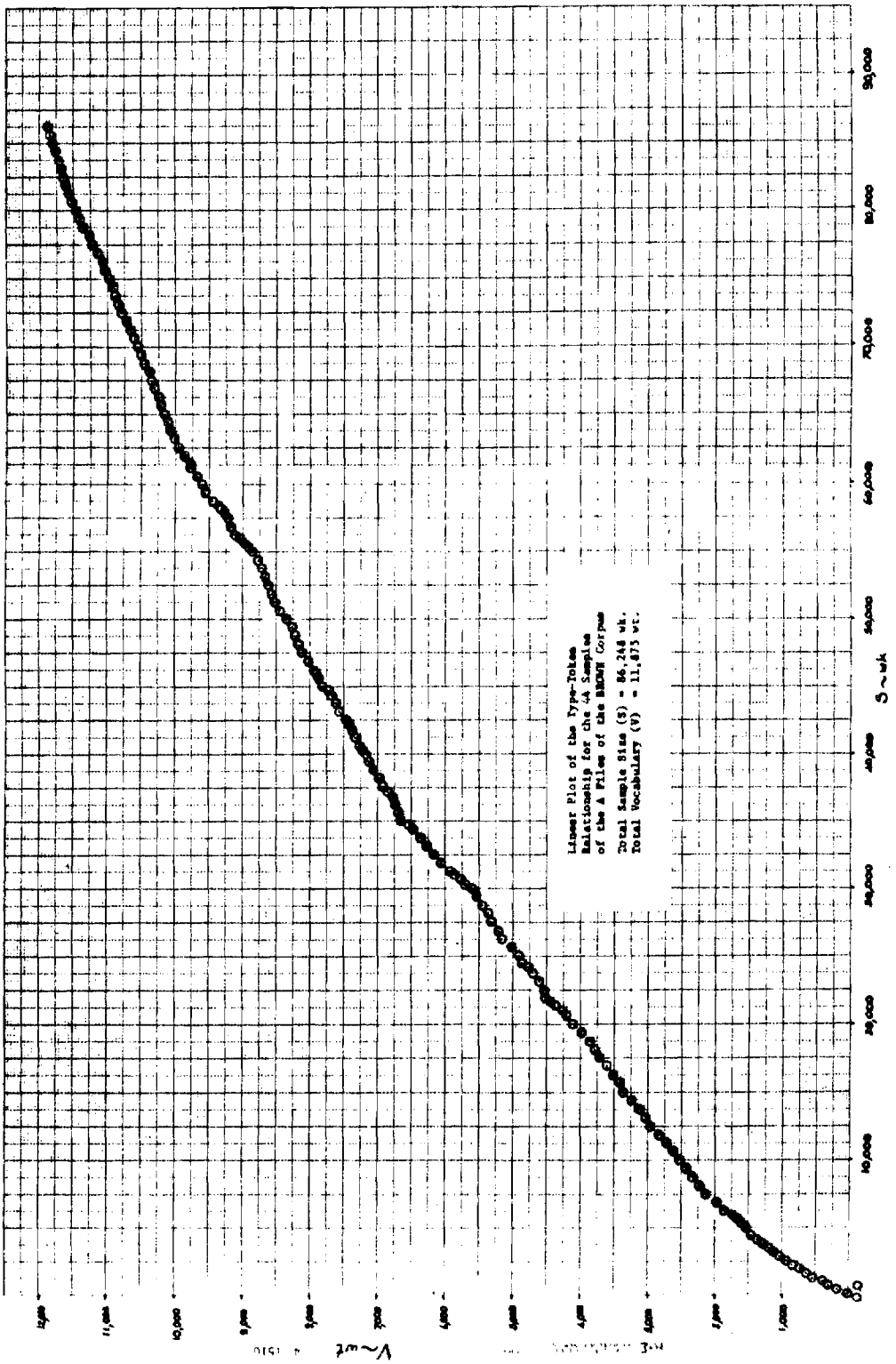


Figure 1. Type-Token Curve

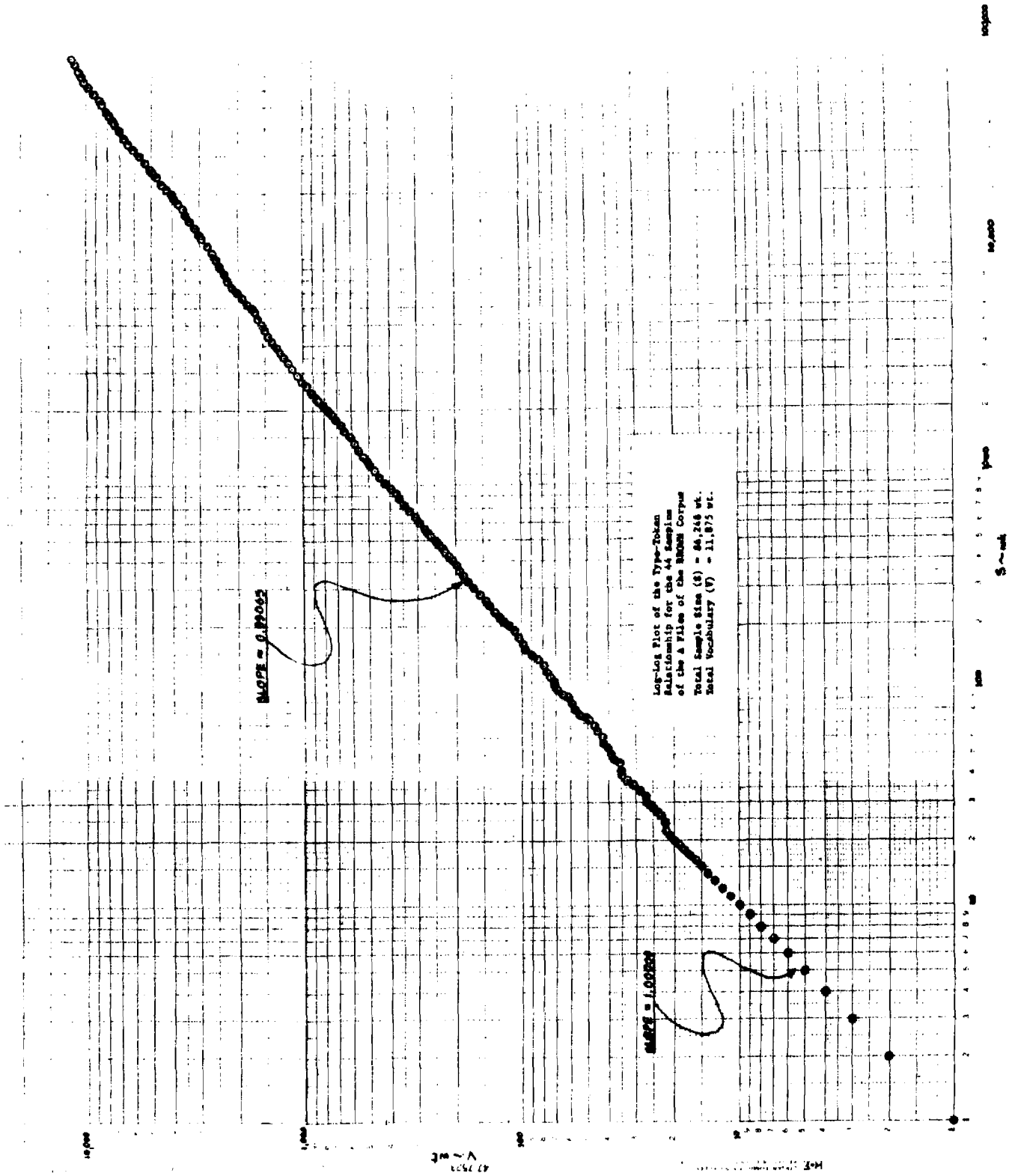


Figure 2. Type-Token Curve Log-Log Transform.

