GEORGIA INSTITUTE OF TECHNOLOGY
OFFICE OF CONTRACT ADMINISTRATION
SPONSORED PROJECT INITIATION

no action
ack 2/15
dd

Date: May 17, 1977

Project Title: Quality of Care and An HMO Automated Medical Record

Project No: G-36-620

Project Director: Dr. Shannon D. Brunjes

Sponsor: DHEW/PHS/NIH National Center for Health Services Research; Rockville, Md. 20857

Agreement Period: From 10/1/76 Until 9/30/77 (01 year)

Type Agreement: Grant No. 7 R01 HS02834-01

Amount:
$58,008 PHS Funds (G-36-620)
3,053 GIT Contribution (G-36-323)
$61,061 Total

Reports Required:
Annual Progress Reports with Continuation applications;
Terminal Progress Reports upon grant expiration.

Sponsor Contact Person (s):

Technical Matters

Contractual Matters (thru OCA)

Ralph L. Sloat
Chief, Grants Administration Branch
Center for Health Services Research
DHEW, PHS, NIH
Rockville, MD 20857
Phone: (301) 443-2810

Defense Priority Rating: None

Assigned to: School of Information and Computer Science (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
- Office of Computing Services
- Director, Physical Plant
- EES Information Office
- Project File (OCA)
- Project Code (GTRI)
- Other ______

GEORGIA INSTITUTE OF TECHNOLOGY OFFICE OF CONTRACT ADMINISTRATION

SPONSORED PROJECT TERMINATION SHEET

Date 3/8/82

Project Title: Quality of Care and An HMO Automated Medical Record

Project No: G-36-620

Project Director: Dr. Brunjes

Sponsor: DHEW/PHS/NIH National Center for Health Services Research; Rockville, MD 20857

Effective Termination Date: 9/30/78

Clearance of Accounting Charges: 9/30/78

Grant/Contract Closeout Actions Remaining:

None

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

Assigned to: ICS (School/Laboratory)

COPIES TO:

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

FORM OCA 10:781

SYNTACTIC STRUCTURE OF INFORMATION AND INFORMATION PROCESSES

Annual Progress Report

National Science Foundation Grant IST-7827002

VLADIMIR SLAMECKA
Principal Investigator

CHARLS PEARSON
Research Scientist

October 1980

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

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

October 24, 1980

Dr. Edward C. Weiss
Director, Research Program
Division of Information Science
National Science Foundation
Washington, DC 20550

Dear Dr. Weiss:

I am pleased to submit this annual progress report on our NSF grant IST 7827002. The report includes seventeen publications which comprise results of our work during the past fourteen months. The publications and our work fall into the following categories: 1) eidometer design and calibration; 2) experimental work, including the Law of Word Interpretation, information decay in immediate memory, and measurement of information transfer rates; 3) type-token aspects of word shape; 4) philosophy of methodology for information science and empirical semiotics; 5) theoretical methodology for information science and empirical methodology; 6) experimental methodology for information science and empirical semiotics; and 7) availability of computer programs developed under grant No. IST 7827002.

1. During the past fourteen months, the Mk V eidometer was designed, built and calibrated. Howell [8] reported on the experiments which led to the particular design philosophy adopted for the Mk V. Pearson [9] reported on the calibration experiments for the Mk V. The S/N 1 instrument achieved an assayed precision of 2.9 bpm. Pearson later [15] reported on an improved Mk V, the S/N 2 instrument and improvements to both the measurement procedures and the calibration procedures that increased the assayed precision of the Mk V to 3.8 bpm. Pearson [15] also reported on the preliminary design of the Mk VI vernier eidometer and the design of the experiments that will be required to obtain the design data for the Mk VI. The Mk VI is the ultimate goal of the eidometer task of this project since it is the instrument required for the final information transfer rate experiments. Its vernier scale is expected to yield an assayed precision of 5.0 bpm. Pearson [15] also defined a new eidontic deviance unit, the deviometer, which takes advantage of the additional measurement structure obtained since the definition of the °ED unit during the design of the Mk IV instrument four years ago. This new unit comprises an interval measurement scale compared to the former ordinal system used with °ED.

2. Much progress has already been achieved in the experimental aspects of the project, although the principal instrument used for this work, the McGowan 4-field teescope, has just been received as of last week and has not been

checked out yet. All experimental work with this instrument remains to be done. Pearson [2] reported the fundamental law of this investigation, the Law of Word Interpretation. He was next able to evaluate the measured deviations from the Law of Word Interpretation because of the assayed metrological data which he obtained on the Mk IV eidometer and the Polymetric 0959 teescope. These deviations are semiotically real and in [6] are attributed to information decay in immediate memory, together with a partial mathematical description of this relation. The background, theory, and progress of these experimental investigations of eidontic factors in information processing were given in [7]. Lo [14] carried out a literature search in connection with the final measurements of information transfer rates and made many suggestions that will be incorporated into the first experimental designs. He has reinterpreted the information decay effect to be a function of both immediate and short-term memory and is developing methods of separating these factors.

3. Part of the information processing aspects of word shape is brought out by the type-token relationship of words in natural language. Although no complete investigation of this can be carried out within the scope of this project certain investigations were necessary in order to meet the goals of the present project. In [11] Pearson announced the invention of a new instrument, the echelon counter, for measuring the type-token relation of words and published the results of a prototype instrument and preliminary experiment that was used to evaluate the concept. This instrument has exciting potential because of its high precision and low measurement noise. The TTKANAL program for measurements of type-token data by classical counting techniques was extended from the Burroughs B5700 computer to the Cyber-70 and [17] is an announcement of the availability of this new program to any interested investigator.

4. The work under this project and all others in the SemLab is guided by an evolving philosophy of methodology for information science and empirical semiotics. An evolving overview of this understanding of an appropriate methodology begins with [3] which discusses the problem of methodology for the trinary (semiotic) sciences as opposed to the binary (physical) sciences, and evaluates the role of the various kinds of scientific paradigms. This overview was further evaluated and placed into the context of all sciences while at the same time contrasting it with that of the binary sciences in [4]. In [13] Pearson outlined some of the challenges of developing a basic science of information science in a milieu in which information science is viewed predominantly as a technological discipline. Finally, [16] extended the overview to the present stage of its evolution. The paper identifies five separate classes of scientific paradigm that are important for the scientific methodology of information science and empirical semiotics, defines and discusses each, and gives contrasting examples of each from physics and information science.

5. Progress in the evolution of an appropriate theoretical methodology for information science and empirical semiotics which began under NSF grant No. GN-40952, continued with [10], discussing the attributes of information and showed that they could all be related to observable aspects of a theoretical concept, called the 'sign'. This was further elaborated in [12], showing that all the fundamental concepts of information science can be defined in terms of the basic concept of the sign and its attributes.

6. A sophisticated methodology for information science and empirical semiotics has been evolving over the last ten years. In [1], Pearson enlists all the cognitive sciences as an inventory of experimental paradigms for information science and empirical semiotics. In [5], a very general approach to experimental methodology for the trinary sciences is presented. Undoubtedly, this approach is not yet complete, and new understanding will develop during the current project.

7. Many computer programs have been developed as part of the ongoing work of this project. All of these which could be of general use to other investigators in information science and empirical semiotics have been made available without charge to any interested investigator. Phongphatar [17] is an example of an announcement that was released to journals in information science and semiotics announcing the release of one of our programs.

At the completion of the first fourteen months, this project is on time for undertaking an investigation of the effect of shape on the syntactic structure of information and information processing, as proposed in the grant document. The principal instrumentation, the McGowan 4-field teescope, has been designed, built and delivered. Three physiosemiotic instruments, an electromyograph, an electrodemograph, and a skin temperature sensor, have been procured. The Mk V eidometer has been designed, built, and calibrated and the preliminary design of the Mk VI accomplished. Mathematical studies of the data on decay of information in immediate memory is progressing.

During the final year of the project we have to checkout the teescope, develop experimental procedures for it, explicate the concept of interpretation errors, carry out a preliminary evaluative series of experiments, design the Mk VI eidometer, refine the mathematical description of the Law of Information Decay, and finally carry out all the final experiments leading to the Redundancy Curve for Natural Language, the Law of Word Interpretation, the Law of Information Decay in Immediate Memory, and the Measurement of Information Transfer Rates. This work is well underway and we anticipate its successful completion.

Sincerely yours,

Vladimir Slamecka
Principal Investigator
NSF Grant IST 7827002

cc: Office of Contract
Administration
Enclosures: Project Bibliography
Paper reprints (17)

<u>Project Bibliography</u>

1. Pearson, C. "The Cognitive Sciences as a Paradigm Inventory for the Development of Experimental Methodology in Semiotics". Presented to the Symposium on Empirical Semiotics sponsored by SIG/ES at the 1978 Annual Conference of the Semiotic Society of America; Providence, RI; October, 1978.

2. Pearson, C. "A New Law of Information: An Empirical Regularity Between Word Shapes and Their Interpretation". Presented to the session on Foundations of Information Science sponsored by SIG/FIS at the 1978 Annual Conference of the American Society for Information Science; New York City; October, 1978.

3. 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.

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. "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.

6. Pearson, C. "Information Decay in Immediate Memory: A Second Order Correction to the Law of Word Interpretation". 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.

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

8. 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, Ind; October, 1979.

9. 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, Ind; October, 1979.

10. 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.

11. 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, Calif; October, 1980. Published in Communicating Information: Proceedings of the 43rd ASIS Annual Meeting, 17(1980), p364-366.

12. 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, Calif; October, 1980. Published in Communicating Information: Proceedings of the 43rd ASIS Annual Meeting, 17(1980), p367-369.

13. 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, Calif; October, 1980.

14. 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.

15. 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.

16. 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.

17. 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.

THE COGNITIVE SCIENCES AS A PARADIGM INVENTORY FOR THE DEVELOPMENT OF EXPERIMENTAL METHODOLOGY IN SEMIOTICS

by Charls Pearson

Georgia Institute of Technology
School of Information and Computer Science

Preliminary version for comments only. Not for publication or quotation. All rights reserved by the author.

October, 1978

THE COGNITIVE SCIENCES AS A PARADIGM INVENTORY FOR THE DEVELOPMENT OF EXPERIMENTAL METHODOLOGY IN SEMIOTICS

by Charls Pearson

I. Motivation and Background

This paper is the second half of a two-part paper, the first half of which was presented in April 1978 at the Second Annual Conference of the Society for the Interdisciplinary Study of the Mind under the title "The Cognitive Sciences: A Semiotic Paradigm".

That paper noted that the problem of representation cuts across all of the cognitive sciences and appears to be a unifying concept for this emerging discipline. Representation is a semiotic concept in that a trinary relation called a 'sign' is involved essentially in the notion of representation. That is, using the word 'entity' in the broadest sense, one entity must interpret a second entity as representing a third entity. This indicates that representation may be a unifying viewpoint for the cognitive sciences and empirical semiotics may even offer our first hope of a unifying methodology for what is currently at best a highly heterogeneous problem area and at worst merely an interdisciplinary collection of somewhat related problems.

The April paper presented a preliminary analysis of representation from the semiotic viewpoint that demonstrated the possibility of using a common language to discuss problems within many of the individual semiotic sciences including all the disciplines presently recognized as impacting the cognitive sciences. One conclusion of this analysis is that there are many more disciplines impacting the cognitive sciences than commonly recognized. These include all of the semiotic sciences. Some major examples are: psychology, linguistics, logic, philosophy, sociology, economics, esthetics, theology, historiology, anthropology, communication science, information science, computer science, artificial intelligence, ethology, neuroscience, and hypnosis.

This analysis of representation treated it as a coding process and used an eighteen component theory of sign structure, called the Universal Structure Theory, to analyze the coding problems involved for each sign component. Each such problem analyzed turned out to be a problem that is of great concern to one or more of the cognitive sciences and to cognitive science in general. The semiotic dimensions of syntactics, semantics, and pragmatics, suggested by Morris, provide a convenient method of classifying and systematizing these problems. An initial attempt to carry this out suggested that cognitive science may be identified with syntactic and semantic semiotics.

We thus see that semiotics presents the possibility of using a common language to unify the cognitive sciences and for discussing their common problems. On the other side of the coin, the cognitive sciences present an inventory of paradigms for the development of experimental methodology in semiotics. It is this other side of the coin that is examined in the present paper.

II. Introduction

Semioticists must not make the mistake that Sherlock Holmes attributed to inspector Lestrade of forming our theories before the important facts are known and failing to revise them in the face of new evidence.* In fact, we do not need to form any theories at all until our experiments have begun to give us empirical laws.

Altho theory and experiment are intrinsically interfused — they cannot be logically separated — and we cannot design experiments until we have some kind of theory, we do not have a chicken and egg problem here. For we are every one of us given a theory of the universe and all of its phenomenas the moment we learn to speak our native language. Language is a theory of all those things we interact with and that are important to us for one reason or another. It is not a theory that has been arrived at by systematic reasoning, nor by controlled experiment. Nevertheless it does serve as our first approximate theory for any field of investigation.

Therefore, one need of any discipline attempting to transform itself from a speculative science into an empirical science is an inventory of experimental paradigms by which it may design and perform its experiments. Since semiotics now finds itself in exactly this situation, let us therefore look at semiotics' inventory of experimental paradigms.

III. The Initial Paucity of Experimental Paradigms

I would hesitate to say that semiotics has no experimental paradigms, but other than direct observation of sign processes, I can think of very few experimental paradigms used by either semioticians or semioticists. We may for purposes of this paper, idealize, and say at this point semiotics has an empty inventory of experimental paradigms. This places semiotics at the level of pre-Archimedian physics. How do we short circuit the approximately 2,000 years it took physics to build up a minimal basic inventory of experimental paradigms? This leads directly to the title of this paper.

IV. The Cognitive Sciences as a Paradigm Inventory for the Development of Experimental Methodology in Semiotics.

Many of the semiotic subsciences are much more mature as sciences than the parent discipline itself and especially is this the case with the cognitive sciences. Examples are experimental psychology, cognitive psychology, perceptual psychology, computer science, linguistics, etc. This claim of maturity is simply a claim that they have developed a more advanced empirical methodology, including an inventory of experimental paradigms, than semiotics.

Is there any way we can avail ourselves of this reservoir of experimental paradigms in the cognitive sciences? Well —— this is just the point I am trying to make. The answer is a resounding YES! The process is called

---

*I wish to thank Tom Sebeok for calling my attention to this example.

inverting the paradigm or "turning the sock inside out". It allows semiotics to adopt every single experimental paradigm in each of the cognitive sciences. The procedure is best understood by a progressive series of examples.

V. Example: Faraday's Electrolysis Experiment

For our first example I choose one that is not semiotic but is one well-known to all high school students and sets the procedure forth in bold relief. Consider Faraday's measurement of electric charge. This is a physical experiment and in fact is one at the base of the theory of electromagnetism. But it is itself based on the discovery that upon electrolysis, water decomposes into hydrogen and oxygen in volumetric ratio of two to one. This is usually regarded as a chemical experiment rather than a physical experiment.

Let us look at some of the parameters of this development.

Amount of charge was not measurable because the physical methodology had not yet been invented. But by using a fixed (even tho unmeasurable) amount of charge, chemists discovered that water, upon electrolysis, decomposes into hydrogen and oxygen. However, the weight and volume of oxygen and hydrogen could easily be measured by the then available chemical methodology. This led Faraday to reverse the point-of-view, or what I call "inverting the paradigm" — or what in more picturesque terms could be called "turning the sock inside out".

He simply used the newly discovered fact that water decomposes upon electrolysis and our well-developed chemical ability to measure the weight and volume of gases to define a unit of charge and develop a method of measuring the charge. The development of the ability to measure charge then led to experiments which ultimately led to such discoveries as Ohm's Law, etc. I am simplifying drastically of course for the purpose of my argument. But let us generalize the essentials of this simplified argument.

A controlled but unmeasurable unit or process (charge) in the parent field (physics) led to the discovery of a general result (electrolysis) in a distinct subfield (chemistry). This general result (electrolysis) was in turn then inverted (measurement of weight and volume of gases) and used as the means of measuring the originally unmeasurable unit or process (charge).

This process has been repeated many times between physics and its subfields. It is a process that is completely generalizable and holds between semiotics and all of its subfields.

VI. Example: Communication Synonymy of Words

My next example is less-known and slightly more complex — but it comes from semiotics and its subfield of psychiatry.

A control and a subject are seated in two separate rooms with no visual communication. There is one-way audio communication from the control to the subject by means of a microphone and loudspeaker. Pairs of words are displayed on a screen in both rooms, the same pair in both rooms. However, in the control presentation only, one word is underlined. The words are not distinguished in the presentation made to the subject. Example:

RUN WALK

The control now utters one word which may not be one of the two words, cannot rhyme with either word, and cannot give the position of the word on the display. For example, for this pair he might say "faster".

This paradigm can be used to measure the communication behavior of an open set of pairs of people by measuring their response times and error rates averaged over a fixed set of word pairs. By this means it has been found that pairs of persons who are good communicators by intuitive standards have a very low average response time and error rates in this experiment while pairs who are poor by intuitive standards have a very high average response time and error rates. For instance, it was discovered that pairs of schizophrenics typically have an exceptionally high average response time and error rate in this experiment.

When comparing the communication behavior of people by the above experiment, it is necessary to maintain the set of word pairs fixed because it is found that as the word pairs change, so will the response time. A pair of words which are intuitively nearly synonymous will take the same pair of persons longer to identify by this technique than a pair of words which are much less related. For example, most pairs of communicators take much less time and are more accurate in distinguishing NURSE from STONE than WALK from STROLL.

This suggests that the paradigm can be inverted and used to measure the communication synonymy of an open set of word pairs by selecting a fixed set of pairs of persons as standard and measuring the response time and error rates averaged over pairs of communicators. By this means semiotic phenomena which varies with synonymy may be investigated experimentally.

VII. Example: Eidometry and the Law of Word Interpretation: the Development of Semiotic Instrumentation

We move now to a more fruitful example. This is an area I have been actually working in for the past several years. It involves semiotics and psychology.

The Miller, Bruner, Postman (M-B-P) experiment measures the behavior of people interpreting artificial words. It is an attempt to measure and understand the Cattell Phenomena, discovered in 1885, that words are interpreted wholistically. In 1954 Miller used the uncertainty calculus, developed by Shannon in 1948, to model the Cattell phenomena in such a way that it could be measured, and hence investigated experimentally.

In a famous experiment, published by Miller, Bruner, and Postman in 1954, they discovered what is now called the Miller, Bruner, Postman Effect, that in one sense the interpretation of words depends drastically on the shape of words, but in another sense the interpretation of words is completely independent of shape. In other words, the interpretation of words depends on their shape in a very particular way. However, in their paper, M-B-P stated they could not measure shape, but could only control it.

By this point in the paper, that statement should sound familiar and the reader should know exactly what I propose to do with it. Word shape is a semiotic phenomena and the M-B-P Effect can be inverted so as to serve as a measure of word shape. An instrument, called the 'eidometer', was invented in order to make this measurement more precise and with its help the Law of Word Interpretation describing the relationship between word shape and interpretations was discovered. This is a semiotic law in that it describes the relation between the shape of signs and the interpretation of these signs.

VIII. Example: Proposed Measurement of Iconicity

In the last section we looked at an example of semiotic instrumentation and research under current active development. In this section we move to a proposal for future semiotic research and instrumentation development that is based on exactly the same reasoning as exhibited in each of the previous examples. However, it is one that is based on sound theoretical considerations, past experimentation, and a desire to answer important semiotic questions, and has been seriously proposed for active concentration, so it has not been made up purely as an illustration of the present concept. Hence, it serves as an excellent example of the fruitfulness of the principle of paradigm inversion.

Bernbach discovered a phenomena involving differences in the ability to access short and long term memory depending on whether iconic or symbolic signs were used for this access. Again, he could not measure the "iconicity" and "symbolness" of his signs; he could only control them. In fact, he used colors and color names in young children as his pure icons, and pure symbols. Our proposal in the SemLab was to use this "Bernbach Phenomena" in inverted fashion to measure the iconicity and symbolness of signs of mixed semantic structure.

In the Law of Word Interpretation, we had not gone directly from the phenomena to the law but had used the M-B-P Effect as an intermediary in developing our instrumentation. In this case the intermediary is missing and we attempted to overcome this disadvantage by following up a discovery of the information scientist P.J. Siegmann who suggested that the Bernbach Phenomena could be intermediated by a visual interference effect that would be affected only by the icons, and not by the symbols.

In a preliminary set of experiments the concepts of iconic squares and iconic circles were developed into a crude system of instrumentation and the existence of Siegmann's visual interference phenomena was confirmed. In addition, it appears that a simple metric may be imposed on iconic similarity, thus enabling a mathematical relation between visual interference and iconic similarity to be determined. Our proposal is thus to invert the Bernbach paradigm and to use the relation between visual interference and iconic similarity to measure the iconicity of signs of mixed semantic structure.

Once again, iconic circles are semiotic instruments, and iconicity is a semiotic measurement, and they have been obtained by inverting a paradigm from developmental psychology.

IX. Summary

Every single experimental paradigm in each of the cognitive sciences can be inverted to supply an experimental paradigm which may be developed for semiotics.

The Principle of Paradigm Inversion also enables us to distinguish between semiotics and each of the cognitive sciences. For instance: Psychology uses known properties of signs as a probe to investigate the structure of behavior, whereas semiotics uses the known structure of behavior as a probe to investigate the structure of signs.

It may be interesting to note in passing that the Principle of Paradigm Inversion is a corollary of the concept of semiotic reinterpretation, a concept that is useful in the development and interpretation of semiotic instrumentation; but it is beyond the scope of this paper to go into this relationship here.

According to our idealization made at the beginning of this paper, semiotics now has a very full bag of experimental paradigms and the job is therefore to determine what questions to ask, design the experiments and carry them out.

A NEW LAW OF INFORMATION: AN EMPIRICAL REGULARITY BETWEEN WORD SHAPES AND THEIR INTERPRETATION

By Charls Pearson
Georgia Institute of Technology

In 1948 [4] Shannon noticed a phenomena discovered earlier by Markov, that information sources producing words by a Markov-chain of finite order produce words that look more like natural language, the higher the Markov-order. In 1954 [2] Miller, Bruner, and Postman discovered another phenomena involving the interpretation of words produced by a Markov information source. As expected, the errors of interpretation under tachistoscopic conditions decreased, the higher the Markov-order; but the surprising result was that the amount of information (in Shannon's sense) obtained from a tachistoscopic exposure is independent of the order of approximation to English letter sequences. Therefore differences in the error rate can be predicted from a knowledge of the information structure of English.