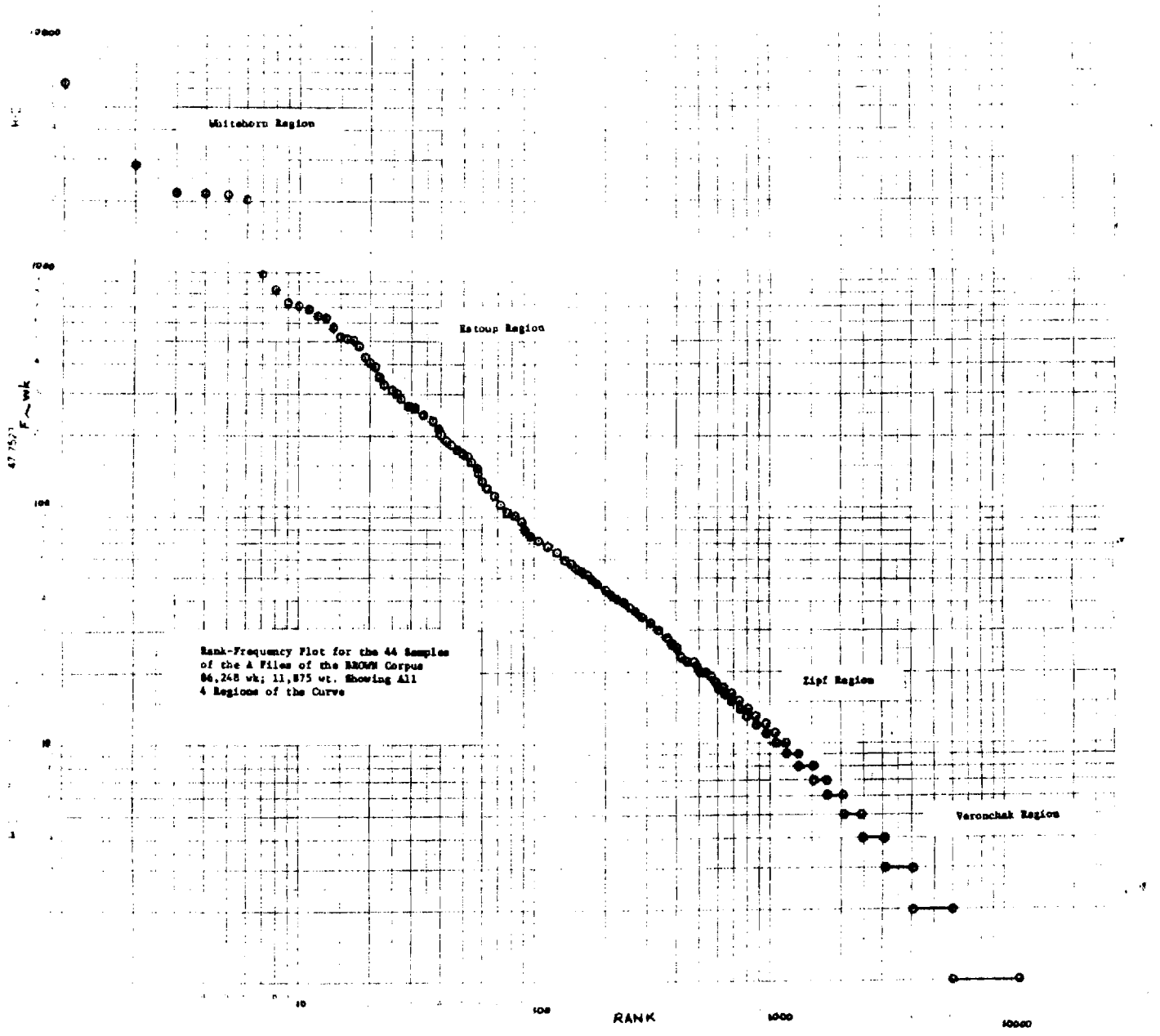


Figure 3. Rank-Frequency Curve.

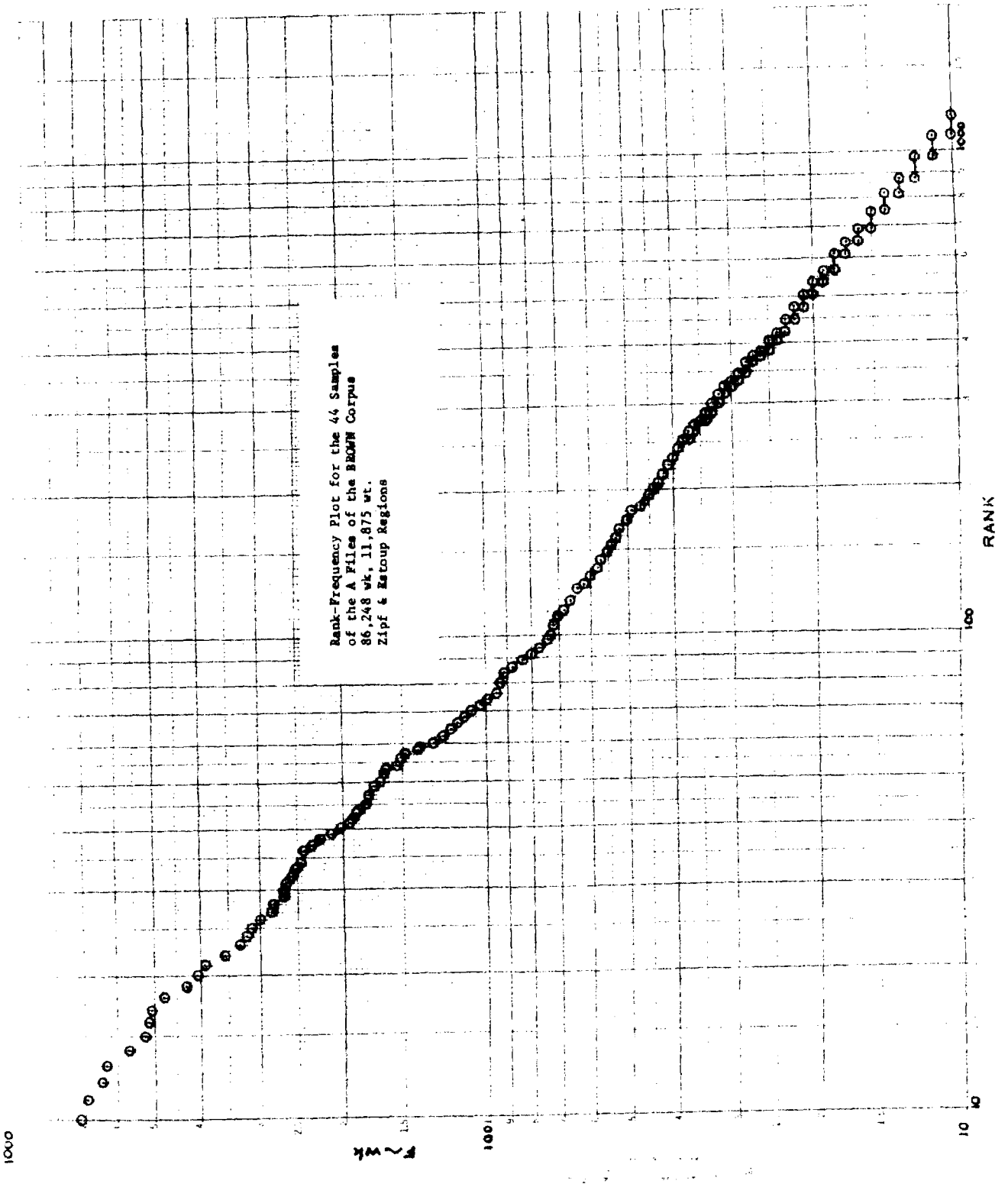
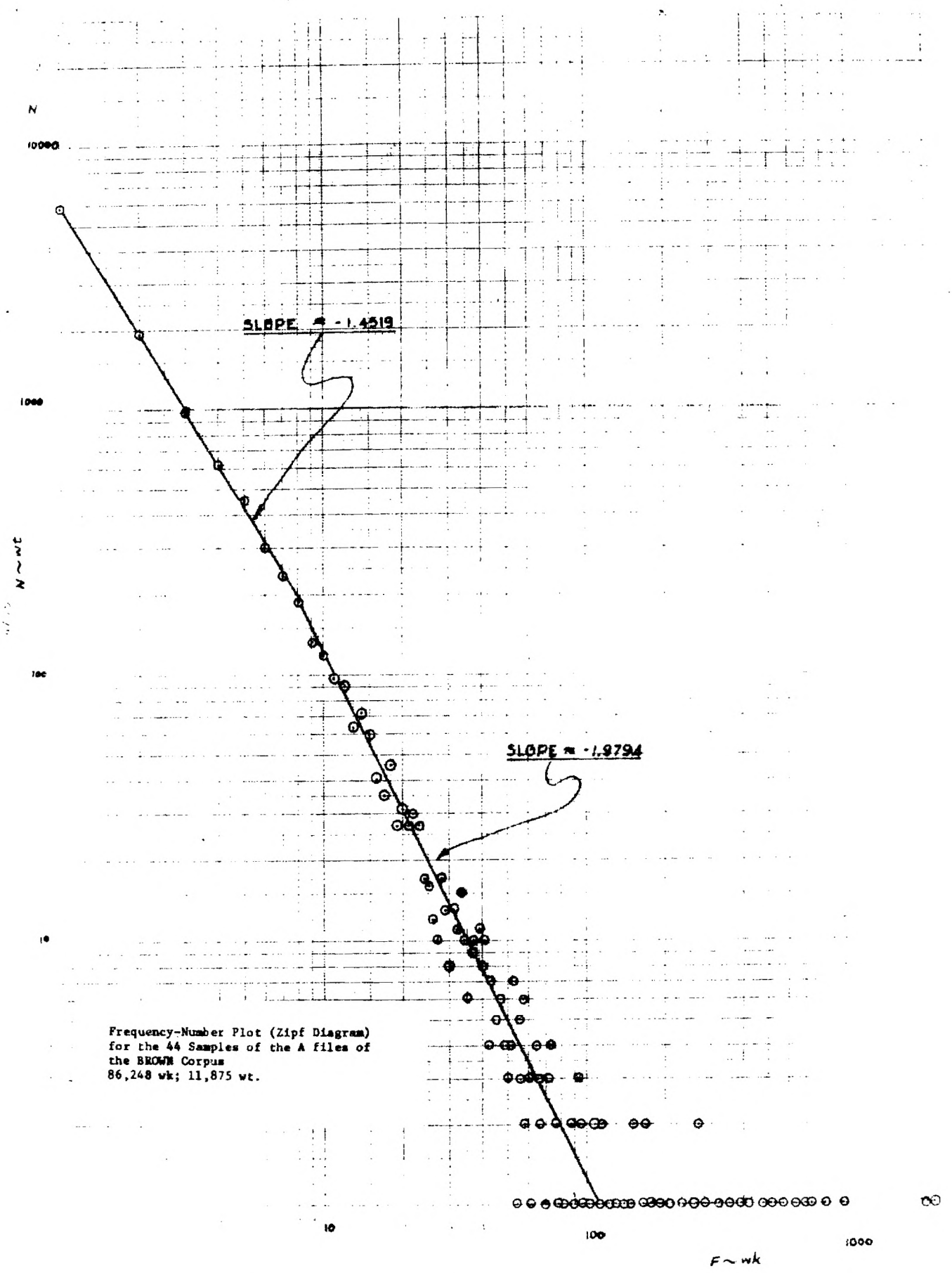


Figure 4. Zipf and Estoup Regions of Rank-Frequency Curve.



Frequency-Number Plot (Zipf Diagram)
for the 44 Samples of the A files of
the BROWN Corpus
86,248 wk; 11,875 wt.

Figure 5. Number Frequency Curve.

While the paper demonstrates a point in mathematical semiotics, it uses an example from the experimental paradigm of Type-Token measurement. It is set within the linguistic-conceptual paradigm of Pearson's Language of Menetics and the theoretical paradigm of Pearson's Universal Sign Structure Theory.

A mathematical relation is obtained for the Type-Token Relation which satisfies all the known boundary conditions exactly and describes the measured values approximately.

2. INTRODUCTION

The potential for applying the calculus of finite differences to experimental semiotics may be adumbrated by its recent applications to problems in some of the other semiotic sciences such as economics, psychology, and sociology. Introductory textbooks usually cover such topics as the difference calculus, the sum calculus, and finite-difference equations. The calculus of finite differences is the study of the general properties of the difference operator, Δ.

Given a function $f(x)$ we may define the DIFFERENCE OPERATOR, Δ, by

$$\Delta f(x) \equiv f(x + h) - f(x) \tag{1}$$

where h is some given number usually positive and called the DIFFERENCE INTERVAL. If in particular $f(x) = x$ we have

$$\Delta x = (x + h) - x = h \tag{2}$$

or

$$h = \Delta x \tag{3}$$

A geometric interpretation of Δ is given in fig. 6.

We note immediately the very strong analogy between the definition of the difference operator and that of the derivative operator from the differential calculus.

$$\frac{df(x)}{dx} = \lim_{h \to 0} \left\{ \frac{f(x + h) - f(x)}{h} \right\} \tag{4}$$

This analogy is very pervasive and powerful. Many of the methods of the finite-difference calculus have familiar analogs in the classical differential calculus altho the methods themselves are grounded on drastically different underlying theories and usually have drastically different results. For instance, the basic concepts of the differential calculus are 'real-valued measurements', 'continuum of the system of real numbers', and the 'limit operation'. The corresponding basic concepts of the finite difference calculus are 'counting processes', 'discreteness of integer numbers', and

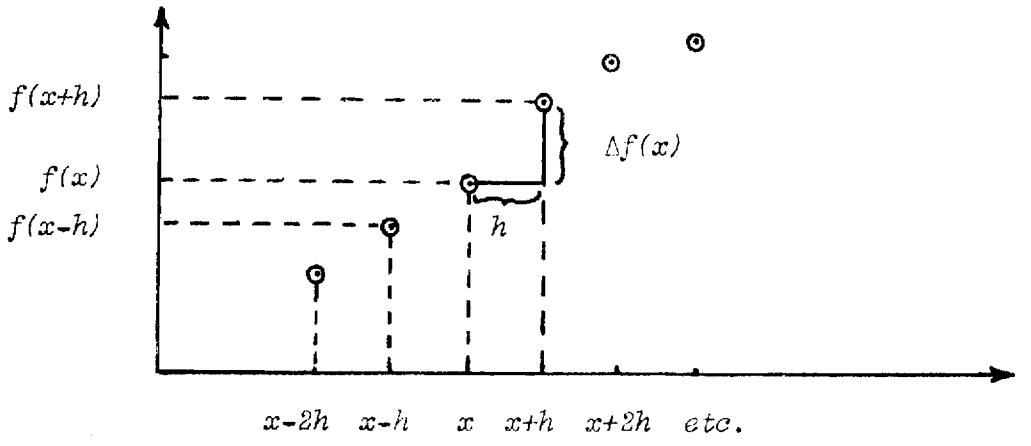


Figure 6. Geometric Interpretation of Difference Operation.