Intuitively it would seem that these two phenomenas are related. However, without a quantitative concept of word shape and an instrument to measure it, it was not possible to study this relation experimentally. An instrument, called the 'eidometer', was invented based on Shannon's observations, and a concept of strangeness of word shape relative to a given natural language, called 'eidontic deviance', was developed based on a series of experiments with the eidometer.

The relation between eidontic deviance and interpretability has been studied with experiments involving the eidometer and a teescope. Thruout a wide range of word shapes this relation appears to be linear, leading to a new law of information which is a quantitative description of the Miller, Bruner, Postman effect.

One suggested application leads to a measurement of the redundancy function for natural language [5]. Shannon and others have made mathematical estimates of the upper and lower bounds for this curve, but it has never been measured before. The eidometer now gives us the capability for carrying out this measurement using the Law of Interpretation.

The teescope (tachistoscope) is a classical instrument of experimental psychology [2]. The one used in our studies to date is a Polymetric model 0959 two-field teescope. A vernier pot yields three digit precision for exposure setting, while a three position range setting switch gives us a total exposure range of 10 to 10,000 milliseconds. The final experiments were carried out on the shortest range of 10 to 100 ms.

The eidometer is an empirical explication of the intuitive judgment that LYDRA looks more like a printed American word than does WQQXZTKL, or the latter looks stranger to a native literate American than the former. The Mk IV design consists of a set of eighteen lists of artificial words. Each list consists of about eight to twelve words which have been chosen to look about equally strange. These lists were constructed by a series of experiments stretching over the last five years.

The lists are mounted on a display panel with the strangest list (0.00 °D) mounted to the extreme left and the least strange list (3.00 °D) mounted to the extreme right. Outside of the extreme lists are two blank lists (one with the value -0.20 to the left of 0.00 and one with the value 3.20 to the right of 3.00) to remind subjects that measured values can in fact fall outside of the instrument span. The list values increase from -0.20 to 3.20 by increments of 0.20 °D, thus resulting in a span of 3.40 °D. The eidontic deviance of a word is measured by finding two adjacent lists such that most of the words on one list look generally stranger than the specimen word even though a few may look better, and most of the words on the adjacent list look less strange although a few may look stranger. The measured value is obtained by interpolating between the values of these two lists, thus resulting in three digit precision.

The design, construction, and operation of the eidometer was reported in [3]. The Mk IV, SN#1 has a precision of 4 bpm, a reliability of 3 bpm, and a validity of about 1 bpm, where bpm stands for bits/measurement. These concepts were defined and the experiments leading to these values reported in [3]. Experiments are now underway to increase the performance of the eidometer and a Mk V design is nearing completion. It is expected to be a substantial improvement over the Mk IV design.

The study was conducted by generating a set of 100 standard specimens, each eight letters long, using ten different information generators. Four of these were of fixed Markov orders 0, 1, 2, and 3. The other six were variable-mask generators of shape (0,1), (1,0), (1,2), (2,1), (2,3), and (3,2), [1]. Ten adult, native literate S's measured each of the 100 specimen words. Each S measured the strangeness of the words in a different random order using the eidometer. The ten readings for each word were averaged to get the final deviance for each word. Ten completely different adult, native literate S's measured the placement errors obtained by viewing the same 100 words in the teescope with the procedure used in [2]. Each S was calibrated on two sets of twenty warm-up words to get a measure of his perceptual response rate and to adjust the exposure time of the teescope for maximum sensitivity to that S. The teescope was adjusted after each warm-up trial of twenty words to obtain as nearly as possible a 50% error rate and the second adjustment was usually very minor. After the final run for each S on the 100 specimen words, any deviation from 50% for that S was corrected for by dividing the error rate for each word by the ratio of that S's overall error rate to the ideal value of 50%. For each word, these corrected values were then averaged to get the final measured interpretation error rate. A linear correlation between the eidometer measurements and the teescope measurements gave a correlation of

$$r = 0.67$$

with r^2 about 45% for a validity of about 1 bit. For 97 degrees of freedom this is significant at the .001 level. Sample data values are shown in Table 1.

Miller, Bruner, and Postman originally attributed the interpretation error effect to perceptual information processes [2]; however,

Table 1
Sample Interpretation Data

#	Word	Order	°D	Error Rate
3	STIONYTH	3	2.02	3.47
70	DENNMSAO	1	1.74	3.71
28	OOKORGED	(3,2)	1.36	4.35
87	PDXPUYJ-	0	0.44	5.21

Sperling has more recently interpreted it as an information process associated with immediate memory. The eidometer, however, still appears to be measuring our perception of the strangeness of shape of the sign relative to a given natural language, and so this law would appear to relate a perceptual information process to an information process associated with immediate memory.

The conclusions to be derived from this study are: 1) that the strangeness of the shape of artificial words relative to a given natural language may be reliably measured; 2) that this measurement gives us the means of describing quantitatively the Miller, Bruner, Postman effect; and 3) that in the range of data studied, this description is best given by a simple linear equation,

$$E = a + bS$$

where E is the tachistoscopic error rate, S is the strangeness measured in eidontic deviance units, °D, and a and b are constants.

Since this is an empirical generalization relating the measurements of two different information processes, I call this a law of information, specifically, the Law of Interpretation.

By claiming the Law of Interpretation as an information law, I do not claim that I have anything like a theory or even a conjectured explanation of this law, but only that I know how to objectively reproduce and quantitatively describe a pervasive regularity in our observation of information processes. However, we may speculate that this law will give us a tool that will eventually help us to decipher the way the human processor codes information in the mind, and/or will lead us to a fuller understanding of the structure of information in that any theory of information structure will eventually have to account for this law. And as mentioned in the introduction, it gives us for the first time a method for measuring the redundancy curve for natural language.

References:

1. Flowers, J. "Progress in Semiotics Software at Georgia Tech's Sem-Lab". The Semiotic Scene, 1(1976).

2. Miller, G.A.; Bruner, J.S.; and Postman, L. "Familiarity of Letter Sequences and Tachistoscopic Identification". Journal General Psychology, 50(1954), p129-139.

3. Pearson, C. "An Empirical Concept of Word Shape". Paper presented at the Second Annual Conference of the Semiotic Society of America; Denver, Colorado; October 15-16, 1977. To appear in the published proceedings.

4. Shannon, C.E. "A Mathematical Theory of Communication". Bell System Technical Journal, 27(1948), p379-423, and 623-656.

5. Shannon, C.E. "Prediction and Entropy of Printed English". Bell System Technical Journal, 30(1951), p50-64.

THE PROBLEM OF COMMUNICATING RESULTS IN EMPIRICAL SEMIOTICS

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

June 1979

Preliminary draft for comments only. Not for publication or quotation. All rights reserved by the author. Intended for presentation at the SIG/ES Workshop on Immediate Problems in Empirical Semiotics to be held at the Second International Semiotics Congress in Vienna, Austria; July 2-6, 1979.

THE PROBLEM OF COMMUNICATING RESULTS IN EMPIRICAL SEMIOTICS

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

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.

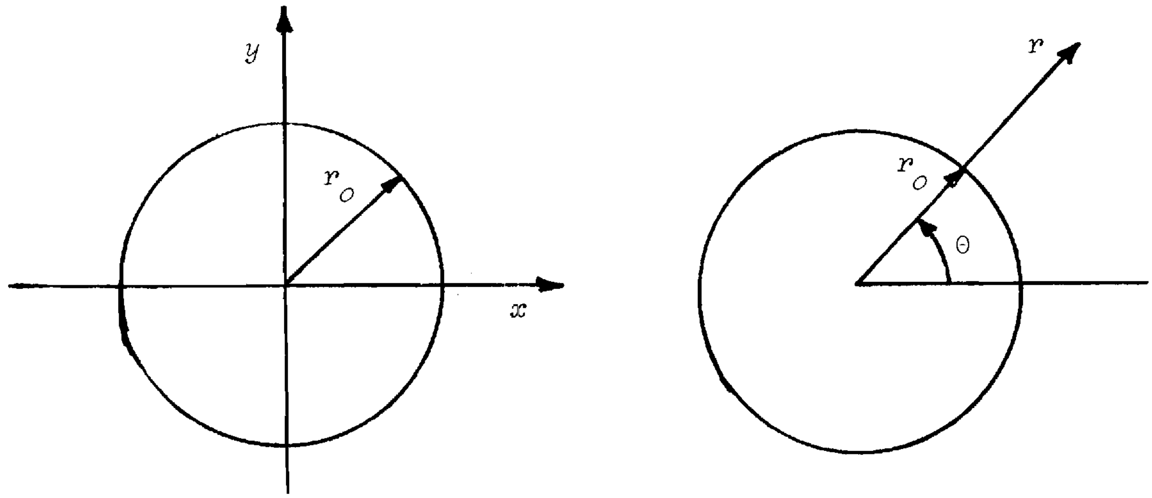
Scientific communication -- the communication of precise and rigorous scientific results -- requires the existence of universally agreed-upon paradigms -- or at least universal agreement on what the disagreed-upon paradigms are -- in order to take place effectively. In the present state of empirical 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.

Background:

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.

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 phenomena are 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. Linguistic paradigms 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 1a shows a circle as described by rectangular coordinates and gives the corresponding algebraic equation. Figure 1b shows the circle as described by polar coordinates, and the much simpler algebraic equation associated with the polar description.



$x^2 + y^2 = r_o^2$

a)

$r = r_o$

b)

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

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 language. 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.

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); Maxwell's Electromagnetic Theory; etc. Theories compete empirically on the basis of their ability to explain known phenomena, 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.

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 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.

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, 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 phenomena 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 technology is playing much the same role as did steam engineering in 19th century physics. We should also 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.

Proposed Solution:

Some way must be found to break the 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. My proposal is exceedingly simple and recognizes all of the above stated problems and objectives. At first we must just do that which we can do. It is not possible, or necessary, to do more.

Each of us in presenting results in empirical semiotics must state our own paradigms. In many 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. Often a single sentence can fulfill this obligation, viz: "My experiment was designed to answer a question stemming from Morris's Theory of Sign Structure as framed within Peirce's Language of Logical Analysis; the experiment was designed using only standard techniques of the statistical theory of inferential models; and data analysis requires only simple algebra and inferential statistics".

Since most papers in empirical semiotics emphasize only one of the five paradigm types of empirical language, theory, experiment, mathematical analysis, or application, my proposal is specifically to mention, or state explicitly, the three or four paradigms other than the one being specifically discussed in the paper. Thus as examples: 1) a paper on the language or philosophy of empirical semiotics would name the theoretical paradigms, the experimental paradigms, and the mathematical paradigms this result was intended to aid; 2) a paper in theoretical semiotics would name the language paradigm in which the theory is set, the experimental paradigm for which the theory is required, and the mathematical paradigm required for stating and using the theory; 3) a paper in experimental semiotics would name the linguistic paradigm in which the paper is presented, the theoretical question the experiment answers and the theoretical paradigm which motivated that question; and the mathematical technqiues used in designing the experiment and in analyzing the data; and 4) a paper in mathematical semiotics would name the empirical language in which the work took place, the theoretical or experimental problem the mathematical technique is intended to solve, and the experimental or theoretical paradigm respectively, related to that problem. Each of the foregoing would name the appropriate applications paradigm, if pertinent. A paper in applied semiotics would name all four of the other paradigms involved.

Outlook:

If this proposal is adopted for the presentation of papers in experimental 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.

THESES OF EMPIRICAL SEMIOTICS

by Charls Pearson

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

Presented at the Theses Session held at the Second International Semiotics Congress in Vienna, Austria; July 2-6, 1979; and submitted for inclusion in the published proceedings of the congress.

©

Copyright December, 1979
by Charls Pearson

THESES OF EMPIRICAL SEMIOTICS

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

ABSTRACT

1. Semiotics is an empirical science and there is no epistomological difference between it and the other natural sciences, such as physics, chemistry, and psychology. The principal difference is one of subject matter and methodology only. Physics and the physical sciences study properties of nature that can be modeled by binary relations while semiotics and the semiotic sciences study properties of nature that must be modeled by trinary relations.

2. As in any other empirical science, semiotic knowledge is gained only by the scientific observation of nature. The scientific observation of nature implies the design, execution, and quantitative analysis of rigorously controlled experiments to answer intellectually interesting questions; all taking place within the framework of an explicit and mathematically specified theory.

3. 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.

4. It is only by the interplay of theses (2) and (3) within the framework of thesis (1) that basic semiotics can develop and mature as an intellectual discipline. All other approaches are at best only supportive of this approach (historical semiotics, semiotic pedagogy, etc.) or at worst distracting, disruptive, and divisive (literary criticism, humanistic essays, and undisciplined speculation).

5. The basic subject matter of semiotics is the general structure of signs and sign processes.

The Universal Sign Structure Theory proposed by Pearson is an initial attempt to develop a scientific theory of sign structure within the framework of theses (1) thru (5).

THESES OF EMPIRICAL SEMIOTICS

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

1. Introduction:

These theses were presented in the format of a "poster session" which consisted of the announcement of a set of theses and the subsequent defense of them in open discussion with the audience. Accordingly, after the above five theses were announced and the following short discussion presented, the floor was opened for discussion. However, almost all of the discussion time was taken up with questions and answers concerning the Universal Sign Structure Theory. It appeared that those in attendance accepted the theses of empirical semiotics as uncontroversial and wanted to know more about recent substantive results in empirical semiotics.

2. Semiotic Methodology:

The first thesis of empirical semiotics is that semiotics is an empirical science and there is no epistomological difference between it and the other natural sciences, such as physics, chemistry, and psychology. The principal difference is one of subject matter and methodology only. Physics and the physical sciences study properties of nature that can be modeled by binary relations while semiotics and the semiotic sciences study properties of nature that must be modeled by trinary relations.

As can be seen from this thesis, the word "empirical" refers to the methodology of empirical science, rather than to the seventeenth century school of philosophy called "empiricism". The adjective form "empirical" is also sometimes improperly confused with "experimental". But this should be carefully distinguished since "empirical" means knowledge methodologically grounded in the phenomena and includes conceptual, mathematical, and theoretical as well as, experimental methodologies. A better name for this approach to doing semiotics may have been "scientific semiotics". However, there are problems with this name also, including such trendy misuses of the word "science" as "library science", "computer science", and "the engineering sciences". Therefore the term "empirical semiotics" was chosen by the SSA special interest group SIG/ES as a compromise to denote the approach to semiotic knowledge thru the methodology of empirical science.

There is a distinct difference between the binary sciences and the trinary sciences -- between physics, chemistry, and meteorology, etc. on the one hand, and semiotics, psychology, sociology, etc. on the other. Just as physics has become the paradigm for the binary, or physical, sciences, so semiotics may be the best paradigm for the trinary, or semiotic, sciences. In fact, the principles of Paradigm Inversion, and Semiotic Reinterpretation go a long way towards enabling semiotics to integrate, unify, and systematize the social and behavioral sciences plus a lot more studies that are not yet widely regarded as sciences, such as theology, esthetics, ethics, historiography, etc.

In adopting methodologies from the physical sciences, semiotics must be careful to distinguish between those methods dependent only on the basic epistomology of science, and those that depend on the binary nature of the particular relations studied. Those methods that depend only on the scientific outlook are likely to work for semiotics, while those that depend on the binary nature of the physical sciences are likely not to work for semiotics. At any rate, all methodologies contemplated by semioticists must be self-consciously tested and analyzed for objectivity, reliability, and rigor.

In borrowing methodologies from the trinary sciences we are on firmer ground. Altho even these methodologies must be tested and analyzed, the Paradigm Inversion Principle and the Principle of Semiotic Reinterpretation give us guidance in the explicit design of empirical methodology for semiotics that is most likely to work in the sense of science (i.e., objectivity of observation, reliability of knowledge, and rigor of understanding).

One of the more important categories of methodology is the mathematical methods. It is here that we are likely to deviate farthest from the physical sciences. We must remember that most of present-day mathematics was developed at the specific instigation of the physical sciences. All of the standard concepts of current mathematics can be reduced to binary relations* except that of "operation", and even this is almost always treated in a pseudo-binary manner. The mathematical methods for treating trinary relations have not yet been invented. We know they will look very much different from the methods used to treat binary relations. For instance, the basic definition of a trinary relation is "any subset of the set of all ordered triples $<a,b,c>$ whose first element a comes from a designated first set A, whose second element b comes from a designated second set B and whose third element c comes from a designated third set C". If we denote this set of all ordered triples in the usual fashion as $A\times B\times C$, we must remember that this is not equal to either $(A\times B)\times C$ or $A\times(B\times C)$. The equality of all three of these sets is often assumed for most binary mathematics. However, even these new mathematical methods can not be developed apriori. They must be developed at the instigation of specific empirical problems and their logic teased out in such a way as to enable practical methods to be developed for solving these problems. It should be recognized that some of these problems may appear first in the applied areas of semiotics such as, for instance, literary criticism, library science, and computer science.

3. Experimental Semiotics:

The second thesis of empirical semiotics is that, just as in all empirical sciences, semiotic knowledge is gained only by the scientific observation of nature. The scientific observation of nature implies the design, execution, and quantitative analysis of rigorously controlled experiments to answer intellectually interesting questions; all taking place within the framework of an explicit and mathematically specified theory.

---

*Excepting, of course, those concepts upon which the concept of 'binary relation' is grounded; these being principally 'set', 'extension', and ordered n-tuple'.

The concept of 'experiment' with its attendant concept of 'rigorous control' is an evolving one. Concepts, such as 'rigorous control', 'measurement tolerances', etc. always depend on the current state of knowledge and technology. The current states of semiotic knowledge and technology are not very mature, but neither are they totally absent. We must apply the available knowledge and technology to the designing of experiments to answer new questions of interest rather than starting over from scratch on each new question. Theoretical questions must be translated into experiment by rigorously disciplined thinking, rather than using only undisciplined speculation. This implies that empirical theories of semiotics must contain within themselves rules of translation from the theoretical realm into the observable, or experimental realm. A semiotic system containing only theoretical terminology and relations among the theoretical terms, but no rules whereby one can translate from theoretical terms into experimental terms, is not a theory of semiotics. It may be a language system, or a model, but it is not a theory in the empirical sense of that word. The state of semiotics today is one of too many ungrounded language systems and models, and not enough empirically grounded theories.

Fortunately we know quite a bit about methodology for the design and execution of semiotic experiments. The Paradigm Inversion Principle and the Principle of Semiotic Reinterpretation tell us how to adopt proven and assayed methodology from each of the other semiotic sciences. Psychology, especially in its information processing aspects, has developed a quite sophisticated experimental methodology, almost all of which can be adapted to the experimental investigations of semiotic questions.

4. Theoretical Semiotics

The third thesis of empirical semiotics is that, 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.

5. The Growth of Semiotic Knowledge:

The goal of empirical semiotics is the scientific understanding of sign processes, also called information processes. Therefore, the fourth thesis of empirical semiotics is that it is only by the interplay of the second and third theses within the framework of the first thesis that basic semiotics, or basic information science, can develop and mature as an intellectual discipline. All other approaches are at best only supportive of this approach (historical semiotics, semiotic pedagogy, etc.) or at worst distracting, disruptive, and divisive (literary criticism, humanistic essays, and undisiciplined speculation).

The scientific understanding of sign processes requires the development of semiotic theories. This is the task of theoretical semiotics. But because of the fallibility of the human mind and the incompleteness of our total state of semiotic knowledge at any given finite time these theories must be subjected to test and confirmation. This is the task of experimental semiotics.

As experimental results become more and more refined, theories must be refined and revised in order to maintain the empirical fit between theory and observation. And since experimental methodology is a function of the current state of theory, as our theories improve, so will the experimental methodology. This closes the circle and shows that the process of obtaining semiotic knowledge is really an interplay between theory and experiment in an environment of empirical methodology which is constantly being criticized, assessed, and refined.

---

*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.

Historical semiotics can search out the past traces of how our present semiotic knowledge was obtained, but it cannot generate new semiotic knowledge. It is of great importance to keep our past firmly in mind in order to lend direction to the future and to give balance and insight to the present. But it is also essential not to confuse assessment of past knowledge with attainment of new knowledge. It is essential to give the highest priority to the main goal.

By a similar argument, pedagogy can enable us to teach current semiotic knowledge to our followers, but it does nothing to develop new semiotic knowledge, except possibly to better prepare our followers to search for new knowledge. If the development of knowledge is to be a continuing, ongoing process, this latter is, of course, necessary; however, it is obvious, that at some point, some emphasis must be placed on the development of new knowledge, or the entire pedagogy process is to no avail.

Literary criticism, classical studies, etc. are studies involving particular, individual facts. As such, these studies do not contain the generality necessary for the scientific approach to knowledge. They represent the natural history, case study, or clinical study stage of the knowledge development process. This is necessary, of course in the prescientific stage of any discipline, but a maintenance of this attitude past the initial stages of scientific development acts as an inhibitor to the growth of scientific knowledge. An example of this is psychotherapy, in which the concentration on clinical knowledge to the exclusion of any interest in scientific experimentation and theory building has deterred this field from reaching anything like its promised potential.

In the 20th century, semiotics has reached the initial stages of scientific formulation, with the development of rudimentary scientific theories and experimental methodology. We have now reached the point where research interest should be dictated by the goal of refining theory and experimental methodology. The continuance of the natural history approach has now become an anachronism.

There is also another, more important, objection to calling such studies as literary criticism, classical studies, etc. semiotic research. But this can be better seen if we first remove the objection stemming from the concentration on individual isolated facts. The study of the general relations involved in literary criticism has been known by various names including structuralism, poetics, etc. In fact, recent works under the rubric of structural poetics have claimed to be concentrating on gaining an empirical, scientific, understanding of the literary criticism process in general. There should be no objection to this approach on the natural history/case study/clinical score.

However, the question is now one of subject matter. The subject matter of semiotics is the structure of signs and sign processes in general (see the next thesis). The subject matter of structural poetics is the structure of a particular sign process -- literature. Therefore this is to be regarded as part of esthetics -- one of the semiotic sciences -- and not part of general (or basic, or "pure") semiotics itself.

This objection is easy to misunderstand. I am not complaining that esthetics in general, or structural poetics in particular, should not be studied. Quite the contrary; I have recently developed considerable interest in structural poetics. I am only maintaining that we need to recognize the distinction between general, or pure, or basic semiotics on the one hand, and the applications of basic semiotics to the other semiotic sciences on the other hand. Within semiotics itself, there is so much need for answers to so many questions that we need to concentrate on the questions of pure semiotics and leave the applications to the applied disciplines as for instance esthetics, sociology, anthropology, psychology, theology, etc.