'summation operations'. Also corresponding to the integral of a continuous polynomial function $\int x^n dx$ we have the summation of a discrete factorial function $\sum x^{(n)}$, with

$$\int x^n dx = \frac{x^{n+1}}{n+1} + C \tag{5}$$

where C is a constant, usually determined by boundary conditions, and

$$\sum x^{(n)} = \frac{x^{(n+1)}}{n+1} + C(x) \tag{6}$$

where $C(x)$ is a periodic constant, again determined by the boundary conditions of the problem. A periodic constant has a constant value for integral values of h. The trick, therefore in taking advantage of this particular analogy is to be able to transform back and forth between polynomial functions x^n and factorial functions $x^{(n)}$ and to be able to translate between periodic constants $C(x)$ and ordinary constants C. Similar tricks abound for utilizing the various other analogies between the two subjects.

The calculus of finite differences is therefore useful for modeling and describing discrete phenomena and discontinuous processes. Such for example are symbol relationships and the observation of symbol processes.

3. THE FINITE-DIFFERENCE CALCULUS:

Having defined the difference operator Δ by eq. 1, we may now list the general rules of the finite-difference calculus.

$$\text{R1:}\quad \Delta[f(x) + g(x)] = \Delta f(x) + \Delta g(x) \tag{7}$$

$$\text{R2:}\quad \Delta[\alpha f(x)] = \alpha \Delta f(x) \tag{8}$$

where α is a constant.

$$\text{R3:}\quad \Delta[f(x)g(x)] = f(x)\Delta g(x) + g(x+h)\Delta f(x) \tag{9}$$

$$\text{R4:}\quad \Delta\left(\frac{f(x)}{g(x)}\right) = \frac{g(x)\Delta f(x) - f(x)\Delta g(x)}{g(x)g(x+h)} \tag{10}$$

Those of you familiar with the differential calculus will immediately notice the resemblance between these rules and the general rules of differentiation. In which case you might also recall there was a fifth rule of differentiation as follows:

$$D[f(x)]^m = m[f(x)]^{m-1} Df(x) \tag{11}$$

where m is a constant.

This rule does not carry thru exactly to the finite-difference calculus as do the other rules of differentiation. In order to perfect the analogy we define the FACTORIAL FUNCTION by

$$x^{(m)} \equiv x(x-h)(x-2h) \ \ldots \ (x-[m-1]h) \tag{12}$$

$$m = 1,2,3 \ \ldots$$

consisting of m factors. The name 'factorial' is motivated because in the special case $x = m$, $h = 1$, we have

$$m^{(m)} = m(m-1)(m-2) \ \ldots \ 2 \cdot 1 = m! \tag{13}$$

i.e., factorial m. In order to make this analogy as complete and systematic as possible and to simplify the resulting calculations, we define

$$x^{(0)} = 1. \tag{14}$$

Also for negative integers we define

$$x^{(-m)} \equiv \frac{1}{(x+h)(x+2h)\ \ldots\ (x+mh)} = \frac{1}{(x+mh)^{(m)}} \qquad (15$$

$$m = 1,2,3, \ \ldots$$

With our treatment of factorial functions complete we may now list the differences of some special functions:

1. $\Delta[c] = 0$ (16

2. $\Delta[x^{(m)}] = mx^{(m-1)}h$ (17

3. $\Delta[(ax+b)^{(m)}] = mah(ax+b)^{m-1)}$ (18

4. $\Delta[b^x] = b^x(b^h - 1)$ (19

5. $\Delta[e^{ax}] = e^{ax}(e^{ah} - 1)$ (20

6. $\Delta[\sin ax] = 2\sin(ah/2)\sin a(x + h/2)$ (21

7. $\Delta[\cos ax] = -2\sin(ah/2)\sin a(x + h/2)$ (22

8. $\Delta[\ln x] = \ln(1 + h/x)$ (23

Note the obvious similarity between these differences and the derivatives of the same functions.

We now prove eq. 17 both as an example of how to prove difference relations in general and as an example of the usefulness of our recently introduced factorial function. Applying definitions 12 and 1 we have

$$x^{(m)} = x(x-h)(x-2h)\ \ldots\ (x-[m-1]h) \qquad (24$$

$$(x+h)^{(m)} = (x+h)(x)(x-h)\ \ldots\ (x-[m-2]h) \qquad (25$$

$$\begin{aligned}\Delta x^{(m)} &= (x+h)^{(m)} - x^{(m)} \\ &= (x+h)(x)(x-h)\ \ldots\ (x-[m-2]h) \\ &\quad - x(x-h)(x-2h)\ \ldots\ (x-[m-1]h) \\ &= [(x+h) - (x-[m-1]h)](x)(x-h)\ \ldots\ (x-[m-2]h) \\ &= mhx^{(m-1)} \qquad (26\end{aligned}$$

In order to apply these tools to solving problems it is necessary to be able to pass back and forth from the differential notation to the difference notation. Since the one analogy we have developed in detail relates derivatives of power functions to differences of factorial functions, we develop a notation for passing back and forth between power functions and factorial functions. This makes use of factorial polynomials. From eq. 12 we find on putting $m = 1,2,3, \ \ldots$

$$
\left.
\begin{aligned}
x^{(1)} &= x \\
x^{(2)} &= x^2 - xh \\
x^{(3)} &= x^3 - 3x^2h + 2xh^2 \\
x^{(4)} &= x^4 - 6x^3h + 11x^2h^2 - 6xh^3 \\
x^{(5)} &= x^5 - 10x^4h + 35x^3h^2 - 50x^2h^3 + 24xh^4
\end{aligned}
\right\} \qquad (27
$$

etc.

If p is any positive integer, we define a FACTORIAL POLYNOMIAL OF DEGREE (p) as

$$a_0 x^{(p)} + a_1 x^{(p-1)} + \ldots + a_p$$

where $a_0 \neq 0$, a_1, ..., a_p are constants. From eqs. 27 we see that a factorial polynomial of degree (p) can be expressed uniquely as an ordinary power polynomial of degree p. In fact if one is a master of many mathematical models he may recognize the numerical coefficients appearing in equations 27 as the Stirling Numbers of the First Kind, s_k^n where we define a STIRLING NUMBER OF THE FIRST KIND recursively by

$$
\left.
\begin{aligned}
& s_k^{n+1} = s_{k-1}^n - n s_k^n \\
\text{with} \quad & s_n^n = 1, \; s^n{}_k = 0 \quad \text{for } k \leq 0, \; k \geq n+1 \text{ where } n > 0.
\end{aligned}
\right\} \qquad (28
$$

This allows us to simplify the transformations of eq. 27 by using the Stirling Number notation as follows

$$x^{(n)} = \sum_{k=1}^{n} s_k^n \, x^k \, h^{n-k} \qquad (29$$

Conversely any power polynomial of degree p can be expressed uniquely as a factorial polynomial of degree (p). We write the first few:

$$
\left.
\begin{aligned}
x &= x^{(1)} \\
x^2 &= x^{(2)} + x^{(1)}h \\
x^3 &= x^{(3)} + 3x^{(2)}h + x^{(1)}h^2 \\
x^4 &= x^{(4)} + 7x^{(3)}h + 6x^{(2)}h^2 + x^{(1)}h^3 \\
x^5 &= x^{(5)} + 15x^{(4)}h + 25x^{(3)}h^2 + 10x^{(2)}h^3 + x^{(1)}h^4
\end{aligned}
\right\} \qquad (30
$$

etc.