There is another possible misunderstanding of this objection. There is a great deal of interaction between pure semiotics and the semiotic sciences and it is often quite hard to draw the line between them. Just where semiotics ends and structural poetics begins is hard to determine and often a matter of personal judgement. I am not saying there is no interaction, or that semiotics cannot benefit by advances in structural poetics. In fact, I think there are tremendous benefits to be obtained in both directions. But the decision as to whether we are to devote the time and resources within semiotics to a particular interactive study should depend not on who needs financial help from the woefully small amount of funds available for semiotic research, but on which direction the benefit lies. A study that uses semiotic knowledge to explore a problem in structural poetics should be justified on poetic grounds and should be supported from funds within poetics. Many interactive studies would still fall within the domain of semiotics under this guideline. A study, for instance, which analyzed a poetic implication of semiotic theory, compared it with the empirical facts from literary criticism (people can recognize plot summaries as being summaries of the same or different plots, etc.) and then used this to refine the basic semiotic theory would find no trouble being recognized as semiotic research.

Needless to say, humanistic essays and undisciplined speculation have no place in the search for, and development of, reliable knowledge of any kind. Such studies may fit within philosophy or literature where they may have some bearing on developments which impact semiotics. But at least they should be so regarded and judged according to philosophical or literary standards and not masquerade as semiotics, or information science.

6. The Subject Matter of Semiotics:

The fifth thesis of empirical semiotics is that the basic subject matter of semiotics is the general structure of signs and sign processes.

The most fundamental theoretical concept of semiotics is that of the 'sign'. Hence the basic subject matter is the structure of this concept and according to the above theses semiotics should concentrate on developing theories of sign structure and structure of sign processes, and experiments to test, refine, and revise these theories.

The only possible confusion pertaining to this thesis involves the word 'structure'. French structuralism has incorporated this important word into the name for a particular approach to investigation. As used within empirical semiotics, 'structure' means simply the mathematical models of experimentally discovered relations involving signs and/or sign processes and the mathematical theories that explain these relations.

7. A Claim for Empirical Semiotics:

Subsidiary to the above theses of empirical semiotics is the claim that the Universal Structure Theory proposed by Pearson is an initial attempt to develop a scientific theory of sign structure within the framework of these five theses. The best known version of this theory establishes a basic abstraction called a Universal Sign Model which contains 18 components. All signs can be derived from this basic structure by various modes of combination that are determined by the three principles of the theory. As an example, Peirce's nine kinds of signs are derived from the three principles by means of nine representation theorems. The theory has been very useful in motivating many experiments on basic structure of signs, incuding syntactic structure, semantic structure, and pragmatic structure.

8. Acknowledgements:

The work reported herein was sponsored in part by grant IST-7827002 from the National Science Foundation's Division of Information Science and Technology. This support is gratefully acknowledged. I also wish to thank my colleague Vladimir Slamecka for various discussions and encouraging support.

EMPIRICAL METHODOLOGY IN INFORMATION SCIENCE AND SEMIOTICS

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

Presented at the Workshop on Fuzzy Formal Semiotics and Cognitive Processes held at the Second International Semiotics Congress in Vienna, Austria; July 2-6, 1979; and submitted for inclusion in the published proceedings of the congress.

©
Copyright December, 1979
by Charls Pearson

EMPIRICAL METHODOLOGY IN INFORMATION SCIENCE AND SEMIOTICS

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

ABSTRACT

This paper attempts to clarify several issues involving scientific methodology in semiotics and information science. As such 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. Since it is a generally accepted part of the methodology of all other sciences that quantification and formalization must be built on a sound basis of empirical investigation, rather than vice-versa, this paper assumes the desirability of also adopting this procedure in semiotics and then examines four specific problems stemming from this adoption: 1) how to design and carry out a concrete, rigorous experiment to measure something so ephemeral as a sign, its meaning, its information, and its structure; 2) how to perform experiments before quantification is established, or a formalized language exists; 3) how to develop quantification and formalize language out of experimental results; and 4) how to interpret a semiotic or information measurement before we have a formalized system.

The development involves a key distinction between 'quantitative' and 'empirical' and introduces two new methodological principles to semiotics. These are the Principle of Semiotic Reinterpretation, and its corollary, the Paradigm Inversion Principle.

EMPIRICAL METHODOLOGY IN INFORMATION SCIENCE AND SEMIOTICS

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

1. Introduction:

The theme of this workshop: the formalization, or quantification, of a part of semiotics, has raised several serious questions concerning ontology, epistomology, and methodology in semiotics, but most especially concerning empirical methodology. While I can't begin to answer these questions in the time available, this discussion is intended to help clarify some of the issues that have been raised and point out the direction in which the answers lie. Insofar as information science has traditionally been associated with that part of semiotics which is concerned with observation and measurement, what I have to say will apply to information science and semiotics equally.

This paper 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.

This discussion is assembled in part from various notes and comments which I have made before other groups, most notably the American Society for Information Science, the Semiotic Society of America, the Society for the Interdisciplinary Study of the Mind, and various of my classes at Georgia Tech. I wish to thank the Semiotic Society of America's Special Interest Group for Empirical Semiotics (SIG/ES) for acting as a sounding board for my ideas and especially Pranas Zunde and Vladimir Slamecka of Georgia Tech's School of Information and Computer Science for various discussions and encouragement. All work reported herein was carried out in the SemLab, Georgia Tech's lab for experimental semiotics, and supported in part by the NSF under research grant IST-7827002. This help is gratefully appreciated.

2. Background:

Despite its long background and eminent history, semiotics has not evolved into a hard, empirical science such as, for instance, chemistry or psychology. In fact, vagueness, ambiguity, and undisciplined reasoning are still pretty much the rule rather than the exception in semiotics and information science. Charles Peirce attempted to correct this in part in the nineteenth century and the opening years of the twentieth century with his logical "semiotic". But even the great Peirce cited the need for rigorous, empirical investigation before formalization could take place.

Attempts to develop a quantitative, empirical approach to semiotics belong to very recent history. So far as I know, none were made before the present century. Therefore the purpose of this workshop, to explore the possibilities of making a portion of semiotics more precise and less ambiguous, by formalizing, or quantifying it, is of the highest merit. However, by the same reason of recentness of concentration, the relationship between quantification, formalization, and empirical methodology has not been explored extensively within semiotics, altho it has been explored somewhat within the philosophy and methodology of other sciences.

One thing that semiotics can learn from the methodology of these other sciences is the necessity for building quantification on a sound base of empirical investigation rather than vice-versa. The question to be discussed in this paper concerns the nature of this empirical methodology, how it can be created out of whole cloth, so to speak, and its relationship to the quantification problem. In particular the question will be discussed of how a concrete, rigorous experiment can be designed and carried out to measure something so ephemeral as a sign, its meaning, its information, and its structure; how experiments can be performed before quantification exists; how quantification can then be developed out of the experimental results; and how we can interpret a semiotic or information measurement before we have a formalized system.

I begin with the first of these questions in section 5, but first I prepare the way in section 3 by introducing the key distinction between the two concepts of 'quantitative' and 'empirical'; and in section 4, I discuss the nature and role of linguistic paradigms in science, also as preparation for the major development of the paper.

The major argument of the paper begins in section 5. In this section I conclude that semiotic measurement is not as ephemeral as it appears at first glance, and that far from being scarce, semiotics starts with a full bag of experimental paradigms. To do this I introduce the Principle of Paradigm Inversion.

In the discussion of the nature of experimental paradigms for semiotics and information science, a new question is raised concerning the distinction between experiments in semiotics and information science and experiments in various of the individual semiotic sciences. In section 6, this question is discussed, and clarification is provided using an example involving similar experiments in semiotics and psychology. In this section the Principle of Paradigm Inversion is further explicated.

Since the conclusion of section 6 is that the same experimental paradigm can be used for either a semiotics experiment or for an experiment in one of the semiotic sciences, depending on the type of measurement involved, a final question is raised concerning the nature of measurement in semiotics and information science. In section 7 semiotic measurement is explicated using the Principle of Semiotic Reinterpretation.

Section 8 summarizes the foregoing results and notes that the results of sections 5 and 6 are actually corollaries of the Semiotic Reinterpretation Principle which thus turns out to be a very powerful principle in the empirical methodology of semiotics.

3. The Distinction Between 'Quantitative' and 'Empirical':

It is important to keep the distinction between empirical semiotics and quantitative semiotics clearly in mind, especially in a workshop on formalization of semiotics, because any study can be made quantitative in a multitude of trivial ways. The quantification problem involves how to make a study quantitative in empirically interesting ways and therefore the empirical question is prior to the quantitative question.

The empirical question involves the search for, and discovery of, empirically interesting, real, objective, generality of nature. The empirical interest of a real, objective, generality is determined by the number of known relations it enters into with other known real, objective, generalities of nature. As mentioned above, the quantification question involves how to make these relations quantitative in empirically interesting ways. Here, the empirical interest involves making the relationships between relations quantitative as well.

We must therefore expect that our languages of empirical semiotics will grow out of our attempts to solve empirical problems, including experiment, theory, and quantification interrelatedly, and not out of purely apriori quantification. Let us look at the nature of this problem a little closer. The decision as to what languages we shall use in going about our empirical business is called the linguistic paradigm.

4. Linguistic Paradigms:

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 phenomena are 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. Linguistic paradigms 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. Solution procedures may be substantially easier to think out in a different language than the usual one, etc.[I] 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 language. Like other empirical paradigms, linguistic paradigms evolve as a result of our experience in using them and occasionally go thru Kuhnian revolutions. But these revolutions are also a result of our empirical experience with using the paradigms. 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; 4) Professor Nowakowska's language of quantification; and 5) my own Language of Menetics.

Semioticists must not make the mistake that Sherlock Holmes attributed to inspector Lestrade of forming our theories before the important facts are known and failing to revise them in the face of new evidence.* In fact, we do not need to form any theories at all until our experiments have begun to give us empirical laws. This admonition applies to language revision as well.

---

*I wish to thank Tom Sebeok for calling my attention to this example.

Insert: But we can only find this out by empirical experience.

Altho language, theory, and experiment are intrinsically interfused -- they can not be logically separated, and we cannot design experiments until we have some kind of language and theory to work with -- we do not have a chicken and egg problem here. For we are every one of us given a theory of the universe and all of its phenomenas the moment we learn to speak our native language. Natural language is a theory of all those things we interact with and that are important to us for one reason or another. It is not a language designed to facilitate scientific investigation nor a theory that has been arrived at by systematic reasoning and controlled experiment. Nevertheless it does serve as our first language and approximate theory for any field of investigation.

A linguistic paradigm can be no more quantitative, or precise, than our empirical understanding of the concepts dealt with. This implies the necessity of developing an experimental understanding of concepts before attempting to quantify them into the language. And this raises the problem of how to design and carry out experiments in semiotics before we have quantification and measurement. This problem is discussed in the next section. The language must fit the outline of past experience and be suitable for generating new theoretical hypotheses, interesting empirical questions, and rigorous experimental designs to test these hypotheses and answer these questions. Therefore the language must evolve within the *Zeitgeist* of current experimental results, current empirical theory, and what is now in the process of evolving: not what will be the ultimate ideal, but what will satisfy the immediate requirements. In fact, there is a very real epistemological problem of how we could ever know what the ultimate ideal is. This problem has been likened to that of trying to wash a dirty glass using nothing but dirty water and a dirty rag; a very apt description of the scientific method.

Therefore, an early need of any discipline attempting to transform itself from a speculative study into an empirical science is an inventory of experimental paradigms by which it may design and perform its experiments. Since semiotics now finds itself in exactly this situation, let us therefore look at semiotics' inventory of experimental paradigms. In doing this we shall also find out how we can experiment before we can measure and how we can measure before we quantify.

5. Experimental Paradigms:

I would hesitate to say that semiotics has no experimental paradigms at all, but other than direct observation of sign processes, I can think of very few experimental paradigms used by either semioticians or semioticists. We may for purposes of this paper, idealize, and say at this point semiotics has an empty inventory of experimental paradigms. This places semiotics at the level of pre-Archimedian physics. How do we short-circuit the approximately 2,000 years it took physics to build up a minimal basic inventory of experimental paradigms, especially in view of the non-concrete, ephemeral nature of signs, their meanings, their information, and their structure?

Many of the semiotic subsciences are much more mature as sciences than the parent discipline itself and especially is this the case with the cognitive sciences. Examples are experimental psychology, cognitive psychology, perceptual psychology, computer science, linguistics, etc. This claim of maturity is simply a claim that they have developed a more advanced empirical methodology, including an inventory of experimental paradigms, than semiotics.

Is there any way we can avail ourselves of this reservoir of experimental paradigms in the cognitive sciences? Well -- this is just the point I am trying to make. The answer is a resounding YES! The process is called inverting the paradigm or "turning the sock inside out". It allows us to adopt every single experimental paradigm in each of the cognitive sciences. The procedure is best understood by a progressive series of examples. The concept of paradigm inversion itself will be further discussed and partially explicated in section 6.

For my first example, I choose one that is not semiotic but is well-known to all high school students and sets the procedure forth in bold relief. Consider Faraday's measurement of electric charge. This is a physical experiment and in fact is one at the very base of the theory of electromagnetism. But it is itself based on the experimental discovery that upon electrolysis water decomposes into hydrogen and oxygen in volumetric ratio of two to one. This latter is usually regarded as a chemical experiment rather than a physical experiment.

Let us look at some of the parameters of this development.

At the time, amount of charge was not measurable because the required physical methodology had not yet been invented. But by using a fixed (even tho unmeasurable) amount of charge, chemists discovered that water, upon electrolysis, decomposes into hydrogen and oxygen. However, the weight and volume of oxygen and hydrogen could easily be measured by the then available chemical methodology. This led Faraday to reverse the point-of-view, or what I call "inverting the paradigm" -- or what in more picturesque terms could be called "turning the sock inside out".

He simply used the newly discovered fact that water decomposes upon electrolysis, and our well-developed chemical ability to measure the weight and volume of gases to define a unit of charge and develop a method of measuring the charge. The development of the ability to measure charge then led to experiments which ultimately led to such discoveries as Ohm's Law, etc. I am simplifying drastically of course for the purpose of my argument. But let us generalize the essentials of this simplified development.

A controlled but unmeasurable unit or process (charge) in the parent field (physics) led to the discovery of a general phenomena (electrolysis) in a distinct subfield (chemistry). This general result (electrolysis) was in turn then inverted (measurement of weight and volume of gases) and used as the means of measuring the originally unmeasurable unit or process (charge).

This process has been repeated many times between physics and its various subfields. It is a process that is completely generalizable and also holds as well between semiotics and all of its subfields.

My next example is less-known and slightly more complex -- but it comes from semiotics and its subfield of psychiatry.

A control and a subject are seated in two separate rooms with no visual communication. There is one-way audio communication from the control to the subject by means of a microphone and loudspeaker. Pairs of words are displayed on a screen in both rooms, the same pair in both rooms. However, in the control presentation only, one word is underlined. The words are not distinguished in the presentation made to the subject. Example:

<u>RUN</u> WALK

The control now utters one word which may not be one of the two words, cannot rhyme with either word, and cannot give the position of the word on the display. For example, for this pair he might say "faster". The object is for the subject to infer as fast and accurately as possible which word is distinguished. Response time and accuracy are measured.

This paradigm can be used to measure the communication behavior of an open set of pairs of people by measuring their response times and error rates averaged over a fixed set of word pairs. By this means it has been found that pairs of persons who are good communicators by intuitive standards have a very low average response time and low error rates in this experiment while pairs who are poor by intuitive standards have a very high average response time and high error rates. For instance, it was discovered that pairs of schizophrenics typically have an exceptionally high average response time and error rate in this experiment.

When comparing the communication behavior of people by the above experiment, it is necessary to maintain the set of word pairs fixed because it is found that as the word pairs change, so will the response time. A pair of words which are intuitively nearly synonymous will take the same pair of persons longer to identify by this technique than a pair of words which are much less related. For example, most pairs of communicators take much less time and are more accurate in distinguishing NURSE from STONE than WALK from STROLL.

This suggests that the paradigm can be inverted and used to measure the communication synonomy of an open set of word pairs by selecting a fixed set of pairs of persons as standard and measuring the response time and error rates averaged over pairs of communicators. By this means semiotic phenomena which varies with synonymy may be investigated experimentally.

We move now to a more fruitful example. This is an area I have been actually working in for the past several years. It involves semiotics and psychology.

The Miller, Bruner, Postman (M-B-P) experiment measures the behavior of people interpreting artificial words. It was an attempt to measure and understand the Cattell Phenomena, discovered in 1885, that words are interpreted wholistically. In 1954 Miller used the uncertainty calculus, developed by Shannon in 1948, to model the Cattell phenomena in such a way that it could be measured, and hence investigated experimentally.

In a famous experiment, published by Miller, Bruner, and Postman in 1954, they discovered what is now called the Miller, Bruner, Postman Effect which says which says:

that in one sense the interpretation of words depends drastically on the shape of words, but in another sense the interpretation of words is completely independent of shape. In other words, the interpretation of words depends on their shape in a very particular way. However, in their paper, M-B-P stated that they could not measure word shape, but could only control it.

By this point in the paper, that statement should sound familiar and the reader should know exactly what I propose to do with it. Word shape is a semiotic phenomena and the M-B-P Effect can be inverted so as to serve as a measure of word shape. An instrument, called the 'eidometer', was invented in order to make this measurement more precise and with its help the Law of Word Interpretation describing the relationship between word shape and interpretations was discovered. This is a semiotic law in that it describes the relation between the shape of signs and the interpretation of these signs.

In the last few paragraphs we looked at an example of semiotic instrumentation and research under current active development. In this example we move to a proposal for future semiotic research and instrumentation development that is based on exactly the same reasoning that is exhibited in each of the previous examples. However, it is one that is based on sound theoretical considerations, past experimentation, and a desire to answer important semiotic questions, and has been seriously proposed for active concentration, so it has not been made up purely as an illustration of the present concept. Hence it serves as an excellent example of the fruitfulness of the Principle of Paradigm Inversion.

Bernbach discovered a phenomena involving differences in the ability to access short- and long-term memory depending on whether iconic or symbolic signs were used for this access. Again, he could not measure the degree of "iconicity" and "symbolness" of his signs; he could only control them. In fact, he used colors and color names in young children as his pure icons, and pure symbols. Our proposal in the SemLab was to use this "Bernbach Phenomena" in inverted fashion to measure the iconicity and symbolness of signs of mixed semantic structure.

In the Law of Word Interpretation, we had not gone directly from the phenomena to the law but had used the M-B-P Effect as an intermediary in developing our instrumentation. In the present case the intermediary is missing and we attempted to overcome this disadvantage by following up an unpublished discovery of the information scientist P.J. Siegmann who suggested that the Bernbach Phenomena could be intermediated by a visual interference effect that would be affected only by the icons, and not by the symbols.

In a preliminary set of experiments the concepts of iconic squares and iconic circles were developed into a crude system of instrumentation and the existence of Siegmann's visual interference phenomena was confirmed. In addition, it appears that a simple metric may be imposed on iconic similarity, thus enabling a mathematical relation between visual interference and iconic similarity to be determined. Our proposal is thus to invert the Bernbach paradigm and to use the relation between visual interference and iconic similarity to measure the iconicity of signs of mixed semantic structure.

Once again, iconic circles are semiotic instruments, and iconicity, is a semiotic measurement, and they have been obtained by inverting a paradigm from developmental psychology.

This series of examples leads to the general conclusion that every single experimental paradigm in each of the cognitive sciences can be inverted to supply us with an experimental paradigm which may be developed for semiotics.

The Principle of Paradigm Inversion also enables us to distinguish between semiotics and each of the cognitive sciences. For instance: Psychology uses the known structure of signs as a probe to investigate the structure of behavior, whereas semiotics uses the known structure of behavior as a probe to investigate the structure of signs.

According to our idealization made at the beginning of this section, semiotics now has a very full bag of experimental paradigms and the job is therefore to determine what questions to ask, design the experiments and carry them out. Measurement, quantification, and formalization may then follow in their proper course.

6. The Concept of Paradigm Inversion:

By relying on the same set of experimental paradigms, we appear to lose much of the distinction that has traditionally been assumed between semiotics and the semiotic sciences. In this section I will explicate the concept of paradigm inversion which is used in the Paradigm Inversion Principle and use it to argue that the differences between semiotics and the semiotic sciences still remain but they are more subtle than previously supposed. My explication involves a careful attempt to distinguish between a psychological experiment and an experiment in semiotics and/or information science. Psychology is used as a particular example only and the argument is completely generalizable to all the semiotic sciences.

Semiotics and psychology start with the same open structure which consists of a basic experimental paradigm, E, a set of messages, M, and a set of subjects, S; see fig. 1. The message signs from M are used as the stimuluses by the experimental paradigm E, and the behavior of the subjects in S in response to the stimuluses in the context of E is the primary observable of the experiment. Associated with either semiotics or psychology is a language, a set of theories, a set of laws, and a set of questions which motivate the design of the specific experiment.

The difference between psychology and semiotics lies in the way they complete this open structure. Psychology completes the design of the experiment by closing, or fixing, the set of messages into what can then be taken as a standard stimulus set. This plays the same metrological role as selecting a fixed meter bar to serve as a standard for the measurement of all lengths. The set of subjects is then left open, and the experimental paradigm and the fixed, standard, set of stimuluses becomes a single experiment which may be run on any subject to determine the experimental response for that individual subject. As a result, regularities which are determined to hold for any subject may then be stated as a law of psychology. The statement of this law contains a universal quantifier over subjects.

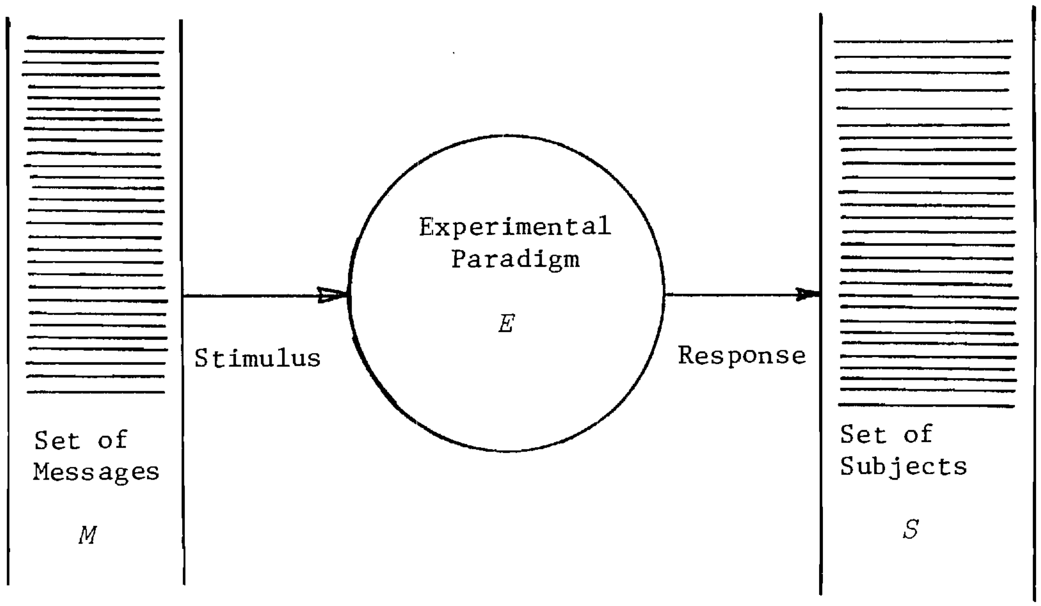


Fig. 1: The Structure of an Experiment in the Semiotic Sciences.