From this example we see that power polynomials can be expressed uniquely in terms of factorial polynomials by

$$x^n = \sum_{k=1}^{n} S_k^n \, x^{(k)_h} h^{n-k} \tag{31}$$

where the S_k^n are STIRLING NUMBERS OF THE SECOND KIND defined recursively as

$$\left.\begin{aligned} S_k^{n+1} &= S_{k-1}^n + k S_k^n \\ \text{with} \quad S_n^n &= 1, \; S_k^n = 0 \quad \text{for } k \neq 0, \; k \neq n+1 \text{ where } n > 0. \end{aligned}\right\} \tag{32}$$

This completes our short introduction to the finite-difference calculus except for one special relation from the sum calculus that will be used in section 6. For completeness, we now state this special result without proof.

$$\sum x^{(-1)} = \sum \frac{1}{x+h} = \frac{\Gamma'(\frac{x}{h} + 1)}{h \, \Gamma(\frac{x}{h} + 1)} \tag{33}$$

The function on the right of eq. 33 is called the 'DIGAMMA FUNCTION' and is denoted by $\Psi(x)$.

4. SYMBOL PRODUCTION PROCESSES:

One of the obvious areas in which to attempt to use the finite-difference calculus as a mathematical model is in the study of symbol production processes, for the reason that symbol production is by nature a discrete process. One can produce a one or two-word text, but not a one-and-a-half word text. It is meaningless to conceive of a one-and-a-half word text. It must be either one word long or two words long because symbols are produced, and exist, discretely. Currently under investigation in the SemLab is a constellation of different relationships and symbol production processes, all intertwined in what may be called the "Type-Token System for Words in Natural Language" or "Type-Token Constellation" for short. These include Zipf's Number-Frequency Law also known as the Zipf Integer Effect, the Rank-Frequency Law of Words and Holophrases which is also known as the Law of Zipf and Estoup, the Type-Token Curve as a function of sample size, the Type/Token Ratio also as a function of sample size, the Vocabulary Growth Rate curve, and many others. In addition, several of the useful regularities of information engineering and library management such as Lotka's Law and Bradford's Law are closely related to the Type-Token Constellation. We will now apply the finite-difference calculus to the study of the Type-Token relationship. You saw examples of some of these earlier.

Because the ordinary Type-Token Relation is now statistically independent, it cannot be described using the methods of statistical estimation. In addition present methods of measurement yield too much measurement noise and

too little precision for statistical estimation methods to be useful even if they were valid [1]. For this reason the SemLab searched for a type-token relationship in which the data satisfies the statistical independence requirements rigorously. The search led us to a little-studied relationship called the Vocabulary Growth Rate. This is defined as the rate at which new vocabulary items, measured in word-types (wt), enter the sample with respect to the increase in size of the sample, measured in word-tokens (wk), [2]. The relative frequency of new words in the *n*th position of a sample is *logically* independent of the relative frequency of new words in any of the other word positions. It turns out that the statistical dependence of the data in each of the other type-token relationships arises because each of these other relations depends on the Vocabulary Growth Rate in a way that destroys independence of the measurements. In addition, and in compensation perhaps, this dependence is such that each of these other relations can be derived from the Vocabulary Growth Rate. For instance, the Type-Token Relation can be expressed as

$$T(K) = \sum_{S=1}^{K} VGR(S) \qquad (34$$

where $T(K)$ is the number of types in a sample expressed as a function of the sample size K, and $VGR(S)$ is the Vocabulary Growth Rate expressed as a function of the sample size S. This allows us to concentrate our experimental investigation on observations of the Vocabulary Growth Rate and later obtain each of the other relation by means of the finite-difference calculus.

However, there is a good reason why the Vocabulary Growth Rate relation has been little studied: it has never been observed. Therefore semioticists do not have even a vague inuition as to an approximate mathematical form for describing this relation. Therefore, altho this relation has been mentioned in the literature, nothing substantive has been learned about it.

The reason the Vocabulary Growth Rate curve has never been observed before is because of the lack of precision of all previous instruments for measuring type-token phenomena. The value of the Vocabulary Growth Rate for any sample size S is a real number between 0 and 1. All previous instruments for measuring type-token phenomena, including instruments employing digital computer techniques, use methods based on raw counting procedures which are precise to the nearest whole integer and whose values therefore increase by either 0 or 1 at each step. Therefore the Vocabulary Growth Rate was completely hidden between the cracks of the instrumentation.

5. MATHEMATICAL DEVELOPMENTS:

The mathematical relations which model the empirical relations of the Type-Token Constellation are related by statistical sampling theory, statistical averaging theory, and the calculus of finite-differences. Given an assumed form for the underlying theoretical distribution of the Vocabulary Growth Rate curve, the model of a single measurement of the curve can be obtained by sampling theory. From this a Vocabulary Growth Rate Number-Frequency curve can be obtained by finite differentiation and the general form of the observed Vocabulary Growth Rate curve can be obtained by averaging theory.

From the assumed theoretical distribution of the Vocabulary Growth Rate, the underlying theoretical distribution of the Type-Token curve can be obtained by Stieltjes integration which reduces in this case to a simple summation. From the theoretical distribution of the Type-Token curve, the mathematical model of a single measurement of the curve can be obtained by sampling theory. And again, from this a Type-Token Number-Frequency curve can be obtained by finite differentiation and the general form of the observed Type-Token curve can be obtained by averaging theory.

From the theoretical distribution of the Type-Token curve, the underlying theoretical distribution of the Rank-Frequency curve can be obtained by a Stieltjes transform which reduces in this case to a summation transform, or what may be called a finite-difference transform. Again, from the theoretical distribution of the Rank-Frequency curve, a single measurement of the curve can be obtained by sampling theory. And again, from this a Rank-Frequency Number-Frequency (or Zipf's Number-Frequency) curve can be obtained by finite differentiation, and the general form of the observed Rank-Frequency curve can be obtained by averaging theory. However, because of the relationship between the transformed variables, the general form of the Rank-Frequency curve is exactly the same as a single measurement of this curve.

Certain key relations can be set out in advance. For intance, the theoretical form of the Type-Token curve $T(K)$ can be determined from the assumed form of the Vocabulary Growth Rate curve $VGR(S)$ by indefinite summation as follows:

$$T(K) = \sum_{S=1}^{K} VGR(S) \tag{35}$$

Also noting that max $R = T$ and that the sum over all types of the frequencies for each type is just the total number of tokens, we get:

$$K = \sum_{R=1}^{T(K)} F(R) \tag{36}$$

which is to be solved for $F(R)$. This last may be seen more easily with the aid of the following diagrams, where $\sum_{R=1}^{T(K)} F(R)$ is the area under the curve of fig. 7, and K is the horizontal coordinate of fig. 8.

In the case where the Rank-Frequency curve is hyperbolic, Zipf [3] obtained the Rank-Frequency-Number-Frequency relation:

$$N = \frac{C}{(F^2 - 1/4)} \tag{37}$$

by carrying out the finite differentiation as follows: let the rank-frequency curve be given by the hyperbolic relation

$$R = \frac{C}{F} \tag{38}$$

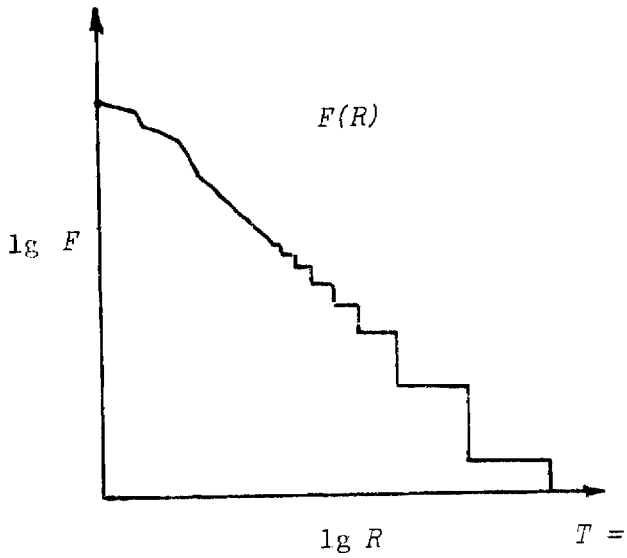


Fig. 7

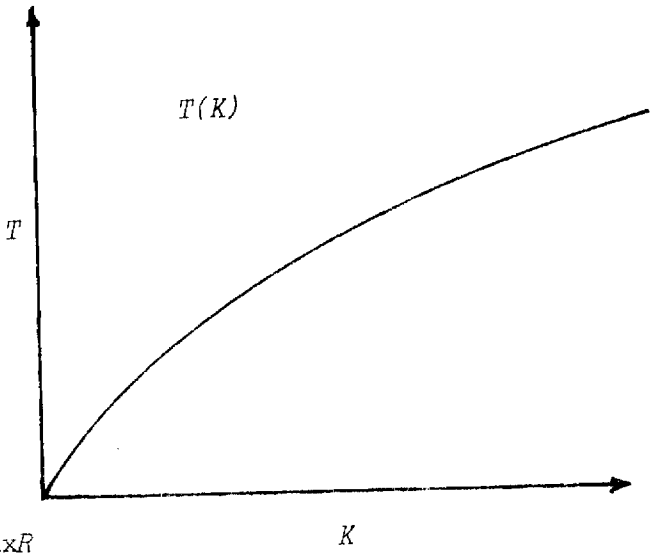


Fig. 8

Then the number of types of the same frequency is given by the requirement for integer occurrences,

$$N = R' - R'' = \frac{C}{F - 1/2} - \frac{C}{F + 1/2} = \frac{C}{(F^2 - 1/4)} \qquad (39$$

where F can assume only integer values, as can be seen by the diagram in fig. 9.

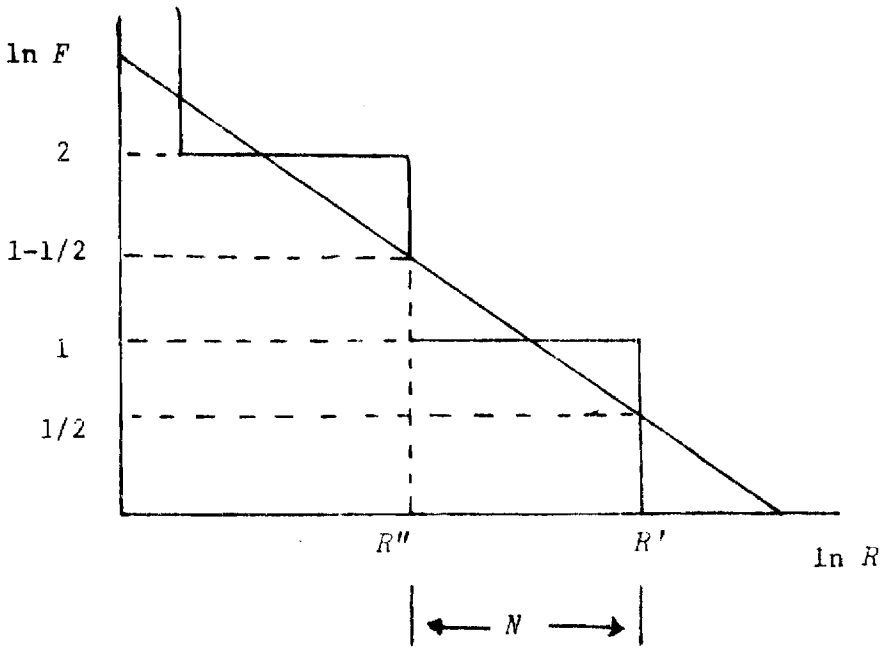


Fig. 9. Zipf Integer Diagram.

Derivation of number-frequency curves for the Type-Token relation and the Vocabulary Growth Rate relation is similar.

Arbitrary constants which appear in these equations are to be evaluated by means of the known semiotic boundary conditions as set forth in [1].

$$T(0) = 0 \tag{40}$$

$$T(1) = 1 \tag{41}$$

$$n \geq m \rightarrow T(n) \geq T(m) \tag{42}$$

$$n \text{ finite} \rightarrow T(n) \text{ finite} \tag{43}$$

These last two conditions can be combined into one more powerful condition by simply noting that both conditions hold at every point of the Type-Token relation. This is shown in fig. 10 where we suppose that we have counted the first m wk. of a sample giving us $T(m)$ as the wt. encountered up to this point. If we count an *additional* n wk. as part of the same sample, we cannot add more than n wt. to the vocabulary even if every word-token is an occurrence of a *new* word-type. This gives us an upper limit for the projection of $T(m + n)$ past $T(m)$ of

$$T(m + n) \leq T(m) + n \tag{44}$$

which is represented in fig. 10 by the dotted line at $\theta = 45°$ projecting from P.

On the other hand we also see that even if all of the new word-tokens are occurrences of word-types already encountered in the sample up to the mth token, we cannot decrease the number of word-types already encountered. This gives a lower limit for the projection of $T(m + n)$ past $T(m)$ of

$$T(m) \leq T(m + n) \tag{45}$$

which is represented in fig. 10 by the dotted horizontal line projecting from P. Since the actual observation P', must lie between these two extremes and can take on either limit, both conditions must hold simultaneously, giving:

$$\forall(m,n \in N) : T(m) \leq T(m + n) \leq T(m) + n \tag{46}$$

where N is the set of natural numbers; i.e., the non-negative integers. This is the restricted monotone condition and from it we can recapture both eq. 42 and eq. 43 as well as another important condition

$$0 \leq T(m) \leq n \tag{47}$$

by substituting $m = 0$ and applying condition 40. This is the first time the restricted monotone condition has been stated for the Type-Token relation.

In summary, the Type-Token Constellation consists of three theoretical distributions; three general observable relations; and three number-frequency, or individual relations: each trio consisting of one each for Vocabulary Growth Rate, Type-Token, and Rank-Frequency. These nine relations form a mathematically consistent system.