Semiotics, on the other hand, completes the design of the experiment by closing, or fixing, the set of subjects into a standard subject set and the set of signs, or messages, is left open. The experimental paradigm and the fixed, standard, set of subjects becomes a single experiment which may be run on any sign, in the set of messages for which the experiment was designed, to determine the experimental structure for that individual sign. Regularities which are then determined to hold for any sign in the message set may be stated as a law of semiotics, or information science, and the statement of this law contains a universal quantifier over signs, or information carriers.

This view and analysis makes obvious the fact that a semiotics experiment cannot be designed without some knowledge of psychology (or some other experimental semiotic science) nor can a psychology experiment be designed without some knowledge of semiotics. The two are thus seen to be interdependent. To date most of the emphasis has been placed on the psychological point of view, using a naive knowledge of semiotics. It would now be of benefit even to psychology to place more emphasis on the semiotic point of view. The payback to psychology would be in increased sophistication in the design of psychological experiments due to the increased knowledge of stimulus structure. Of course an emphasis on semiotic experimentation is absolutely necessary for progress in semiotics, or information science, or even for its existence as a science.

Semiotics interfaces with all the semiotic sciences and the example with psychology was used only as an illustration, psychology being the most sophisticated, empirically, of the semiotic sciences. The same example could be worked out for each of the semiotic sciences: sociology, economics, esthetics, etc. In fact, semiotics is the common ground between all the semiotic sciences, and it is only by taking the semiotic point of view for each of them that their unity can be understood, and their methodology and knowledge become inter-adaptable. We thus have:

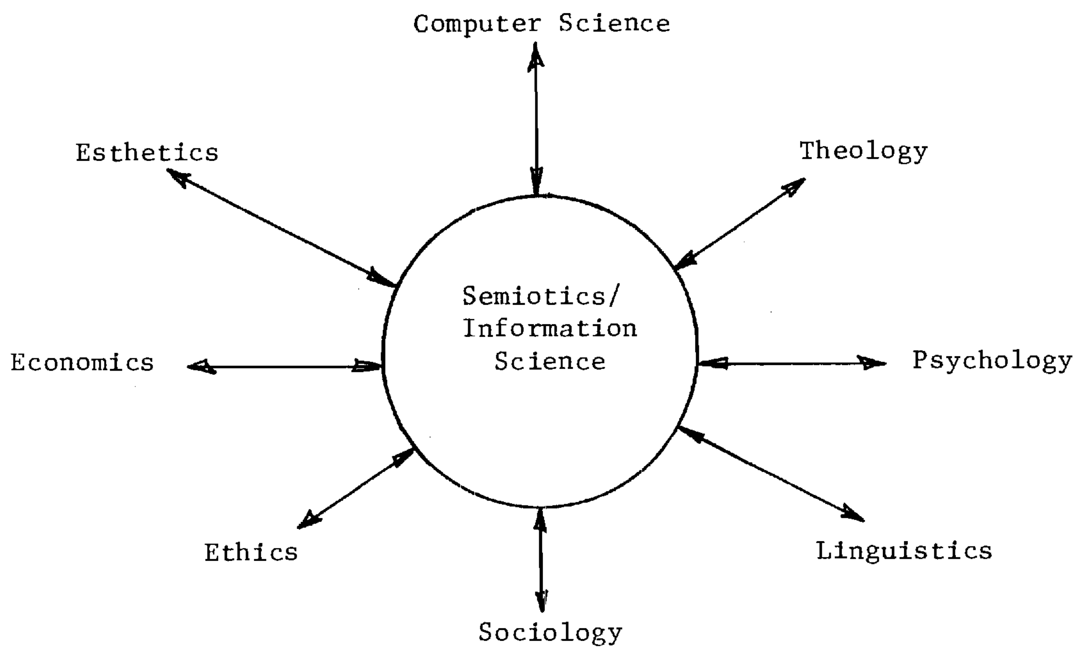


Fig. 2: The Interdependence of the Semiotic Sciences.

We may now specialize fig. 1 to each of the semiotic sciences individually while completing it. I do this below for semiotics in fig. 3 and psychology in fig. 5.

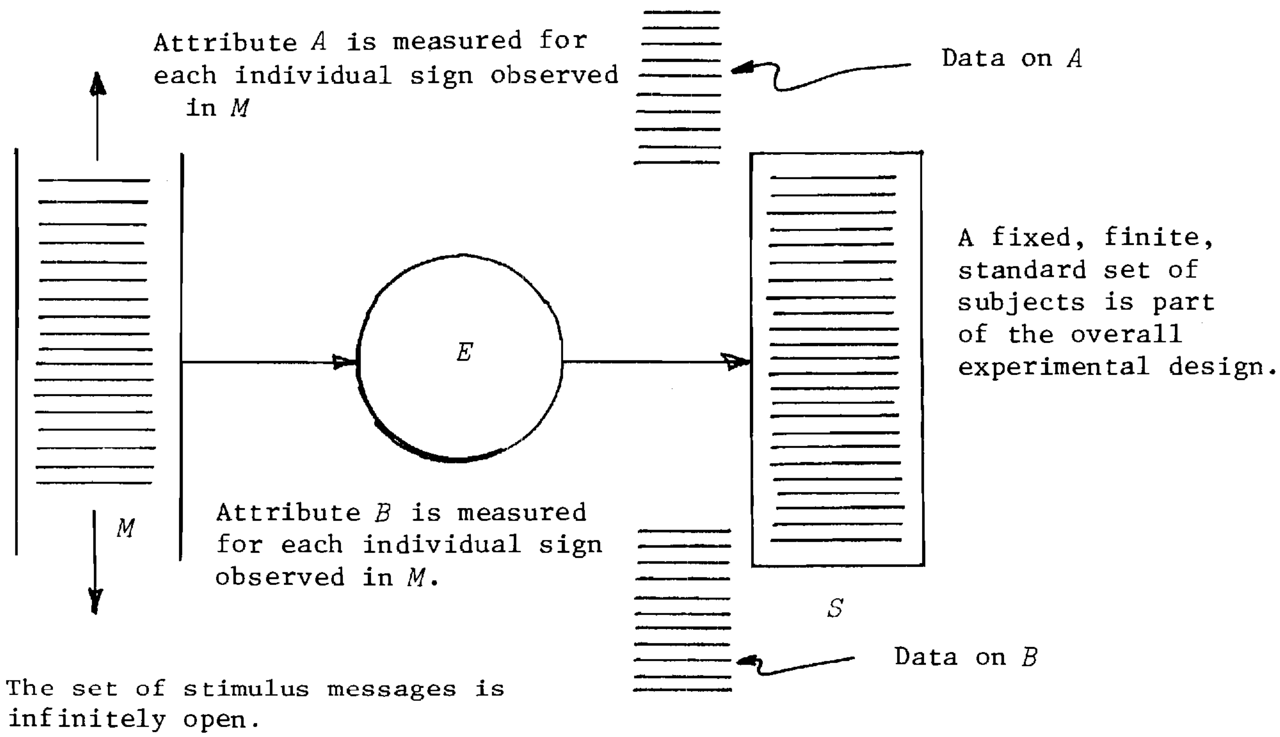


Fig. 3: A Semiotic Experiment.

The measurements on *A* and *B* for the same individual sign are identified as the two components of a vector and a search is made for regularities (this is grossly simplified). If a regularity results, we can set up *A* and *B* as two coordinates.

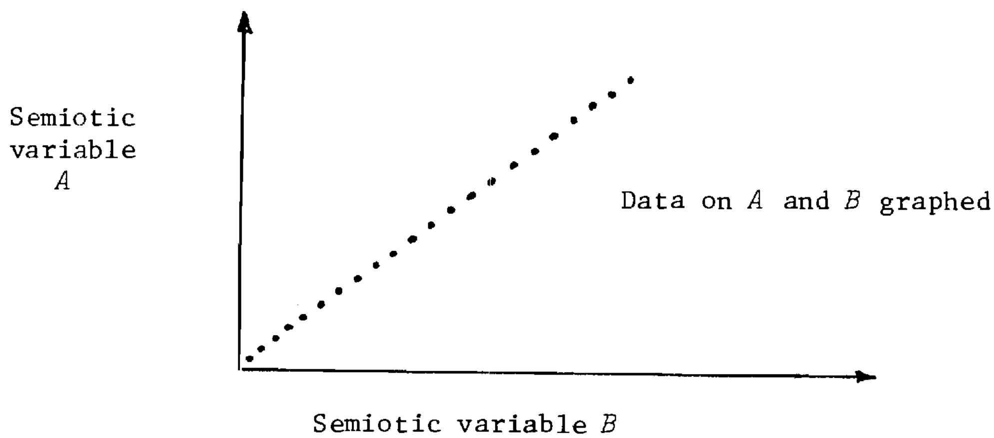


Fig. 4: A Law of Semiotics.

Since both *A* and *B* are semiotic variables and the individuals are signs, and the subjects are standaridzed, this is a law of semiotics.

Measurements are made by observing the behavior of the standard set of subjects, perhaps by averaging some response over all the subjects in the standard set.

We next do the same for psychology.

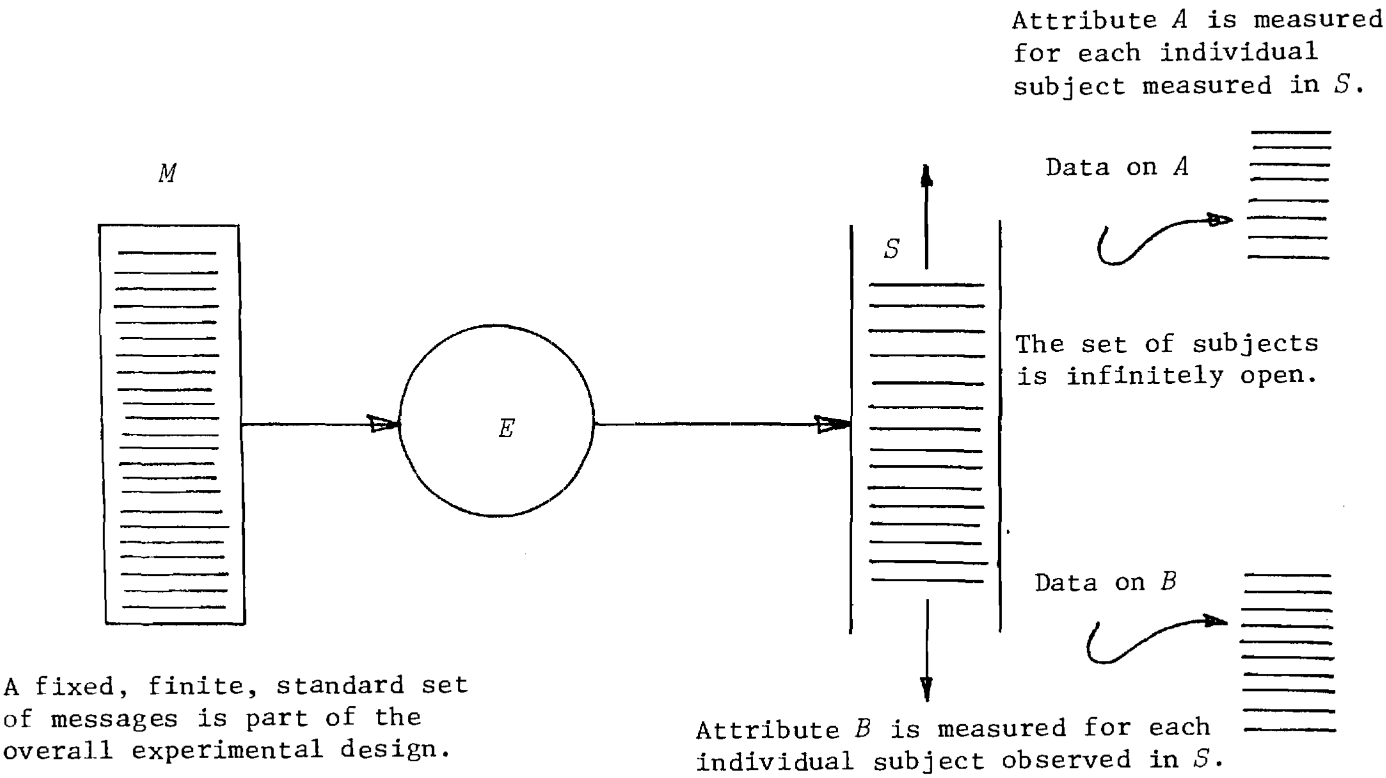


Fig. 5: A Psychology Experiment.

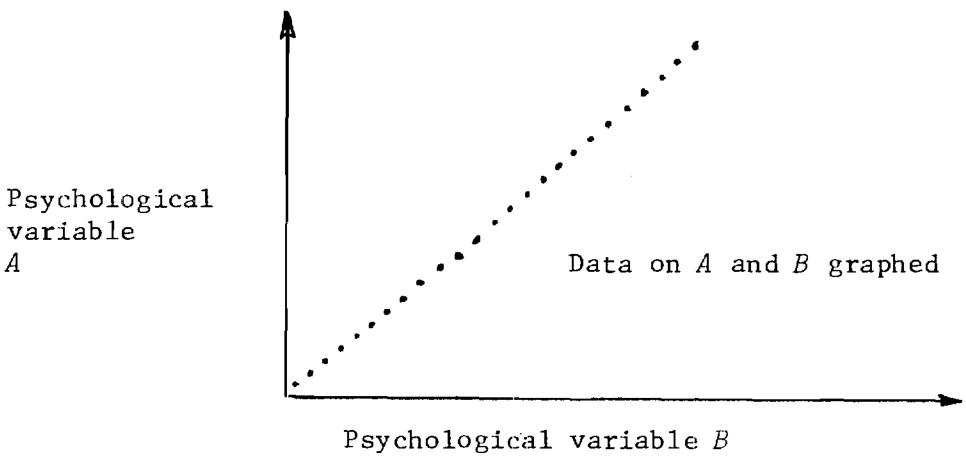


Fig. 6: A Law of Psychology.

Psychological measurements are made by observing the subject's behavioral response over the standard set of signs, perhaps by averaging some responses over all signs in the standard set.

Since both *A* and *B* are psychological variables and the individuals are subjects and the messages, or stimuluses, are standardized, this is a law of psychology.

The statement that humans are too subjective, or too variable, to use as measurement standards just does not hold up. The use of a fixed set of subjects 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 temperature involving ideal gasses. It will also take many experiments before we arrive at the proper idealizations to replace concrete subjects in semiotics but in the meantime experimentation and measurement are both possible and necessary.

7. The Nature of Semiotic Measurement:

In the last section we saw that the same experimental paradigms generate experiments in both semiotics and the semiotic sciences depending on the type of measurement involved. This raises a question as to the nature of semiotic measurement. In this section the nature of semiotic measurement is explicated using the Principle of Semiotic Reinterpretation; at the same time we are able to display the empirical foundation and explain the usefulness or lack of usefulness of individual semiotic measures, or information measures.

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.

Let *S* be the scale of a semiotic measure (the information measure to be reinterpreted) and suppose that *S* has some intuitive relationship with some other domain (semiotic, or nonsemiotic). Let us take as an example that there is some intuitive relationship between *S* and the psychological domain. Then we let *A* be the scale signifying this domain -- the psychological domain in this example.

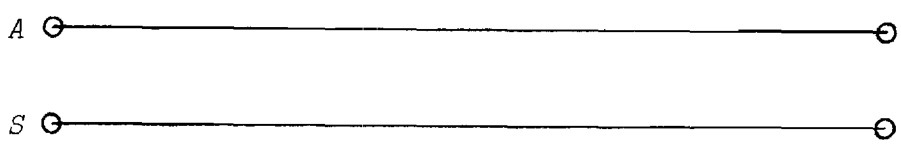


Fig. 7: Intuitive Relation Between Two Domains.

Now the relationship between *A* and *S* can take three forms. Our procedure depends on which form of the relationship we have. If the intuitive relationship between *A* and *S* involves an established measure in *A* we can think of *A* as the scale for this measure. We may then rotate the *A* scale 90 degrees to form a rectangular coordinate system. We then investigate a set of objects which allow independent measurements on both the *S* and the *A* scales, performing both measurements on each object.

If the intuition that told us there was a relation between the *S* and the *A* scale was correct, upon completion of this investigation we will have discovered a natural law -- an empirical regularity holding between the two domains. For instance, figure 8. (Again, we are simplifying drastically in order to clarify the argument.)

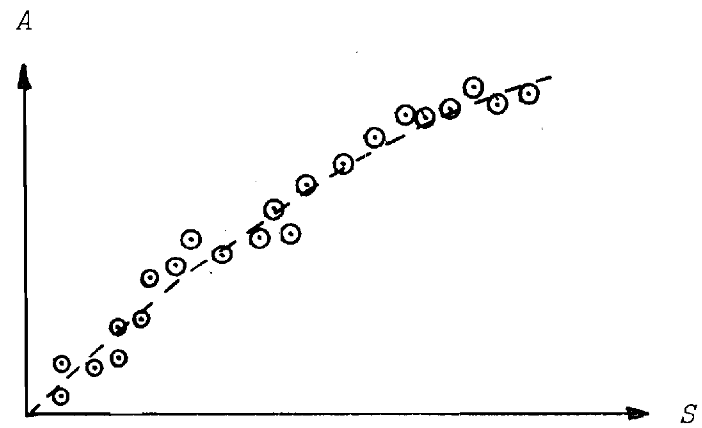


Fig 8: A Clear-cut Empirical Regularity.

Often, however, the results will not be clear-cut as in Figure 9 and one or more of the two scales must be refined or modified somewhat to bring out the relationship in a clearcut fashion. Many tools are available for determining such relationships or regularities including such statistical methods as significance testing, correlation analysis, and regression analysis.

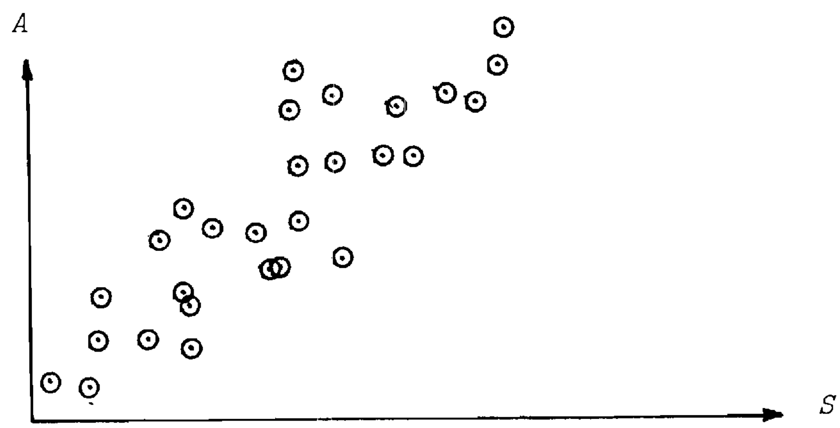


Fig. 9: A Less Clear-cut Empirical Regularity.

Which scale should be modified, or at least investigated first? A good working hypothesis is to examine first the scale which is least established. That is the scale which enters into the fewest other known natural laws. How should it be modified? There is no answer to this question. This is one place where the creativity and ingenuity of a good scientist shows itself. One explicates his concepts and measures in a way that is

1. consistent,
2. does no violence to all other known results, and
3 clarifies the regularities in the data.

An example of this approach to semiotic reinterpretation is the Hick-Hyman Law, where the relationship between the comentropy of an alphabet (a semiotic or information measure) and the time required to psychologically react to the signs in that alphabet (a psychological measure) was discovered to be given by

$$RT = aH + b$$

no matter how the comentropy of the alphabet was varied.

The relationship between *S* and *A* takes the second form if the intuitive relationship between *S* and *A* involves no established measure in *A*. In this case we can use the semiotic measure and the intuitive relationship to establish a measurement by <u>FIAT</u> in *A*. By playing around with our two scales we can examine whether the intuitive relationship demands a linear scale for *A*, an inverse scale for *A*, etc. Figure 10 illustrates this idea.

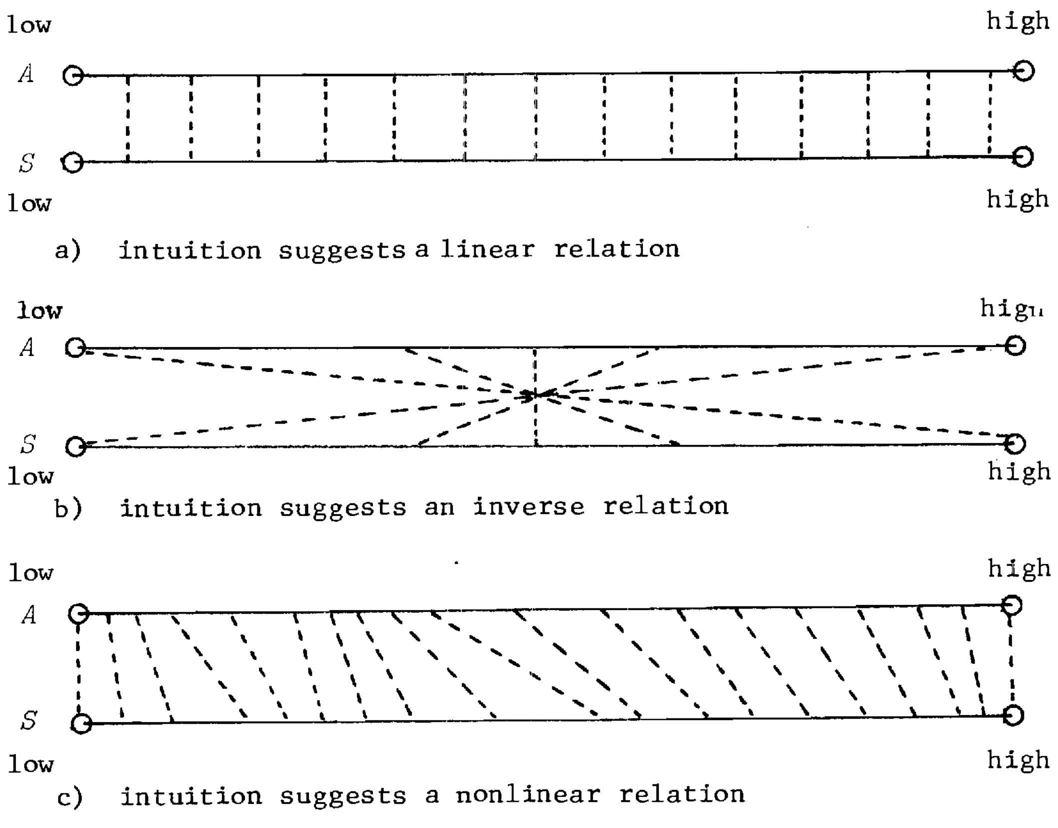


Fig. 10: Examining Our Intuition When No Established Scale For the A Domain is Involved.