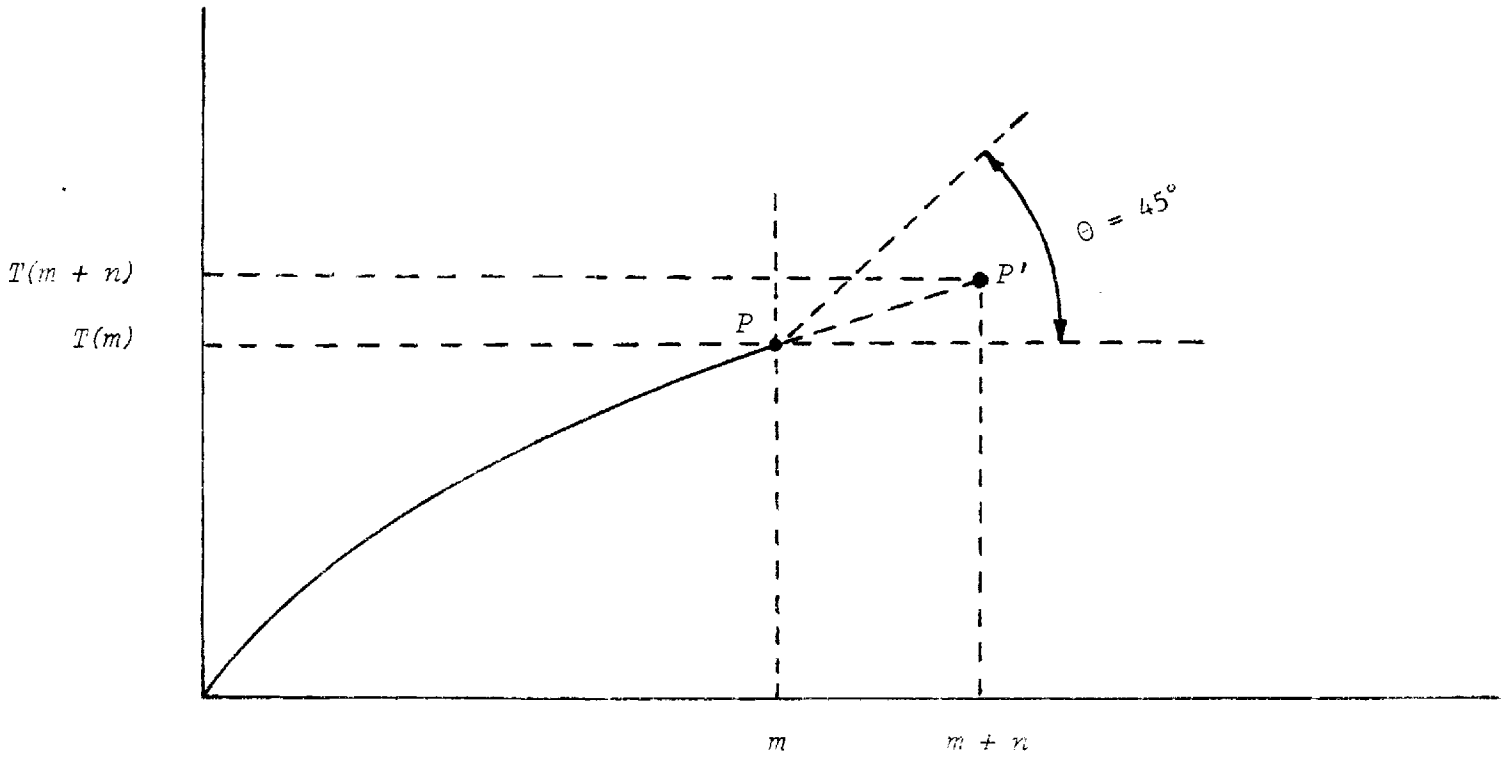


Fig. 10. The Restricted Monotone Boundary Condition for the Type-Token Relation.

6. THE ECHELON COUNTER AND THE FINITE-DIFFERENCE MODEL:

With this much as a preliminary model and using the better understanding of our observational limitations that it yielded, we invented an instrument for counting types and tokens, called an echelon counter, that yielded much higher precision and noise suppression. The Echelon Counter was described in detail in a patent disclosure and its design and performance was reported publicly in [2]. A comparative example of results of measuring type-token data with classical counting instruments and with the Echelon counter is shown in Figs. 11 and 12. Using the Echelon Counter, the Vocabulary Growth Rate data was clearly observable and in preliminary studies, $\frac{\Delta T}{\Delta S}$ appears to be approximately equal to $(S + 1)^{-1}$; using this and the fact that the spacing between text sample sizes is

$$\Delta S = h = 1 \text{ wk}, \tag{48}$$

we get

$$\Delta T = (S + 1)^{-1} \tag{49}$$

From this we get by a Stieltjes integration

$$T(S) = \sum_{i=1}^{S} \Delta T = \sum_{i=1}^{S} (i + 1)^{-1} = \sum_{i=1}^{S} i^{(-1)}$$

$$= \frac{\Gamma'(S + 1)}{h\Gamma(S + 1)} + C(S)$$

$$= \Psi(S) + C(S) \tag{50}$$

where $\Psi(S)$ is our old friend the digamma function from eq. 33.

Let us now attempt to evaluate $C(S)$ by applying the boundary conditions eqs. 40, 41, and 46. For $S = 0$ we have

$$\frac{\Gamma'(1)}{\Gamma(1)} + C(S) = \frac{-\gamma}{1} + C(S) = 0 \tag{51}$$

yielding

$$C(S) = \gamma(S) \tag{52}$$

where γ is Euler's Constant

$$\gamma \approx 0.5772 \ldots \tag{53}$$

and $\gamma(S)$ is appropriately called a '<u>PERIODIC EULER'S CONSTANT</u>'.

We have now used up all of the undetermined factors in the solution but still have two boundary conditions left to satisfy. How shall we take care of these? Let us see what needs to be done yet to satisfy them. Let us calculate T(1), we have

$$\frac{\Gamma''(2)}{\Gamma(2)} + \gamma(2) = \frac{1 - \gamma}{1} + \gamma = 1. \tag{54}$$

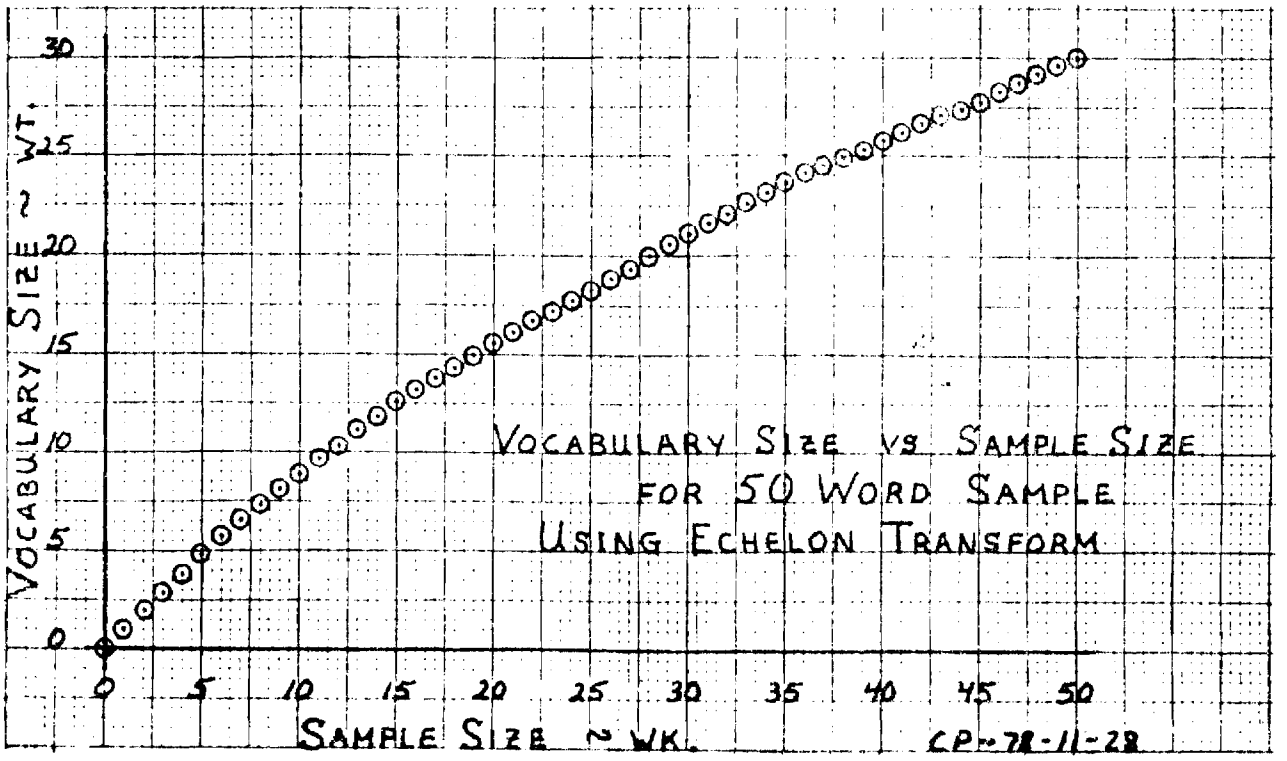


Fig. 11. A 50-word sample measured with the echelon counter.