The lines between the *A* scales and the *S* scales in the above diagram represent the intuitive feelings of the investigator. A good working procedure in this case is to think of as many different quality names in the *A* domain that vary with any relationship to *S* at all. Often it can be seen that these qualities cluster around some unnamed quality and tend to delimit its possible relations to *S*.

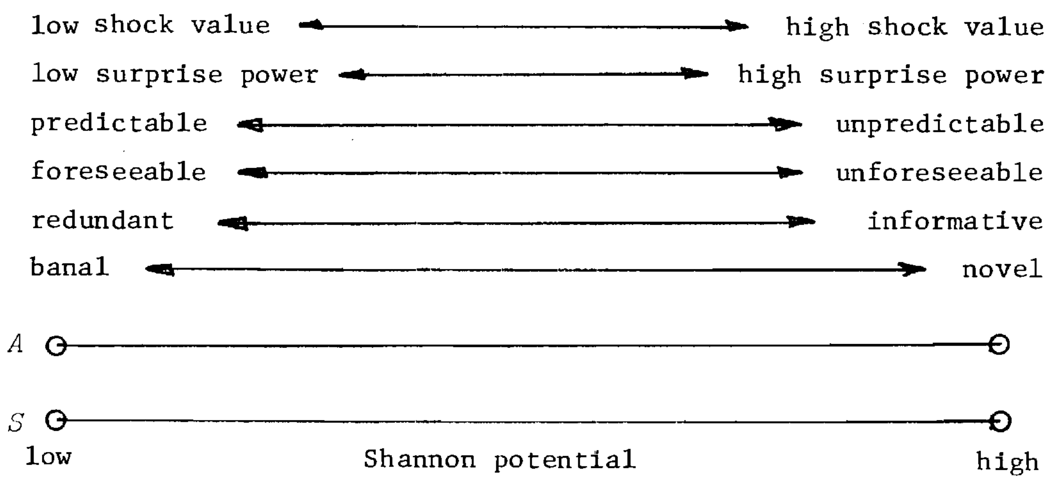
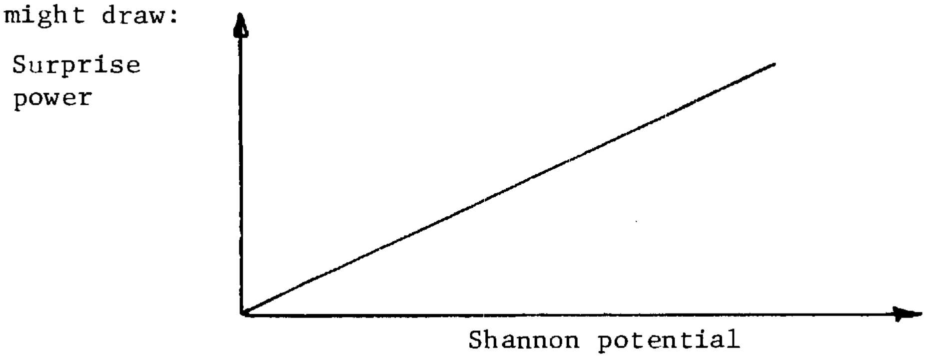
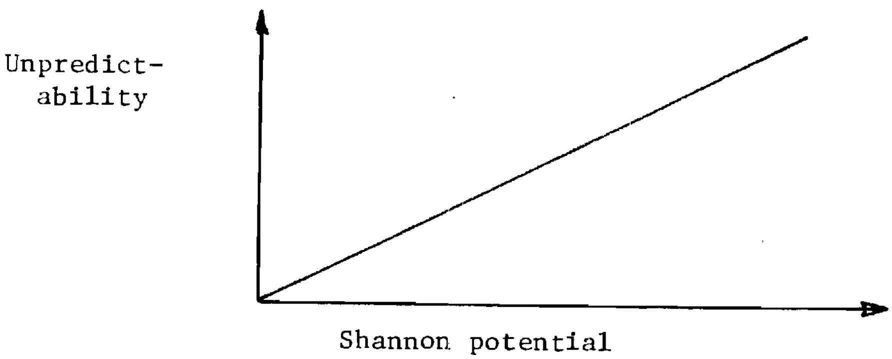


Fig. 11. Attribute Analysis: A Linguistic Exercize.

Now no matter what terms are used to describe the psychological factor, we recognize an intuitive empirical correlation between the two scales. Thus: we might draw:



a) using surprise power as an approximation to the unknown quality



b) using unpredictability as an approximation to the unknown quality

Fig. 12: Intuitive Relation Between Two Domains Illustrated By Two Different Attributes or Qualities.

Remember, at this time these are just intuitive relationships and the lines in the above two examples are not experimental results but are drawn arbitrarily to satisfy intuitions. One does the best he can at this point. Scales are only linguistic conventions and one can always come back in the light of later data and improve the conventions.

Again as in the alternative case, the procedures are not clear-cut. Ingenuity, creativity, and scientific insight are needed to come up with a good definition. But when a good definition for a measurement by fiat system has been achieved it can be used to develop an independent measurement system in the *A* domain. Later investigation with this new measure can then be used to check or revise and modify the arbitrary measurement by fiat definition.

An example of this approach to semiotic reinterpretation is Shannon's original definition of information of a sign occurrence as being equal by definition to the Shannon potential of the sign. Information is a psychological concept and this thus produced a psychological measurement by fiat. But this definition led to the later discovery of the relationship between information and uncertainty. A way was found to measure uncertainty independently of the Shannon potential and it was discovered that information as measured by fiat was equal to the uncertainty removed by the sign as measured independently. This is known as the Shannon-Fano Law. The Shannon-Fano Law can be seen to say that when uncertainty is measured by an independent psychological measurement and the Shannon potential is measured semiotically the two measurements on the same sign always have the same numerical value. This then is an independent check on the original arbitrary definition of information measurement by fiat and the definition is seen to not need revision (altho many workers would like to see the name of the measurement revised because it is philosophically confusing).

The third form of semiotic reinterpretation has been employed in the situation where no semiotic measure existed but a natural law was known to relate a measure in one domain to the semiotic domain. This was the manner in which Hartley and Gabór defined their measures of channel capacity, independently of each other, starting with the physical measure of action, which is the product of distance times momentum, and the Heisenberg uncertainty principle which is one of the fundamental laws of quantum mechanics. They proceeded to show that Heisenberg's principle implies that the smallest possible quanta of action determines a maximum limitation to the number of signs that any physical device can embody and they then defined the channel capacity as a measure of this maximum limit.

Note: Just as we were free to either define measures for occurrence frequency and uncertainty and then observe the empirical regularity between the two; or to define the occurrence frequency as a measure of the surprise power by fiat and then observe the linearity between this and uncertainty as measured by counting predictions, Einstein has said that any natural law can be REINTERPRETED as a definition and vice versa (except for the purpose of analyzing the empirical foundations of a discipline).

The Thesis of Semiotic Reinterpretation

1. All 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.

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.

Note: It is then this natural law which gives the empirical basis for the usefulness of such measures.

8. Summary:

Elsewhere it has been indicated that it is necessary to base the quantification of semiotics on an empirical approach rather than the other way around. This paper has examined some of the methodological issues involved in doing this and has countered the obvious objections that semiotics has no experimental paradigms, that semiotic measurement is impossible without previous quantification, and that it is impossible to measure something as ephemeral as a sign, its meaning, and its structure.

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 semiotics, and use this to design semiotic experiments and make semiotic 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. Finally, we do not lose the distinction between semiotics and the semiotic sciences when we do this. In fact, we can use the relationship of Paradigm Inversion to determine how to measure the ephemeral semiotic quantities in terms of concrete properties of behavior evidenced in some one of the semiotic sciences such as psychology.

It should be noted that the Principle of Paradigm Inversion and the general arguments of section 6 distinguishing between psychological experiments and semiotic experiments are all corrolaries of the Principle of Semiotic Reinterpretation. This is seen by noting that within the principle, all arguments are symmetric about the two axes involved. The argument can in principle be turned around to go from the psychological domain to the semiotic domain, and this is just what occurs in Paradigm Inversion. The Principle of Semiotic Reinterpretation is thus seen to be a very powerful principle of semiotic methodology, highly useful in experimental design, the development of measurement scales, and the interpretation of experimental results, as well as in philosophical analysis.

As we saw in section 6, the Principle of Paradigm Inversion indicates the existence of a relationship between semiotics and each of the semiotic sciences. The two principles discussed here may be used to examine the details of this relationship but this investigation did not fall within the scope of this workshop. It was explored in more depth in the Experimental Semiotics

Workshop held at Georgia Tech in conjunction with the 1976 Conference of the Semiotic Society of America and in a paper titled, "The Cognitive Sciences: A Semiotic Paradigm", delivered to the 1978 Conference of the Society for the Interdisciplinary Study of the Mind.

INFORMATION DECAY IN IMMEDIATE MEMORY:
A SECOND ORDER CORRECTION TO THE LAW OF WORD INTERPRETATION

by Charls Pearson

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

Presented at the session on Experimental Semiotics held at the Second International Semiotics Congress in Vienna, Austria; July 2-6, 1979; and submitted for inclusion in the published proceedings of the congress.

© by author

December, 1979

INFORMATION DECAY IN IMMEDIATE MEMORY:
A SECOND ORDER CORRECTION TO THE LAW OF WORD INTERPRETATION

by Charls Pearson

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

ABSTRACT

Empirical explication of gross semiotic phenomenas is more than just "quantification of the obvious". It has deeper significance for the intellectual understanding of the more subtle semiotic phenomenas.

This paper exemplifies the above argument by showing how one "quantification of the obvious", the Law of Word Interpretation, has been used to observe and measure the second order quantities associated with word interpretation and thereby open up for study the subtle, and even unsuspected, semiotic phenomenas associated with immediate memory.

The study is a statistical analysis of the data from a series of word interpretation experiments run several years ago. The original word interpretation experiments were run in order to answer some questions raised by the Universal Sign Structure Theory, but the purposes of the present analysis were mainly methodological. The results have some theoretical implications involving the relation between the frequencies of the individual letters making up the shape of the word and the information decay in immediate memory associated with the whole word. The methodological argument of second order quantities has implications for the empirical paradigm of philosophical semiotics.

The theory that second order effects in the Law of Word Interpretation are due to information decay in immediate memory partially explains these residuals; but an experimental re-explication of Miller-Bruner-Postman's concept of 'placement errors' is also found to be required. Information decay in immediate memory is found to depend on the frequencies of the individual letters and their positions in the specimen words in a way that is not yet fully clear.

INFORMATION DECAY IN IMMEDIATE MEMORY: A SECOND ORDER CORRECTION TO THE LAW OF WORD INTERPRETATION

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

Introduction

It seems to be difficult for nonempirical semioticians (literary critics, speculative philosophers, etc.) to understand what empirical semiotics is all about. One complaint often heard is that they (the empirical semioticists) merely quantify the obvious; while the semioticists reply that by quantifying the obvious they are replacing a vague understanding of semiotic nature with a precise scientific understanding -- an argument that cuts little weight with the semioticians.

The two sides of this question constitute an age-old controversy between scientists and nonscientists in all fields of knowledge, a controversy that has arisen at least once in the history of every scientific discipline. It is no wonder then that it is arising now in semiotics, as this discipline struggles to make a science of itself.

This paper does not argue the merits of the two sides of this controversy. It is unashamedly biased, assuming as it does, the correctness, and all the intellectual advantages, of the scientific approach to knowledge. Instead, this paper exemplifies, within empirical semiotics, a new argument to use in convincing nonscientists of the propriety of the empirical approach.

By quantifying the obvious, empirical semioticists enable the measurement, and therefore the observation, of the inevitable and inherent second order quantities, or small differences between the quantified obvious and the actual quantities of semiotic nature, and thereby open the nonobvious up to investigation and intellectual understanding.

I will not discuss the methodological implications of second order quantities at this point, nor analyze the philosophical arguments this leads to in the scientific/humanistic and empirical/rationalistic battles. I intend to attempt this in another place. This paper will only exemplify this argument by presenting the analytical details of one investigation in eidometry.

The present study is a statistical analysis of the data from a series of word interpretation experiments run several years ago. It relies primarily on linear regression, multiple regression, and partial correlation techniques; but also makes use of some concepts from canonical correlation and non-linear regression. The original word interpretation experiments were run in order to answer some questions raised by my Universal Sign Structure Theory, but the purposes of the present analysis were mainly methodological. The results have some theoretical implications involving the relation between the frequencies of the individual letters making up the shape of the word and the information decay in immediate memory associated with the whole word. The methodological argument of second order quantities has implications for the empirical paradigm of philosophical semiotics.

Background:

In 1948 [4] Shannon noticed a phenomena discovered earlier by Markov, that information sources generating words by a Markov-chain of finite order produce words that look more like natural language, the higher the Markov-order. In 1954 [1] Miller, Bruner, and Postman discovered another phenomena involving the interpretation of words generated by a Markov information source. As expected, the errors of interpretation under tachistoscopic conditions decreased, the higher the Markov-order; but the surprising result was that the amount of information (in Shannon's sense) obtained from a tachistoscopic exposure is independent of the order of approximation to American letter sequences. Therefore differences in the error rate can be predicted from a knowledge of the information structure of American.

Intuitively it would seem that these two phenomenas are related. However, without a quantitative concept of wordshape and an instrument to measure it, it was not possible to study this relation experimentally. An instrument, called the 'eidometer', was therefore invented based on Shannon's observations; and a concept of strangeness of word shape relative to a given natural language, called 'eidontic deviance', was developed based on a series of experiments with the eidometer, [2].

The relation between eidontic deviance and interpretability has been studied with experiments involving an eidometer and a teescope, and in 1978 Pearson announced his Law of Word Interpretation [3],

$$E = a + bS$$

where E is the tachistoscopic error rate measured in letters per word (lpw); S is the strangeness measured in eidontic deviance units (°ED); and a and b are constants. This law is a quantitative description of the Miller, Bruner, Postman Effect and states that the above relation appears to be fairly linear thruout a wide range of word shapes.

Miller, Bruner, and Postman originally attributed the interpretation error effect to perceptual semiosis [1]; however, Sperling has more recently interpreted it as a semiotic process associated with immediate memory, [5; and 6]. The eidometer, however, still appears to be measuring our perception of the strangeness of sign shape relative to a given natural language, [2], and so this law would appear to relate a perceptual semiotic process to a semiotic process associated with immediate memory.

A linear correlation between the eidometer measurements and the teescope measurements in the Word Interpretation Experiment give a correlation coefficient of

$$r = 0.66$$

with r^2 about 44% for a validity of about 0.82 bit/measurement. For 97 degrees of freedom this is significant at the $\alpha = .00001$ level. There is much scatter left in the data, however, and with an evaluated precision of 4 bits/measurement and an evaluated reliability of 3 bits/measurement, [2], the scatter is far too much to be explained by instrument and measurement effects. It apparently represents real semiotic differences in the specimen words. Accordingly, the present study was designed to investigate these effects.

Purpose of Present Analysis

According to the interpretation of Miller, Bruner, and Postman's 'placement error' concept given by Sperling these real semiotic differences should be due to information decay in immediate memory. It is therefore of interest to learn how to design a set of experiments that would test this interpretation. This requires a search for regularities in the errors involved in measuring placement errors. Current psychological theories of immediate memory suggest these residuals should be a function of the individual letter frequencies and their position in each word.

In analyzing these regularities it should also be possible to determine, at least in part, whether the present data requires a re-explication of Miller, Bruner, and Postman's concept of 'placement-error'. If a simple mathematical function of the letter frequencies and their positions completely describes the measurement residuals, this would suggest both the correctness of the immediate memory information decay interpretation and the satisfactoriness of the 'placement error' concept. If no mathematical relationship at all can be found between the letter frequencies and the measurement residuals, this would suggest some other interpretation of the 'placement-error' effect and would therefore also have nothing to say about the satisfactoriness of the 'placement-error' concept. Finally, if a significant but partial description can be found and if the remaining residuals are also still significant, this, while not proving anything, would suggest both the correctness of the immediate memory information decay interpretation and also the need to re-explicate the 'placement-error' concept. It was suspected that this latter would prove to be the case, especially since several major refinements of the 'placement error' concept were already obvious from our initial experiments.

If the re-explication of Miller, Bruner and Postman's measurement concept should prove necessary, it was hoped to gain enough information from the present analysis with which to design such experiments and carry out this re-explication. This in turn would lead to an improvement in the experimental methodology of eidometry by leading to better control over random effects in the variables not under investigation and better control over the variation, and more precise measurement, of the parameters under study.

Finally, each one of the above goals of this analysis leads to a further, major, goal of the study: improving the precision and validity of the Law of Word Interpretation. There were thus six goals of this study and they are summarized in table 1.

Table 1. Goals of the Present Study

1.	To determine how to design a set of experiments that would test the interpretation of the Law of Word Interpretation residuals as being due to information decay in immediate memory.
2.	To search for a regularity in the errors involved in measuring placement errors. Current psychological theory suggests these residuals should be a function of the individual letter frequencies and their positions in the word.
3.	To determine if the present data requires a re-explication of the Miller-Bruner-Postman concept of 'placement-error'.
4.	To determine how to design experiments and carry out this re-explication, if necessary.
5.	To determine how to improve the experimental methodology in eidometry by maintaining better control over random effects in the variables not under investigation and better control over the variation and more precise measurement of the parameters under study.
6.	To improve the precision and validity of the Law of Word Interpretation.

Results:

1. Our best measurement to date of the Law of Word Interpretation is given by

$$E_c = a + bS$$

where E_c is corrected placement errors measured in letters per word (lpw), S is strangeness of word shape measured in eidontic deviance units (°ED) and a and b are constants given by

$$a = 4.61 \pm .08 \text{ lpw}$$

$$b = -0.87 \pm .08 \text{ lpw/°ED}$$

The tolerances quoted in these measurements are one standard error of measurement.

2. The correlation coefficient for the corrected Law of Word Interpretation is

$$r = -0.777$$

with r^2 in excess of 61% for a validity in excess of 1.33 bit/measurement (bpm).

3. The standard error of estimate in the Law of Word Interpretation is now ±0.59 lpw.

4. The residuals do not seem to have anything to do with eidontic deviance, placement errors, or the Markov generator of the specimen words.

5. We now know how to design a series of experiments to test the immediate memory interpretation of the CORTRM effect.

6. Immediate memory seems to affect the measurement of placement errors in a manner apparently described by an absolute value function of the frequencies of the individual letters in the word and their positions, which is approximated by our CORTRM function.

7. However, CORTRM does not account for all of the scatter in the data. The variance in the corrected residuals after correcting the Law of Word Interpretation for CORTRM is still semiotically real.

8. This suggests that the concept of 'placement error' needs to be re-explicated. There are two glaring deficiencies in the Miller-Bruner - Postman explication of this concept. It does not take into account either of the following:

a) misplaced strings: ACDEFGHX for ABCDEFGH

b) confusion errors: (visual): E for F, H for N, C for G, O for Q, etc.
(verbal): C for Z, E for T, etc.

9. A series of experiments can now be designed to explicate the perturbations in the measurement of the placement errors due to the misplaced string phenomena.

10. A series of experiments can now be designed to explicate the perturbation in the measurement of placement errors due to the problem of confusion errors.

11. We now know how to design a series of experiments to determine a better approximation to the CORTRM function -- i.e., we know to control for the maximum and minimum letter frequency in each letter position as well as the maximum and minimum difference between the maximum and minimum frequency thruout all eight letter positions.

12. All future word-shape experiments must incorporate such controls.

Discussion

The basic Law of Word Interpretation was announced and discussed in [3] and is shown in fig. 1, where the regression line has been drawn in by hand. Its equation is given by

$$E = a + bS$$

where E is placement errors measured in letters per word (lpw), S is strangeness of word shape measured in eidontic deviance units (°ED) and a and b are constants. Under the conditions of the experiment described in [3], the best measurements of a and b were

$$a = 5.51 \pm .19 \text{ lpw}$$

$$b = -0.98 \pm .11 \text{ lpw/°ED}$$

where the tolerance figures represent one standard error of measurement. The scatter is indicated by the standard error of estimate, which was 0.71 lpw.

The above values were determined by a regression analysis of the raw experimental data (fig. 1). In such cases once the analysis is complete, it is necessary to test the assumptions of regression analysis to determine if the analysis is a valid one. This test cannot be made until after the analysis has been completed since it involves the residuals themselves. The basic assumptions to be tested are that the residuals are normally distributed about zero, that they have no systematic relation to the independent variable, and that the variances are constant with respect to the independent variable. Pearson performed these tests by means of a scattergram analysis [3]. Fig. 2 shows the results of one such analysis for the Law of Word Interpretation. The results of all such analyses were negative indicating the basic validity of the regression analysis used in isolating the law and also suggesting that the residuals are not related to eidontic deviance, placement errors, or the Markov generator of the specimen words.

The negative result of the regression residuals analysis does not indicate that the residuals are completely without semiotic meaning, however; but merely that the residuals have no semiotic meaning that would invalidate the validity of the regression analysis. To determine whether the residuals represent any real semiotic differences in the specimen words, they were compared to the precision of the eidometer and teescope evaluated for the procedures used in the Word Interpretation Experiment. Fig. 3 illustrates a typical result of such an analysis. The stars represent actual measured values and the diamonds represent the evaluated instrument precision for several typical points in terms of standard error of measurement limits. For each point analyzed there was less than one chance in ten thousand that that residual was really zero; thus indicating that the residuals represent real semiotic differences in the specimen words. This is one example of the necessity and also the usefulness of evaluating the instrument performance parameters (precision, repeatability, etc.) for all instruments and all procedures used in semiotic experiments.

FILE SDATA1 (CREATION DATE = 79/04/26.)
SCATTERGRAM OF (DOWN) TSCOPE (ACROSS) ED

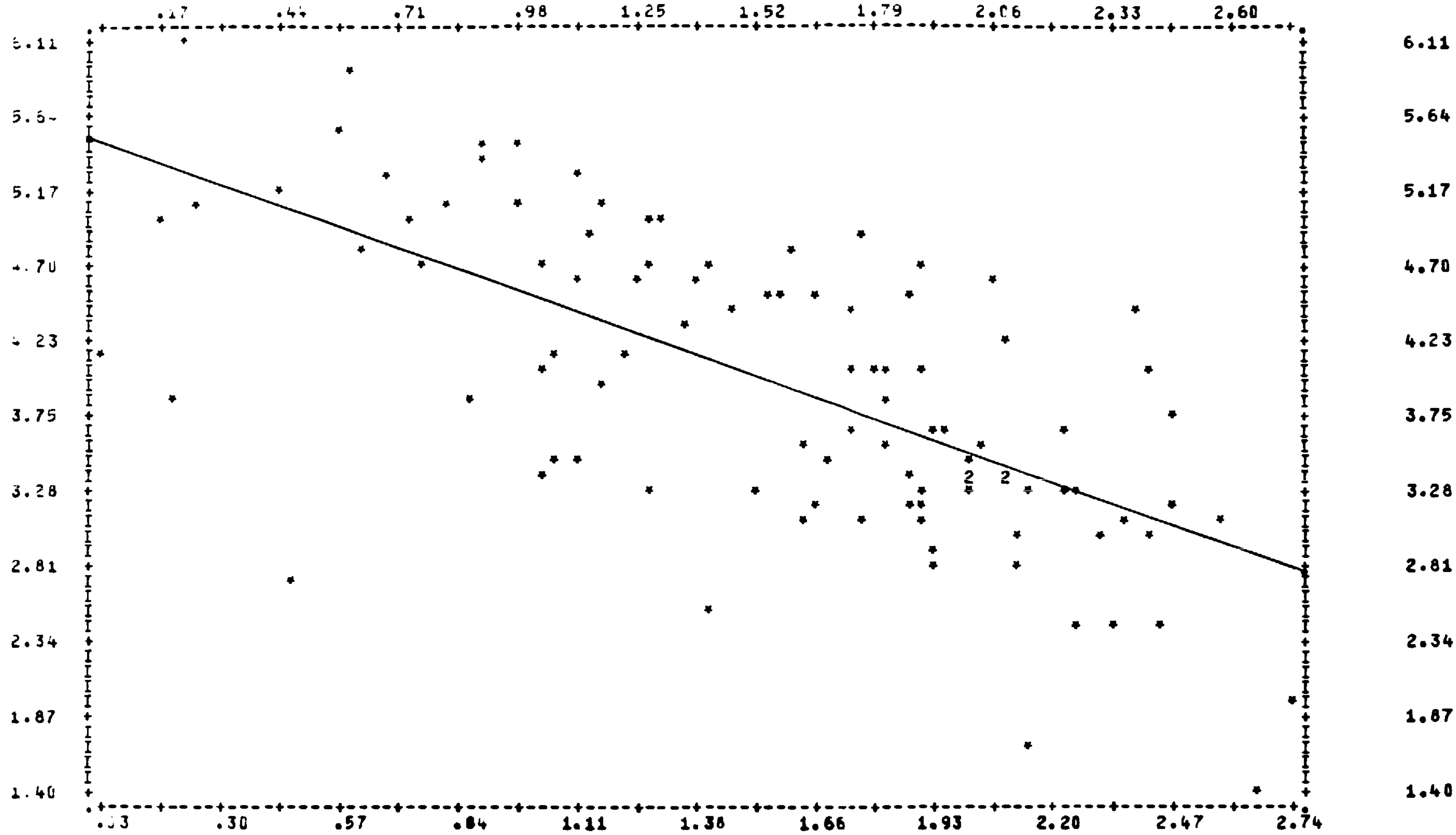


Figure 1. The Basic Law of Word Interpretation

Raw Teescope Readings Vs. Raw Eidometer Readings

TEESCOPE READINGS VS. EIDOMETER READINGS 79/06/19. 14.07.54. PAGE 52

STATISTICS..

CORRELATION (R)-	-.66055	R SQUARED -	.43633	SIGNIFICANCE R -	.00001
STD ERR OF EST -	.70882	INTERCEPT (A) -	5.50892	STD ERROR OF A -	.19016
SIGNIFICANCE A -	.00001	SLOPE (B) -	-.98012	STD ERROR OF B -	.11253
SIGNIFICANCE B -	.00001				

SCATTERGRAM OF (DOWN) RESID (ACROSS) FD

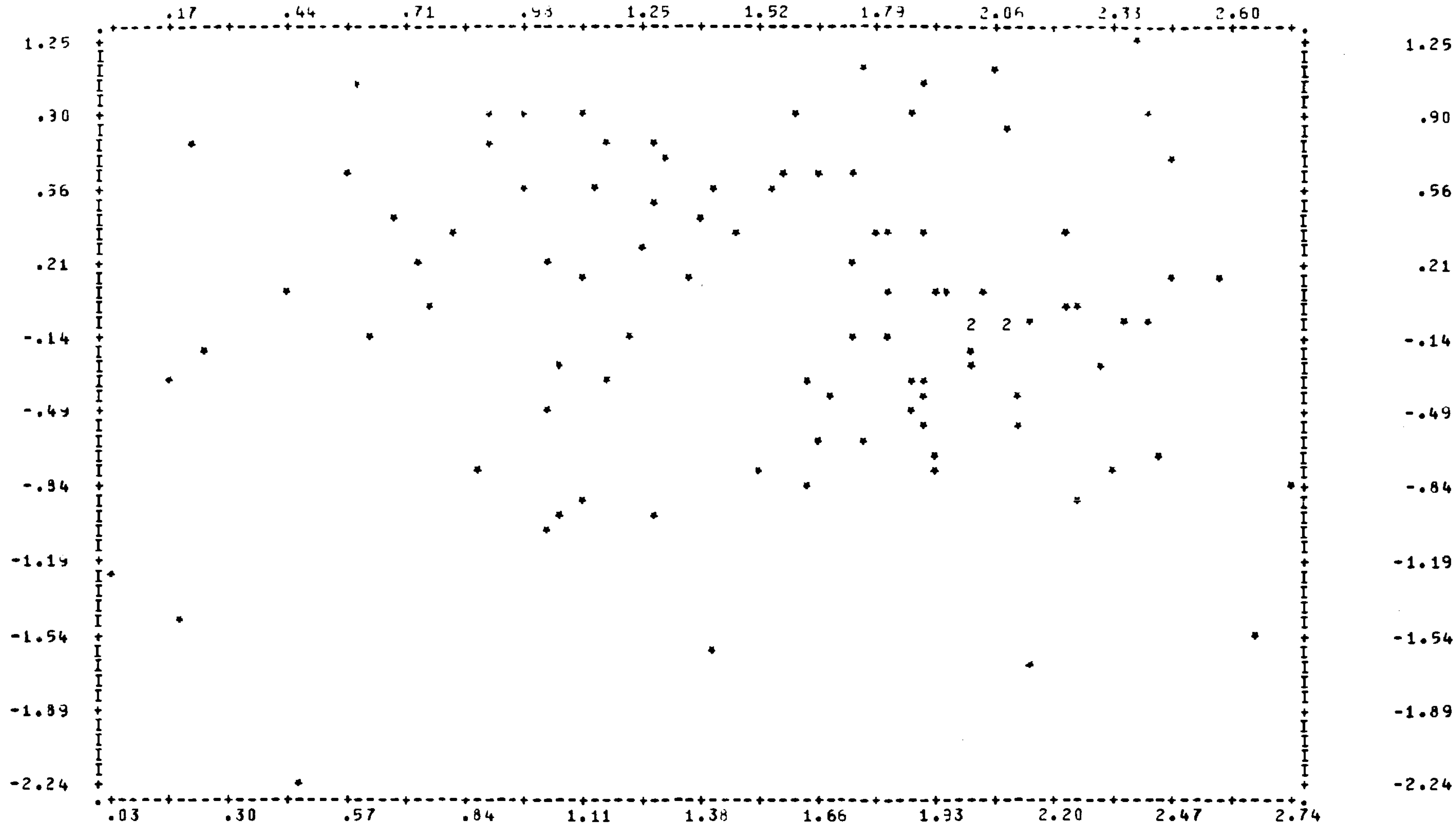


Figure 2. A Test of the Regression Assumptions.

TEST 80/02/26. 13.24.38. PAGE 5

STATISTICS..

CORRELATION (R)-	.00002	R SQUARED -	.00000	SIGNIFICANCE R -	.49993
STD ERR OF EST -	.70885	INTERCEPT (A) -	.00002	STD ERROR OF A -	.19017
SIGNIFICANCE A -	.49996	SLOPE (B) -	.00002	STD ERROR OF B -	.11254
SIGNIFICANCE B -	.49993				
PLOTTED VALUES -	100	EXCLUDED VALUES-	0	MISSING VALUES -	0

******** IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

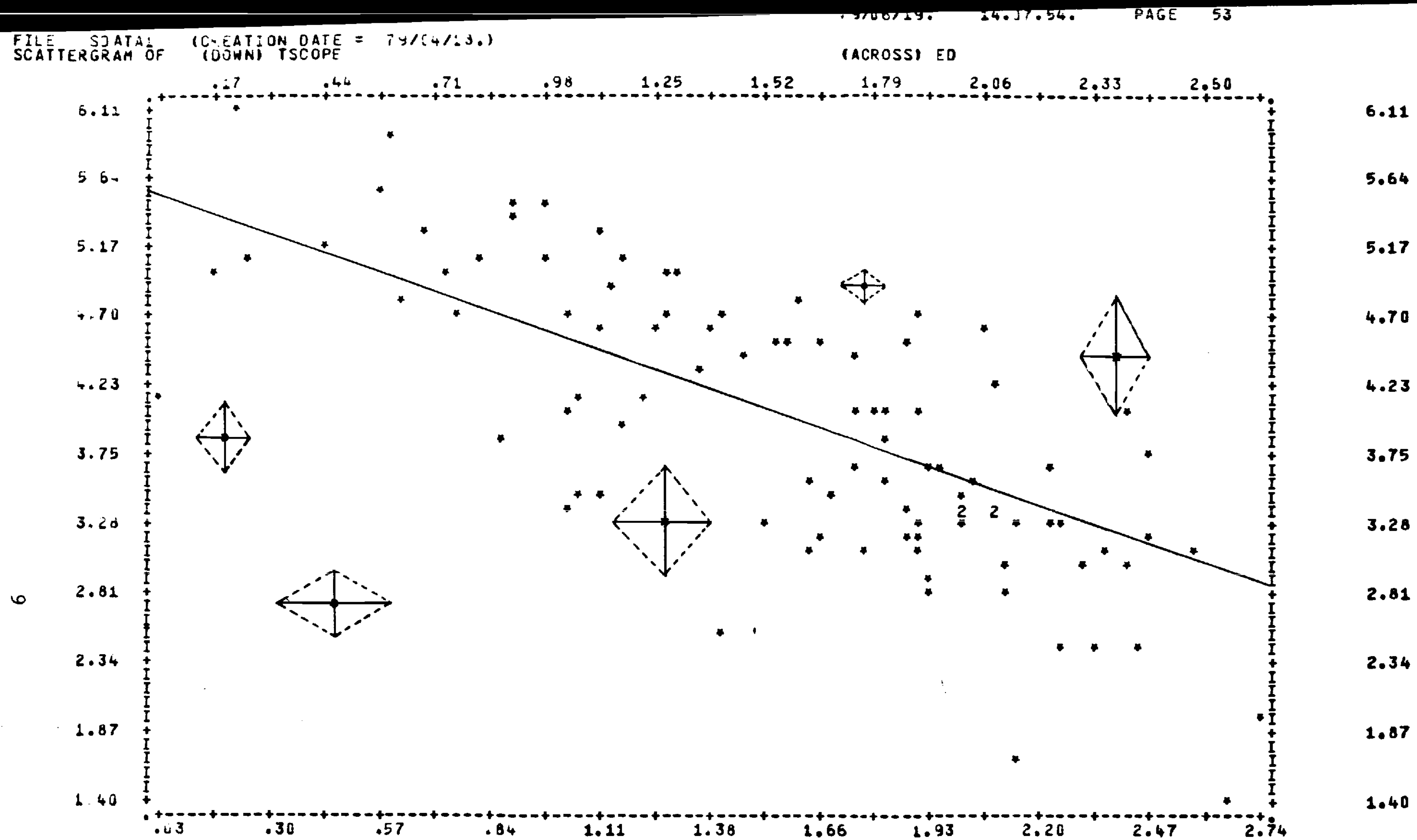


Figure 3. The Law of Word Interpretation and Instrument Precision

Diamonds Indicate the Sample Limits for the Standard Error of Measurement

Residuals are semiotically real.

The present analysis was undertaken in order to analyze these differences and search for objectively observable systematic relations. The method used employed primarily linear regression, multiple regression, and partial correlation; but also made use of some concepts from canonical correlation and non-linear regression. No non-linear regressions were actually run and a great deal of time was expended by having to use linear regression techniques as approximations for non-linear regression. In an exploratory study such as this, however, much time would have been required even if non-linear regression techniques had been available.

The work of Sperling [5; and 6] suggests that the placement errors made in reading the teescope are due to information decay in immediate memory. An analysis of the experimental procedures used in [3] showed the only appreciable source of large error (of the size required for these residuals: standard error of estimate = 0.71 lpw) to be that part of the procedure when the S, after seeing the word exposed in the teescope, looks away to write his interpretation result on the data sheet. This takes typically about two to three seconds. Psychological experiments suggest that immediate memory begins to lose information after about 500 msec. Thus the analysis of these procedures, in conjunction with Sperling's analysis, agree in suggesting semiotic processes in immediate memory as the source of the residuals in the Law of Word Interpretation.

Experiments in cognitive psychology suggest that semiotic processes in immediate memory may structurally relate information decay to the frequencies and position of the individual letters making up the stimulus word. Thus this analysis concentrated on searching for relations between the residuals and the letter frequencies and positions, but the Markov-order of the generator for each word as well as the placement error and eidontic deviance were examined as well.

The residuals were treated as corrections to the placement error value since the error source was thought to lie in the teescope readings rather than the eidometer readings. Thus correction terms were sought such that

$$E_c = E + C$$

when E_c is the corrected placement error value, E is the raw placement error value, and C is the correction term. Four correction terms were found, three of which depended on absolute value functions of the individual letter frequencies and their positions, while the fourth depended on the Markov-order of the source generator of the specimen word. Partial correlation suggested, however, that this last effect was due to confounding of letter frequencies with Markov-order, and that the actual cause of the correction term was due to letter frequency. Since three partially dependent, partially independent correction terms were found, troth of which depended on different absolute value functions of the letter frequencies and their positions and a fourth term was found with a suggested dependency on letter frequency, it is felt that these are all approximations to some other absolute value function of the letter frequencies and their positions whose form I did not happen to chance upon in the limited time at my disposal for this investigation. It is possible, then, that this other term could achieve a better correction to the residuals than all four of the present terms combined.

It should be noted that in the discussion of these correction terms which follows, figures are shown which include a computed significance value. Due to the posthoc nature of such an analysis as this, no meaning can be attributed to this significance value and it should therefore be ignored. The purpose of this analysis was not to prove any result, but to discover possibilities for the best design of future experiments and for improvements in experimental methodology.

The four individual correction terms are called POLYTRM, MAMITRM, INVSTRM, and SORSTRM, while the overall term is called CORTRM. Thus we have

$$\text{CORTRM} = \text{POLYTRM} + \text{MAMITRM} + \text{INVSTRM} + \text{SORSTRM}$$

Fig. 4 shows the scattergram of raw residuals versus the final POLYTRM. The scattergram shows poor distribution and definition, but good correlation with an r^2 of 0.144. The equation for POLYTRM in terms of the letter frequencies and their ordinal positions, indicated by the subscripts, is

$$\text{POLYTRM} = |f_1 - .370f_2 - .088f_3 - .078f_4 + .028f_5 + .100f_6 + .933f_7 - .065f_8 - .0806|.$$

Fig. 5 shows the scattergram of raw residuals versus the final MAXMINTRM (programed as MAMITRM, and V33, to fit the length and system requirements of the SPSS language). Again the scattergram shows poor distribution and definition, and good correlation with an r^2 of 0.123; however, the distribution here seems complementary to that of fig. 4, in that most points in the MAXMINTRM plot are heavily skewed toward the upper right hand area of the graph while in the POLYTRM plot most points are heavily skewed toward the lower left area of the graph. The equation for MAXMINTRM is given in terms of the maximum and minimum letter frequencies over all eight letter positions by

$$\text{MAXMINTRM} = |\max \{f_i\} - 0.700| - |\min \{f_i\} - .0125|.$$

CORTRM also contained two other correction terms, INVSTRM and SORSTRM. INVSTRM was an absolute value function of the sum of the inverse letter frequencies of the word

$$\text{INVSTRM} = |\textstyle\sum^8 f_i^{-1} - 1650|$$

and SORSTRM was simply the Markov-order of the source generator of each word.

Fig. 6 shows the scattergram of raw residuals versus the total correction term CORTRM which was obtained via multiple regression using POLYTRM, MAMITRM, INVSTRM, and SORSTRM. CORTRM is given by

$$\text{CORTRM} = 8.4*\text{POLYTRM} + 11.0*\text{MAMITRM} + .00045*\text{INVSTRM} - 0.20*\text{SORSTRM}.$$

The correlation here is excellent, with r^2 equal to 0.305; but now the distribution and definition are also good, showing that the poor distribution of figs. 4 and 5 was indeed complimentary and strongly hinting that POLYTRM and MAXMINTRM both approximate the true correction term.

79/06/14. 15.17.57. PAGE 51

FILE SDATA1 (CREATION DATE = 79/04/18.)
SCATTERGRAM OF (DOWN) RESID (ACROSS) V44

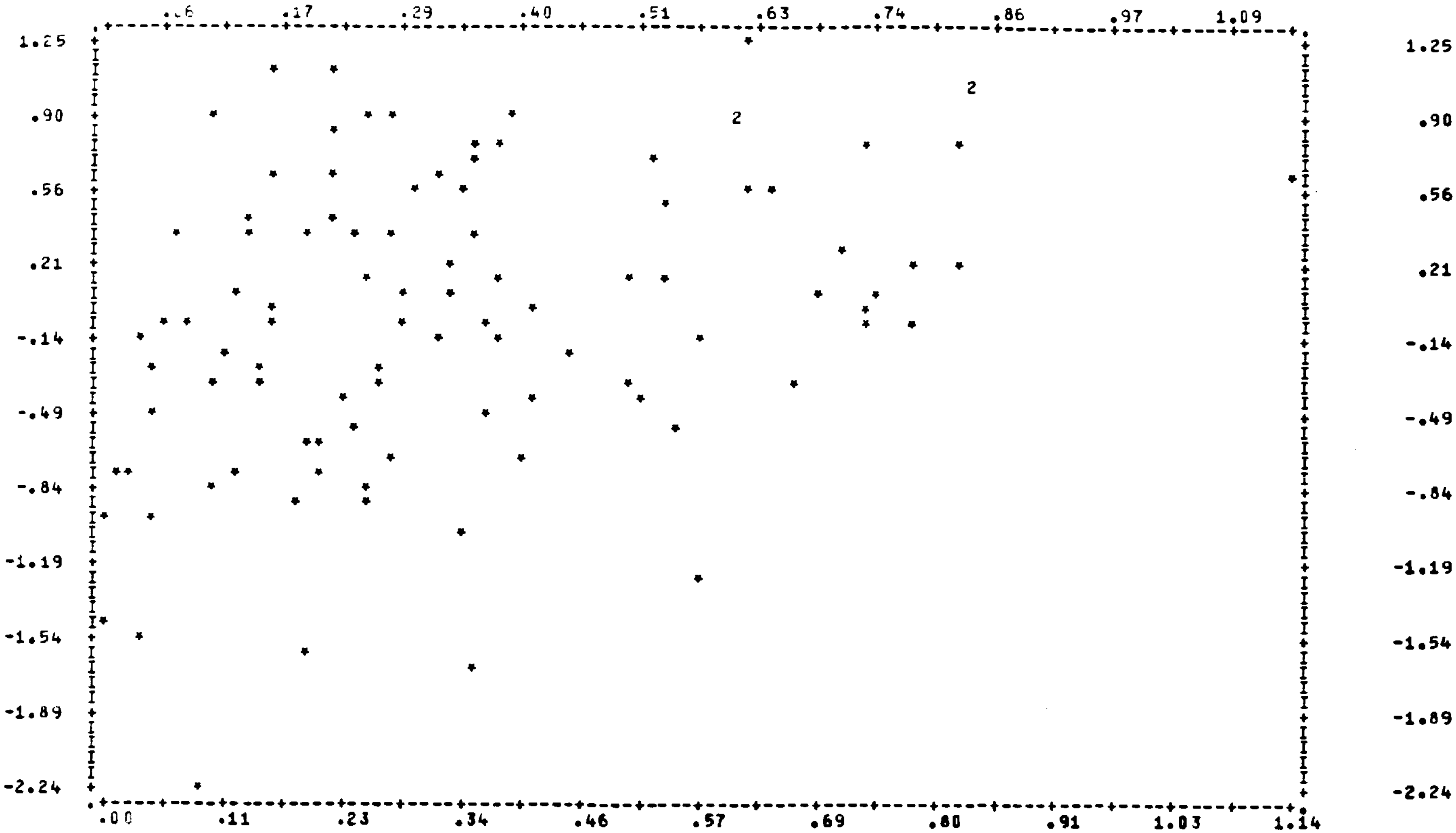


Figure 4: **Raw Interpretation Residuals Vs. POLYTRM**

where POLYTRM is the Absolute Value Frequency Polynomial Correction term, and

POLYTRM = $|f_1 - .370f_2 - .088f_3 - .078f_4 + .028f_5 + .100f_6 + .933f_7 - .065f_8 - .0806|$

STATISTICS..