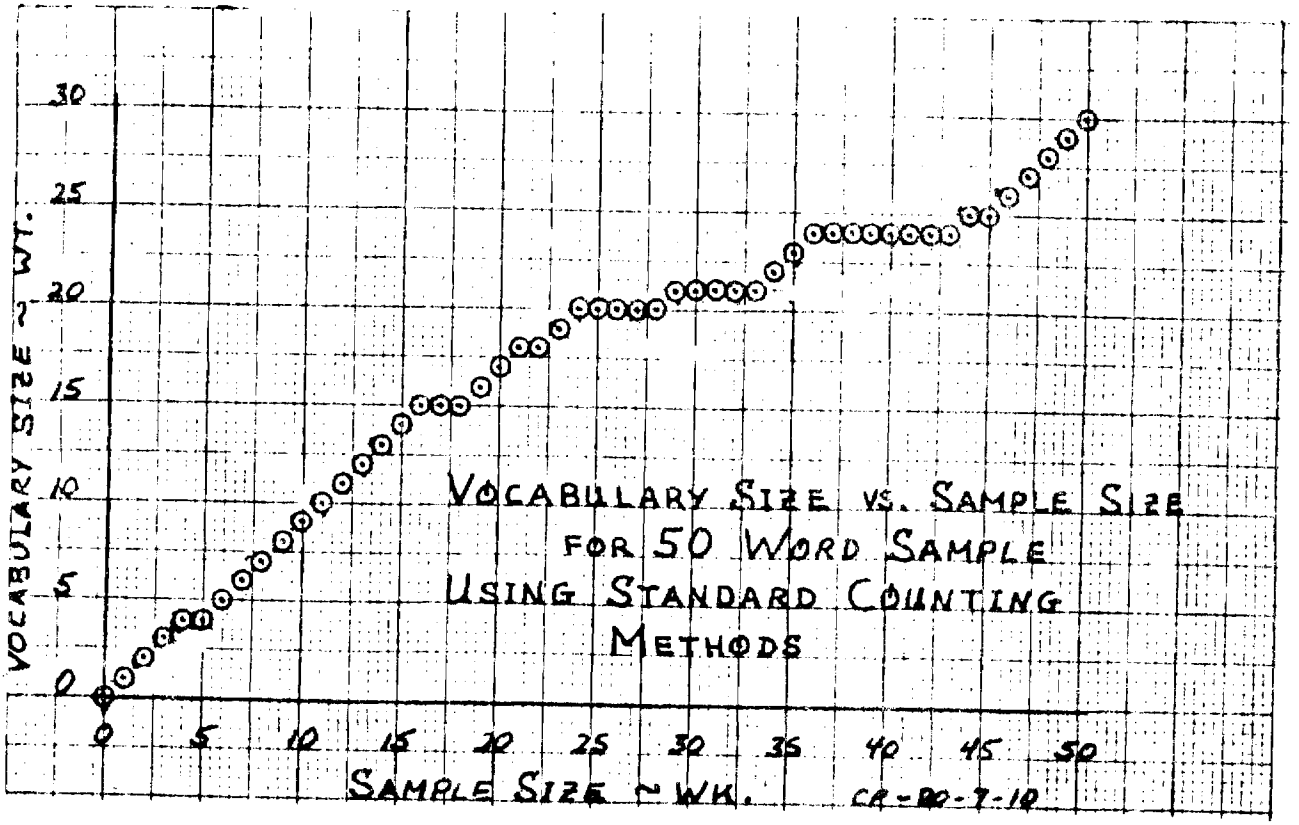


Fig. 12. The same 50-word sample using classical measuring techniques.

In other words, the second boundary condition is already satisfied by our solution, merely by its form. Likewise the digamma function is monotonically increasing for all positive values of S in a restricted way that satisfies eq. 46 thus guaranteeing the satisfaction of the last boundary condition *merely by the form of the solution*. It should be emphasized here that this is the first proposed form of the Type-Token Relation that satisfies all three boundary conditions by any means, let alone by its form alone. This is a significant achievement for the finite-difference calculus. We thus have our final relation

$$T(S) = \Psi(S) + \gamma(S) \tag{55}$$

7. CAVEATS:

It must be emphasized in the strongest terms that eq. 55 is not our final proposal for the Type-Token Relation. It is our final step in this motivation of the semiotic usefulness of the finite-difference calculus and with its achievement we certainly have accomplished that; however, there are still many problems associated with eq. 55 that remain to be cleared up by detailed investigations. For one thing, this development takes into account neither the peculiar nature of individual languages nor of individual authors, both of which are known to affect the Type-Token Relation. Nor does it take into consideration the grammatical constraints of natural language as opposed to a random string of words. In addition for all values of S greater than 1, the function produces too small a value. For instance,

$$T(2) = \frac{\Gamma'(3)}{\Gamma(3)} + \gamma(3) \approx 1.512 \tag{56}$$

a value which is approximately 0.48 too small; and 0.48 is easily detected by our present instrumentation. It is suggested that the proper Type-Token Relationship may be given by a sum of terms

$$T(S) = \Psi(S) + \gamma(S) + L(S) + G(S) + A(S) \tag{57}$$

where $L(S)$ is a term determined by the particular language, $G(S)$ is a term determined by the grammatical constraints of the language, and $A(S)$ is a term determined by the particular author. If this were the case, the present analysis has succeeded in isolating the first two of these terms.

9. SUMMARY:

The finite-difference calculus allows us to obtain the Type-Token Relation in terms of the Vocabulary Growth Rate as

$$T(K) = \sum_{S=1}^{K} VGR(S) \tag{58}$$

with boundary conditions given by

$$T(0) = 0 \tag{59}$$

$$T(1) = 1 \tag{60}$$

$$T(m) \leq T(m + n) \leq T(m) + n \qquad (61$$

If $VGR(S)$ is close to $(S + 1)^{-1}$ as appears to be the case in our initial measurements then the Type-Token Relation can be expressed as

$$T(S) = \Psi(S) + \gamma(S) \qquad (62$$

However, it is more likely that there are several additional terms to account for the individual language, author, and grammatical constraints and the expression may be more like

$$T(S) = \Psi(S) + \gamma_1(S) + \gamma_2(S) + \gamma_3(S) + \gamma_4(S) \qquad (63$$

In any case it is obvious that the finite-difference calculus is an exceedingly powerful tool for the study of symbol production processes.

9. ACKNOWLEDGEMENTS:

This work was supported in part by grant #IST-7827002 from the National Science Foundation, Division of Information Science and Technology. I would also like to thank my colleagues Pranas Zunde, for his intellectual stimulation and challenge, and Vladimir Slamecka, for his continuing support and encouragement.

10. REFERENCES

[1] Pearson, C. "Quantitative Investigations into the Type-Token Relation for Symbolic Rhemes". Proceedings of the Semiotic Society of America, 1(1976), p312-328; Ed. by Pearson, C; and Hamilton-Faria, H.

[2] Pearson, C. "The Echelon Counter: A New Instrument for Measuring the Vocabulary Growth Rate and the Type-Token Relationship". Proceedings of the ASIS Annual Meeting, 17(1980), p364-366.

[3] Zipf, G.K. "Homogeneity and Heterogeneity in Language". Psychol. Record, 2(1938), p347-367.