CORRELATION (R)-	.37900	R SQUARED -	.14364	SIGNIFICANCE R -	.00005
STD ERR OF EST -	.65597	INTERCEPT (A) -	-.38378	STD ERROR OF A -	.11517
SIGNIFICANCE A -	.00061	SLOPE (B) -	1.10647	STD ERROR OF B -	.27291
SIGNIFICANCE B -	.00005				

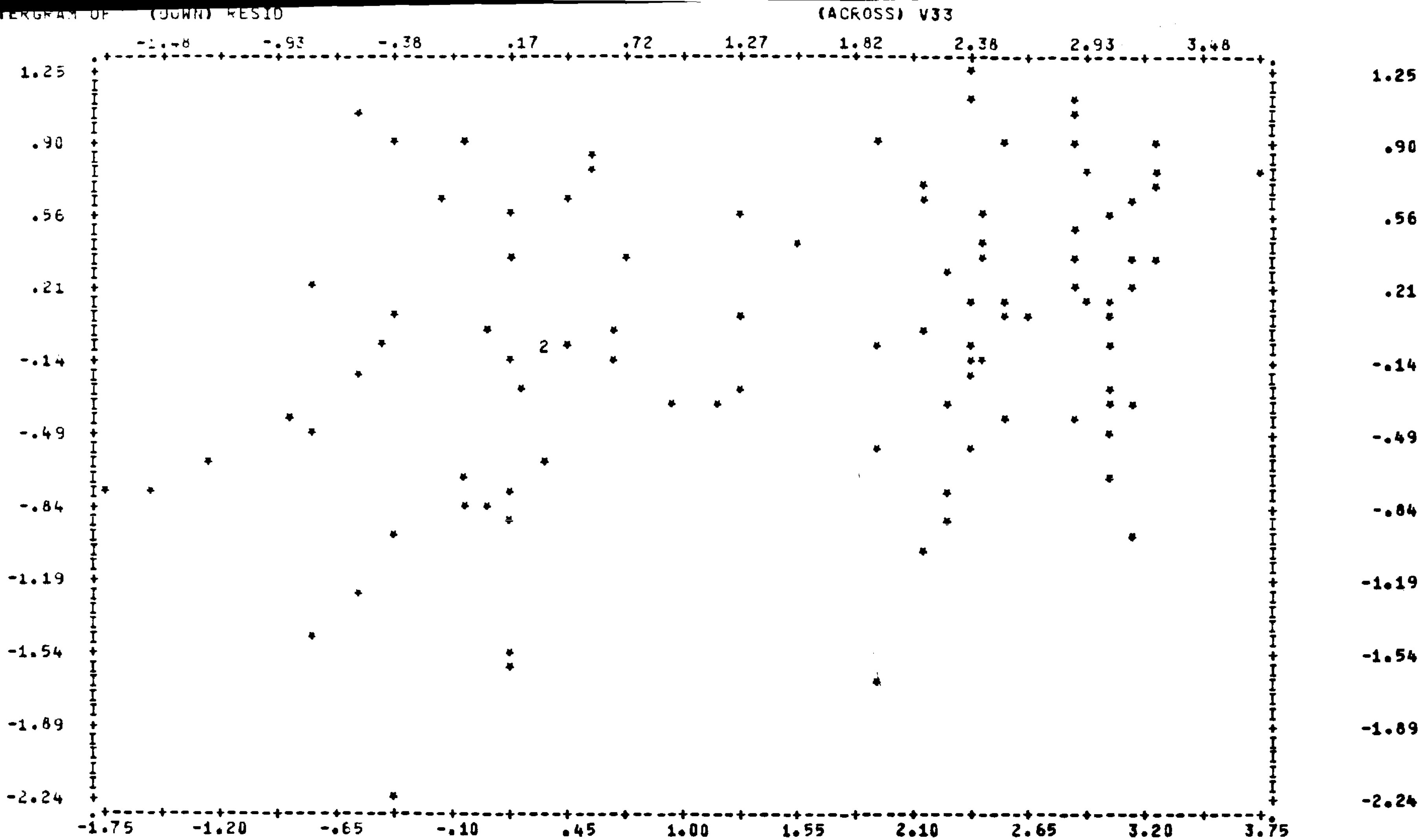


Figure 5. **Raw Interpretation Residuals Vs. MAXMINTRM**

where MAXMINTRM is the Absolute Value Max - Min Correction term, and

MAXMINTRM = $|\max\{f_i\} - 0.700| - |\min\{f_i\} - .0125|$

STATISTICS..

CORRELATION (R)-	.35072	R SQUARED -	.12301	SIGNIFICANCE R -	.00017
STD ERR OF EST -	.66382	INTERCEPT (A) -	-.25819	STD ERROR OF A -	.09622
SIGNIFICANCE A -	.00428	SLOPE (B) -	.17250	STD ERROR OF B -	.04653
SIGNIFICANCE B -	.00017				
PLOTTED VALUES -	100	EXCLUDED VALUES-	0	MISSING VALUES -	0

FILE SDATA1 (CREATION DATE = 79/04/18.)
SCATTERGRAM OF (DOWN) RAWRES (ACROSS) CORTRM

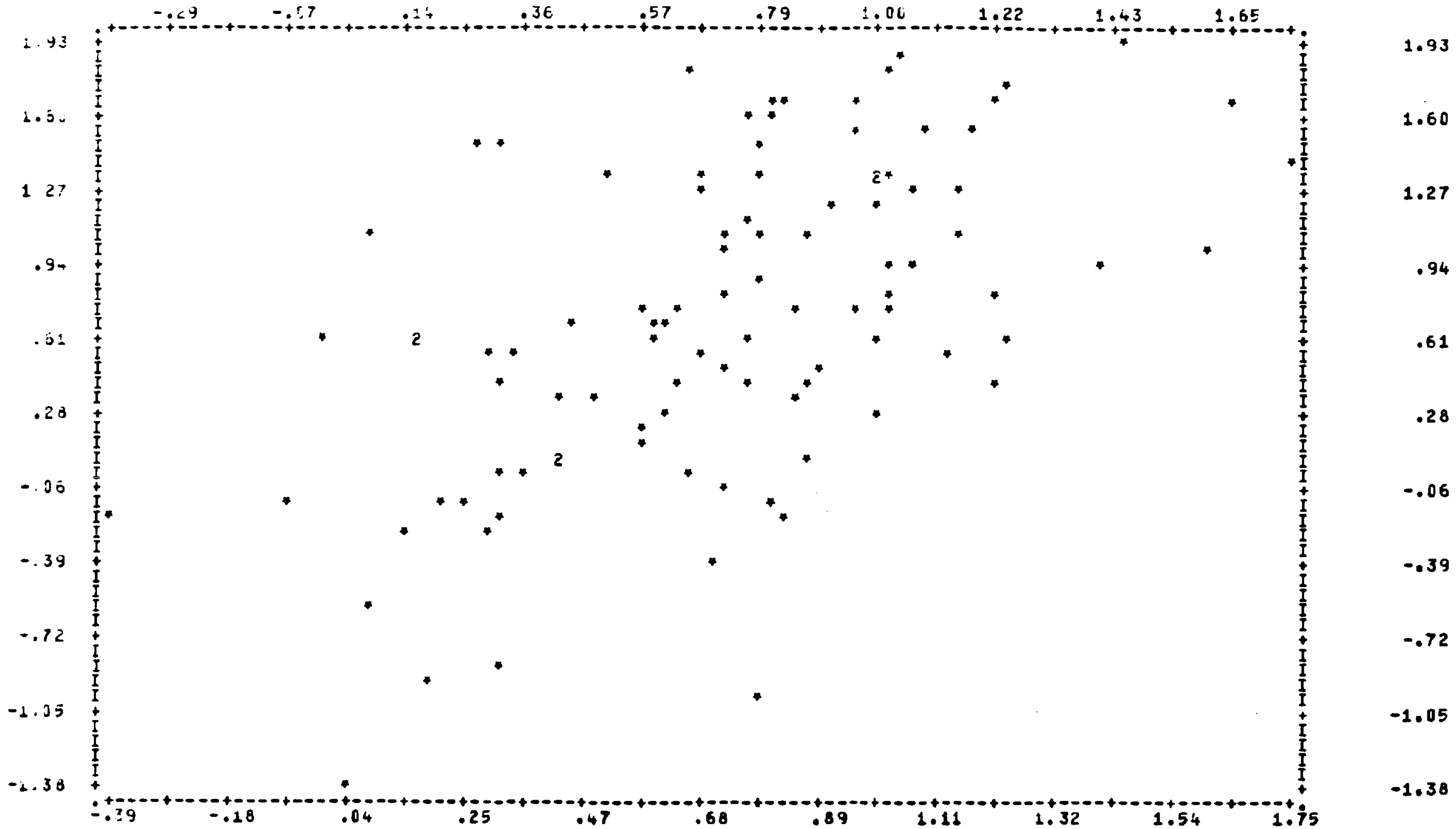


Figure 6: Raw Word Interpretation Residuals Vs. CORTRM where CORTRM is the Final Correction Term, and CORTRM = POLYTRM + MAMITRM + INVSTRM + SORSTRM

RAW INTERPRETATION RESIDUALS VS. TEESCOPE CORRECTION TERM 79/06/19. 14.07.54. PAGE 4

STATISTICS..

CORRELATION (R)-	.55234	R SQUARED -	.30508	SIGNIFICANCE R -	.00001
STD ERR OF EST -	.59376	INTERCEPT (A) -	.03552	STD ERROR OF A -	.12624
SIGNIFICANCE A -	.48259	SLOPE (B) -	1.00293	STD ERROR OF B -	.15290
SIGNIFICANCE B -	.00001				

Fig. 7 shows the final corrected Law of Word Interpretation, altho for technical reasons associated with the SPSS language this is shown as placement errors (TSCOPE) vs. corrected eidometer readings (CORED). In the results section this has been rewritten into the proper form with corrected placement errors as a function of the raw eidometer readings. The difference in the scattergrams of fig. 1 and fig. 7 is dramatic. The immediate memory correction yields a large improvement in the correlation coefficient of the Law of Word Interpretation from -0.661 to -0.777 or an increase in r^2 from 0.436 to 0.604. This yields a better than 61% improvement in the validity of the Law of Word Interpretation from 0.825 bpm to 1.331 bpm.

Of course, once the regression analysis of the residuals is complete, the regression assumptions must be tested to assure validity of the analysis, just as for the regression of the raw data. Fig. 8 shows one of the test runs. It yields a negative result: significance of R = significance of B = 0.48385 > .05, with normal distribution of the residuals about zero and constant variance. All test runs were negative, thus assuring that the regression assumptions were satisfied.

Conclusions

The conclusion has already been suggested above that the various partially correlated terms of the correction term are interdependent approximations to the one, true correction term. If this term could be found, it should lead to an improvement in the analysis of residuals and validity of the Law of Word Interpretation. However, the variance in the corrected residuals is still semiotically real, the residuals are still so large compared to the standard error of measurement for the eidometer and teescope (standard error of estimate = 0.59 placement errors per word compared to standard error of measurement = 0.05 placement errors per word) that even with the improvement that could be expected from an improved correction term a large residual term would still result. This leads to the conclusion that the concept of placement error itself needs to be re-explicated. Two glaring deficiencies are already obvious. The present concept has no way of accounting for misplaced strings, or for confusion errors. If

ABCDEFGH

is displayed, and

ACDEFGHX

is recorded by S, he is given a score of one letter correct, altho he very obviously has seen and remembered a great deal more than one letter. Also if

FQFQFQFQ

is displayed, and

EOEOEOEO

is recorded by S, he is given a score of zero even tho it is again obvious

FILE SDATA1 (CREATION DATE = 79/04/28.)
SCATTERGRAM OF (DOWN) TSCOPE (ACROSS) CORED

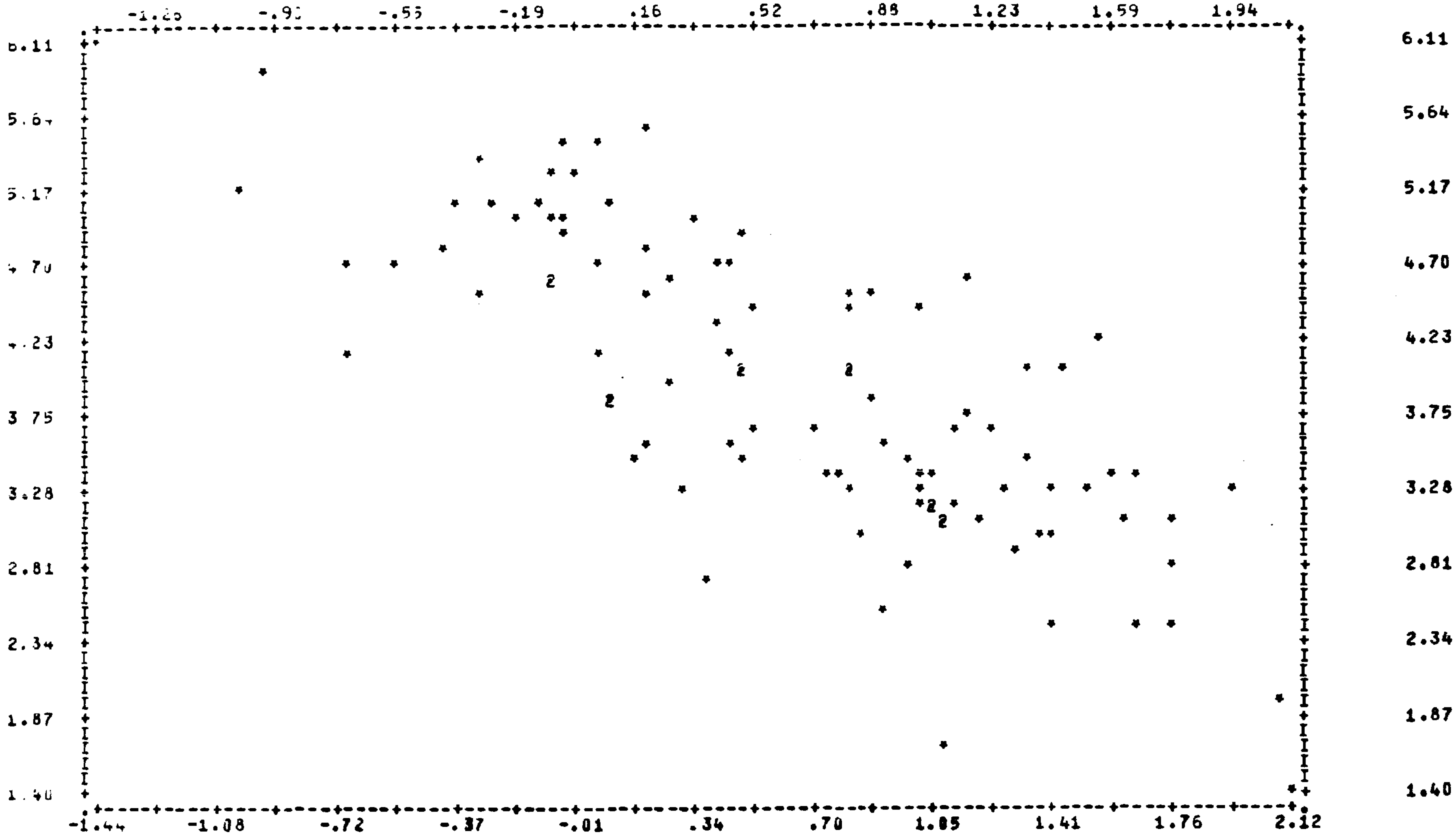


Figure 7: The Corrected Law of Word Interpretation

Corrected Teescope Readings Vs. Eidometer Readings

STATISTICS..

CORRELATION (R)-	-.77748	R SQUARED -	.60447	SIGNIFICANCE R -	.00001
STD ERR OF EST -	.59376	INTERCEPT (A) -	4.63980	STD ERROR OF A -	.07900
SIGNIFICANCE A -	.00001	SLOPE (B) -	-1.03336	STD ERROR OF B -	.08199
SIGNIFICANCE B -	.00001				

ANALYSIS OF CORRECTED INTERPRETATION RESIDUALS 79/06/19. 14.07.54. PAGE 12

FILE SDATA (CREATION DATE = 79/04/18.)
SCATTERGRAM OF (DOWN) CORRES (ACROSS) ED

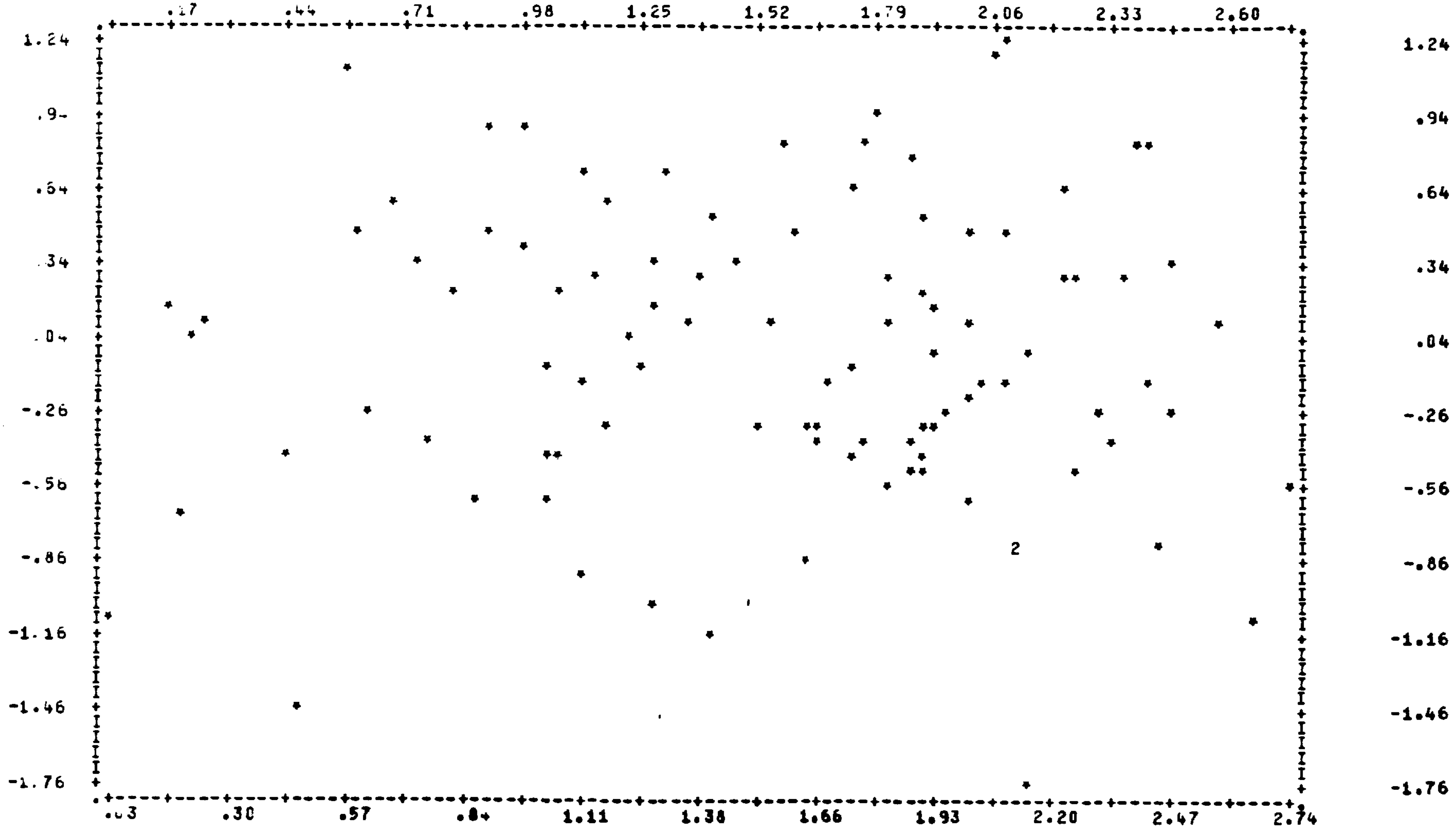


Figure 8: One Test of the Regression Assumptions

Corrected Residuals Vs. Raw Eidometer Readings

ANALYSIS OF CORRECTED INTERPRETATION RESIDUALS 79/06/19. 14.07.54. PAGE 13

STATISTICS.

CORRELATION (R)-	-.00410	R SQUARED -	.00002	SIGNIFICANCE R -	.48385
STD ERR OF EST -	.59376	INTERCEPT (A) -	.01366	STD ERROR OF A -	.15929
SIGNIFICANCE A -	.46591	SLOPE (B) -	-.00383	STD ERROR OF B -	.09426
SIGNIFICANCE B -	.48385				

that he has seen and remembered a great deal more than nothing. In this case we would suspect visual confusion to be at fault. On the other hand, if

KTZZTKTZ

is displayed, and

AECCEAEC

is recorded, we would suspect verbal confusion to be at fault. The name 'interpretation error' has already been adopted for this re-explication of M-B-P's placement error concept.

The theory that information decay in immediate memory leads to a small perturbation of the placement error measurement with the teescope explains a large percentage of the measured residuals from the Law of Word Interpretation. This correction term is apparently described by an absolute value function of the frequencies of the individual letters in the word, which is approximated by the correction term used in this analysis altho the best approximation to the true correction was by no means found.

A series of experiments should now be designed to explicate the perturbations in the measurement of placement errors due to the misplaced string phenomena.

A series of experiments should now be designed to explicate the perturbation in the measurement of placement errors due to the problem of confusion errors.

A series of experiments should now be designed to further explore the nature of the CORTRM function. This experiment must control for the maximum and minimum letter frequency in each letter position as well as the maximum and minimum difference between maximum and minimum frequency thruout all eight letter positions.

All future eidometric experiments must incorporate such controls.

And finally we must observe that the Law of Word Interpretation, a quantitative measurement of an obvious semiotic relation, the Miller-Bruner-Postman Effect, has made possible the quantitative study of a non-obvious semiotic relation, the decay of information in immediate memory. In fact, altho this has been discussed in psychological terms before, it has never been measured before, and has not even been discussed or mentioned in the semiotic literature previously. The empirical approach therefore opens up whole new vistas of semiotic nature for study and analysis.

Acknowledgements

This research was partially supported by NSF under grant IST-7827002. This support is gratefully appreciated. I would also like to thank my colleague Vladimir Slamecka for his advice, criticism, and constant support. Much help of a technical nature in learning to code and debug the SPSS language was supplied by Mary Trauner whose patience exceeded all bounds. I thank her very much.

References

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

2. Pearson, C. "An Empirical Concept of Word Shape". Paper presented at the Second Annual Conference of the Semiotic Society of America; Denver, Colorado; October 15-16, 1977. To appear in the published proceedings.

3. Pearson, C. "A New Law of Information: An Empirical Regularity Between Word Shapes and Their Interpretation". Paper presented to the 1978 Annual Meeting of the American Society for Information Science, New York; November 13-17, 1978.

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

5. Sperling, G. "The Information Available in Brief Visual Presentations". Psychol. Monogr: Gen. & Appl., 74(1960), p1-29.

6. Sperling, G. "Successive Approximations to a Model for Short Term Memory." Acta Psychol., 27(1967), p285-292.

SEMIOTICS AND THE MEASUREMENT OF SHAPE

by Charls Pearson

Georgia Institute of Technology
School of Information and Computer Science

July 1979

Preliminary draft for comments only. Not for publication or quotation. All rights reserved by the author. Intended for seminar presentation at the Technische Universität Berlin; July 1979.

SEMIOTICS AND THE MEASUREMENT OF SHAPE

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

ABSTRACT

Surveys the theoretical and experimental aspects of shape in empirical semiotics concentrating primarily on eidontic deviance, a quantitative concept of strangeness of word shape, and the eidometer, an instrument for measuring eidontic deviance. Distinguishes between theoretical aspects of the sign related to shape and the observable aspects of the sign which generate the attributes of shape. Covers studies from Markov's, Shannon's, and Miller, Bruner, and Postman's observations about semiotic aspects of shape to Pearson's discovery of the Law of Word Interpretation. Discusses the semiotic application of these studies.

SEMIOTICS AND THE MEASUREMENT OF SHAPE

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

I. Introduction:

Semioticians have for centuries talked about various attributes of signs. Some attributes often mentioned, for example, are the shape of signs, the embodiment or the medium of signs, the objects of signs, the ground of signs, the interpreter of signs, the interpretant of signs, and the interpretability of signs. One thread that continuously runs thru all these discussions is the notion of meaning. Meaning, however, is an ill-defined collection of disparate concepts and much talk about meanings has been inconsistent, speculative, nebulous, and confused because all of these concepts have been labeled by the same term -- 'meaning' --, the various relations between them were not well understood, the discussion took place in many incompatible languages, and what was said was usually nonempirical -- that is, not testable because of not being reducible to observation.

The Language of Menetics and the Universal Sign Structure Theory was a double pronged attack which attempted to solve some of these problems. The main thrust was the Language of Menetics in which I surveyed a large cross-section of the 20th century literature on meaning and attempted to design a single integrated and consistent language which would contain terms for each of the major concepts of meaning discussed in the literature, provide a grammar that would allow discussion of meaning in empirically testable ways, and be explanatorily adequate in Chomsky's sense. The successful result of this language design effort was presented in Towards An Empirical Foundation of Meaning [6] in 1977.

Falling out of this, almost as a byproduct, however, was the Universal Sign Structure Theory which postulated the principle that to each major concept of meaning there corresponds one of the major attributes of the sign. As developed, the theory attempts to explain the relationship between all the major concepts of meaning (which is calls 'internal sign components'), all the major attributes of signs (which it calls 'external sign components'), and to predict which of these will enter into empirically important regularities (laws of semiotic nature). The theory as it exists is only a prototheory in the sense that it is nonquantitative and is only useful for predicting where to look for quantitative regularities and for a first attempt at understanding the relation between the various meaning components, the various observable attributes of signs, and the various kinds of sign structure. The theory also says nothing about how this structure is processed during semiosis.

The theory predicts that shape is one generator of many observable and quantifiable aspects of signs and that these aspects will enter into empirically interesting regularities with other quantifiable aspects of signs. These aspects of shape are also of great current interest to information science. The theory also predicts the empirical importance of quantifiable aspects of the sign interpreter and the process of interpretation. This paper discusses one measurable aspect of signs, generated by semiotic shape, and a very simple and pervasive regularity which has been found to hold between it and the interpretability of artificial words. The technical term for this shape concept is eidontic deviance, which is a metrological explication of the strangeness of the shape of a sign. The regularity which holds between eidontic deviance and interpretability is called the Law of Word Interpretation. The discovery of the Law of Word Interpretation depended in an essential way on two observations: one by Shannon regarding the shape of artificial words; and one by Miller, Bruner, and Postman regarding the interpretability of artificial words.

Section II presents an introductory summary of the Universal Sign Structure Theory as a background for the following discussions of shape and interpretability. Section III introduces the technical concept of semiotic shape and discusses its relation with other semiotic observables. Section IV discusses Shannon's observation regarding the shape of artificial words and introduces the concept of eidontic deviance. Section V discusses the invention and development of the eidometer, a precise and reliable instrument for measuring eidontic deviance. Section VI discusses the Miller, Bruner, Postman Effect, and their observation regarding the interpretability of artificial words. Section VII presents the Law of Word Interpretation and finally section VIII suggests some semiotic applications of these results. All references are cited together in section IX.

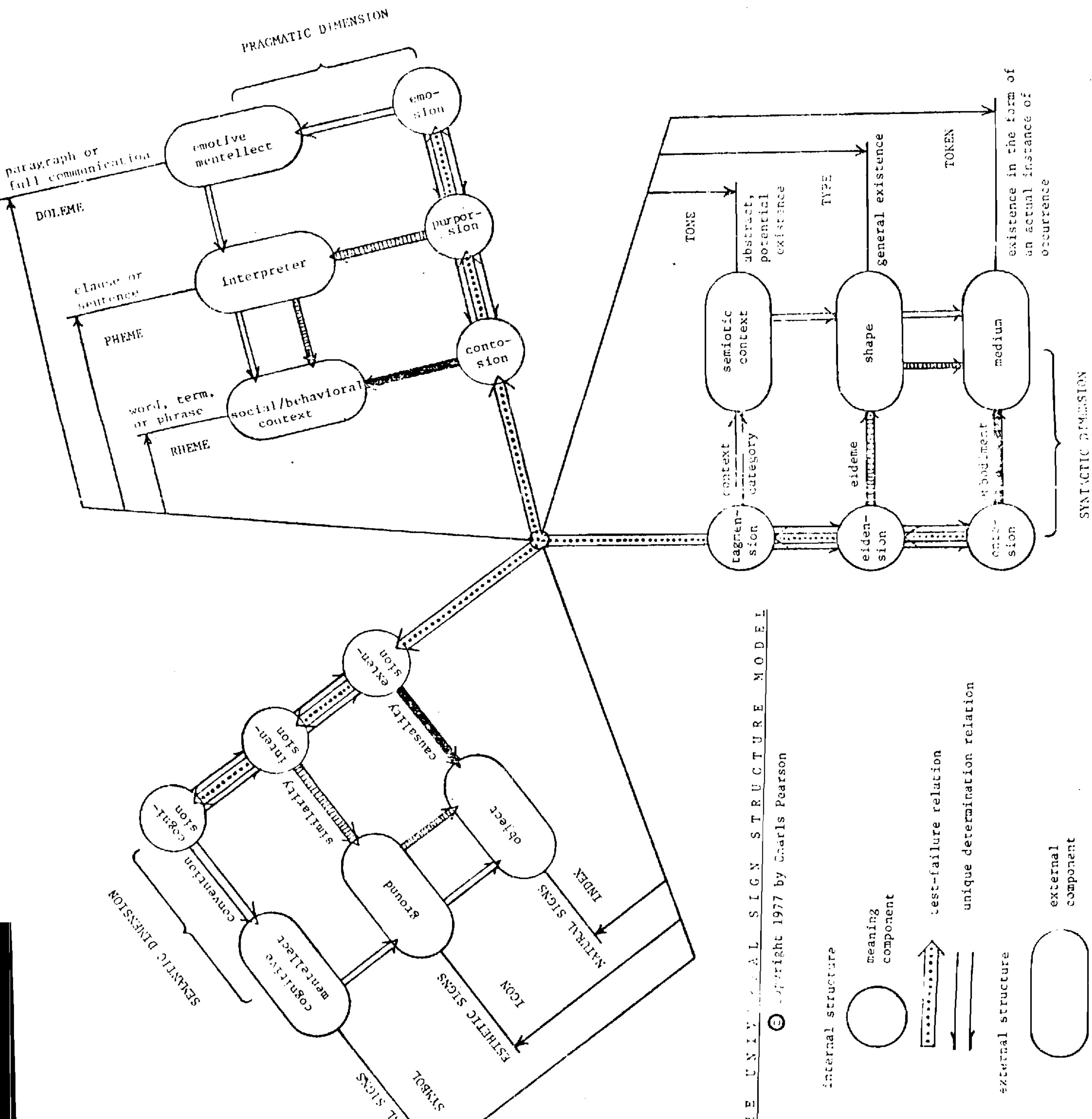
The linguistic paradigm in which the present paper is set is that of the Language of Menetics which is an empirical version of the Peirce-Morris logical-structural approach to semiotics. The theoretical paradigm is the Universal Sign Structure Theory. The experimental paradigms will be discussed at length in the paper, and the only mathematical paradigms required are those of elementary descriptive and inferential statistics including stochastic processes (Markov chains) and Shannon's uncertainty calculus (classical information theory), along with one linear algebraic equation describing the Law of Word Interpretation, and one nonlinear algebraic function defining the weighted mean familiarity of the letters making up a word.

II. Universal Sign Structure Theory:

The Universal Sign Structure Theory is composed of several primitive concepts, a relational model, three principles, and several obvious rules of interpretation or translation between the theoretical vocabulary and the observational (or less theoretical) vocabulary. The most important primitive concept is that of 'sign', which is not defined in the theory. Other important primitive concepts are 'consists of', 'stands for', and 'is interpreted (by)(within)'.

The relational model is called the Universal Sign Structure Model and is shown in Figure 1. Since one of the main purposes of the theory is to explain the interrelationships between various concepts of meaning, various theoretical and observable aspects of signs, and various information concepts, these

Fig. 1. The Universal Sign Structure Model



concepts and aspects are treated by the model as components and appear in the model as the nodes of a directed graph. Some of the relationships between the concepts and aspects then appear as the arrows of the model. These will be discussed in more detail later.

The three principles of the theory are called 1) the Trinarity Principle; 2) the Principle of Internal/External Balance; and 3) the Principle of Additional Structure. These are stated as follows:

<u>The Trinarity Principle</u>: *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.

<u>The Principle of Internal/External Balance</u>: *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.

The study of the physical, or binary, sciences can be neatly divided between theory and observation with the theoretical components of these sciences containing only theoretical concepts. However, the semiotic sciences cannot be approached in such a simple manner, perhaps because of the trinary structure of the sign, and we find the basic theoretical concept of the sign itself composed of both theoretical and observational components. The theoretical concepts are the internal components and the observational concepts are the external components. The totality of the internal components with their relationships makes up the internal structure and the totality of the external components with their relationships makes up the external structure.

We thus see that concepts of meaning are treated by the theory as theoretical concepts and generators of information are associated with observational concepts, as are the various information measures themselves. Note that the term 'information measure' is used here in the empirical sense and not in the mathematical sense. A mathematical information function can be, but need not necessarily be, a model of an empirical information measure.

<u>The Principle of Additional Structure</u>: *Whenever a sign has more than the minimum structure, the additional structure is built up from the center out (as per Figure 1), and for each dimension independently.*

The Peircean Taxonomy of signs can be explained by the theory by means of nine representation* theorems which are derived from the three principles and the diagram of Figure 1.

For instance, we will be dealing in this paper with rhematic symbolic tokens. From [10, p8] we have theorem 3: A sign is a token iff it has all three levels of syntactic structure. It therefore has three components of

---

*'Representation' is used here in its mathematical rather than its semiotic sense.

syntactic meaning (tagmension, eidension, and ontosion) and three external syntactic components (the semiotic context, the shape of the sign, and the medium in which it is embodied). Also from [10, p12] we have Theorem 6: A sign is a symbol iff it has all three levels of semantic structure. It therefore has three components of semantic meaning (extension, intension, and cognesion), and three external semantic components (the object, the ground, and the cognitive mentellect of the sign). Again from [10, p13] we have Theorem 7: A sign is a rheme iff it has exactly one level of pragmatic structure. It therefore has one component of pragmatic meaning (contension) and one external pragmatic component (the social/behavioral context of the sign). Combining theorems 3, 6, and 7, we have the obvious corollary that the structure of a rhematic symbolic token is given by Figure 2. The other representation theorems can be found in [10].

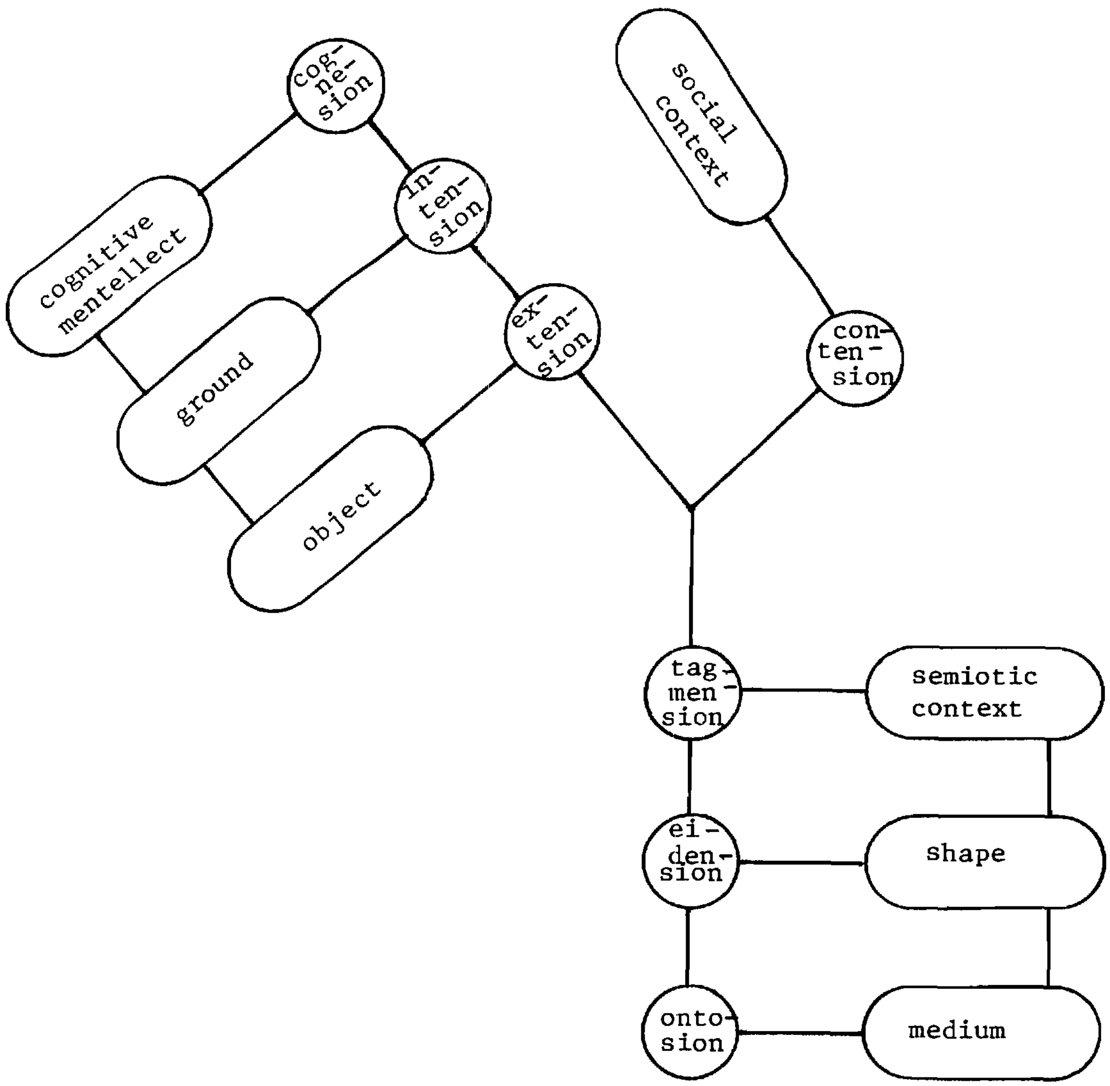


Figure 2. The Structure of the Rhematic Symbolic Token

Other terms which are useful in using the theory, and which may be defined in terms of the primitive concepts of the theory are: dimension, meaning, information, and levels. A dimension is a maximal, independent, substructure of the sign. The three dimensions of the sign are 1) syntactic; 2) semantic; and 3) pragmatic as indicated in Figure 1. Meaning is a theoretical subconcept of the sign comcept. The meaning components are the nine internal components of the sign and thus every possible sign has at least three components of meaning while dolemic symbolic tokens have all nine components. Meaning includes such components as tagmension, extension, intension, contension, and purporsion.

Information is the observable, or measurable, aspect of the sign and is generated by the information generators or external components of the sign. Each of the nine information generators generates an infinite number of information measures. For instance shape, an external component or information generator, generates length, pattern, complexity, strangeness, etc. -- all information measures. In some situations length can be modeled mathematically by Shannon's selective information function, and complexity can be modeled by Kolmogorov's algorithmic information function.

Level refers to a combination of one meaning component and its associated information generator. Thus a token has three levels of syntactic structure while a type has only two levels of syntactic structure.

III. Shape:

Shape and eidension make up the eidontic level of syntactic structure. In the Universal Sign Structure Theory the technical concept of shape is an explication of our nontechnical or intuitive concept of shape. Shape is that collection of attributes by which we distinguish one kind of sign from another kind of sign when both are embodied in the same medium within similar semiotic contexts. For instance, in the case of the letters X and O embodied in printers ink on white paper, it is our nontechnical, intuitive concept of shape by which we distinguish them. Here the technical concept of shape matches our nontechnical concept of shape. In the case of printed words in an alphabetic language shape refers to orthographic shape and is determined by spelling. In the case of spoken words shape refers to phonemic shape and is determined by the phonemic patterns. Finally, an example in which the technical concept deviates drastically from the intuitive concept -- marine communication in which we distinguish one ship's signal-flag from another by its color. In this case, the color of the flag would be the shape of the sign.

Since we will be dealing shortly with words printed in an alphabetic language, we should look briefly at the appropriate concept of shape. As mentioned above shape refers in this instance to the spelling, or orthographic patterning. It is obvious that such aspects are used to distinguish written words as length (in letters per word), vowel-consonant patterning (as for instance CVC), complexity (measured in phonemic generation rules), and strangeness (measured in °ED). However, since shape is an empirical concept as opposed to a mathematical concept, the number of empirical aspects of shape is unlimited. We cannot determine apriori what properties of shape will be discovered, by some experiment not yet invented, to be semiotically interesting. Weissmann referred to this as the open-ness (*porosität*) of nature.

Eidontic properties (generated by shape) can be experimentally related in empirically interesting (simple regularities) ways to other eidontic properties as well as to properties generated by each of the other external components; for example: strangeness (eidontic deviance) is probably related to eidontic complexity by eidemic production rules (phonemic production rules, graphemic production rules, etc.); on the other hand shape is related to the embodiment thru the type-token relation, etc.; shape has an especially strong relationship to the ground of a sign in the case of icons; and shape is related to the emotive mentellect of signs thru Osgood's Semantic Differential. One ought, therefore, to expect an empirically interesting relation to show up between measurable aspects of shape and the interpretation of the sign. As mentioned in section I, the key to discovering this relation was found in two observations, one by Shannon, and one by Miller, Bruner, and Postman.

IV. Shannon's Observation and the Markov Effect:

In 1948 [12] Shannon remembered an effect discovered earlier by Markov, that information sources producing words by those stochastic processes now called Markov-chains of finite order produce words that look more like natural language, the higher the Markov-order.

This effect has sometimes been called Markov's Law, but it doesn't have the interval measurement structure required of empirical laws and should therefore be called the Markov Effect because of its ordinal measurement structure. However, I hesitate to even call it this because it lacks a necessary empirical consistency. It is a relation between a mathematical parameter and an empirical property.

The requisite empirical consistency must be acquired by empirical explication. The object is to achieve a relation between two empirical properties. If we try to explicate 'Markov-order' independently of the shape concept we find that there is no empirical content to the relation other than the trivial one which says the more artificial words look like words of a natural language, the more they look like words of that natural language. We therefore explicate shape in terms of Markov-order using it to develop a quantitative concept of the deviation of word shape from a given natural language and an instrument for measuring it. The Markov Effect completely disappears in this process but can be partially recaptured by pseudo-statistical procedures involving "average Markov-orders".

This leaves us with an empirically measurable concept of strangeness but no empirically independent relation whose description requires its use. A new empirically independent relation was, however, suggested by the Miller, Bruner, Postman observation which is discussed next. The invention of the eidometer and the explication of the eidontic deviance concept is discussed in section VI.

V. Cattell's Phenomena and the Miller, Bruner, Postman Effect:

In 1885 Cattell [1] discovered that words exposed in a teescope (tachistoscope) are interpreted wholistically, that is four or five letters making up a word are perceived by a unit process in about the same time as a single letter is perceived. This is called the Cattell Phenomena.

In 1954 [5] Miller, Bruner, and Postman discovered how to add ordinal measurement structure to the Cattell Phenomena and thereby discovered the Miller, Bruner, Postman Effect (M-B-P Effect) involving the interpretation of words produced by an information source of finite Markov-order. By using Markov-chains of finite order to control the shape of the words in their experiments and by using Shannon's information calculus to model the statistical uncertainty of the individual letters in the artificial words, Miller, Bruner, and Postman discovered a surprising result. As expected, the average errors of interpretation under tachistoscopic conditions decreased, the higher the Markov-order; but the surprising result was that the amount of information (in Shannon's sense) obtained from a tachistoscopic exposure is independent of the order of approximation to English letter sequences. Therefore differences in the interpretation error rate can be predicted from a knowledge of the information structure of English.

Again this was a relation between a mathematical parameter and an empirical concept. Miller, Bruner, and Postman could control shape but they could not measure it, and so the Miller, Bruner, Postman Effect also lacks empirical consistency. This time, when we explicate the Markov-order concept in terms of the quantitative concept of strangeness developed previously, the Miller, Bruner, Postman Effect does not disappear, but becomes a relation between two empirical, measurable concepts, possessing universal regularity, interval measurement structure, and an extremely simple algebraic description. It becomes the Law of Word Interpretation governed by

$$E = a + bS.$$

This is discussed further in section VII.

VI. Development of the Eidometer and the Concept of Eidontic Deviance:

Intuitively it would seem that the Markov Effect and the M-B-P Effect are related. However, as mentioned above, both are only semiempirical and without a quantitative concept of word shape and an instrument to measure it, it was not possible to study this relation empirically.

Shannon observed that it was *obvious* that the higher the Markov-order of an information source, the more the signs that were generated by that source looked like the language from which the frequency statistics for the source were compiled. For instance LYDRA was generated by a third-order source and PBXQQZKTW by a zero-order source. Now there are two immediate problems with using this observation empirically. First, just how much more do the words produced by one source look like the appropriate language than those produced by another? Do the words produced by a word source for American of Markov-order 2 look twice as much like American words as those produced by a source of order 1? Or, is the amount by which words produced by an order 3 source look more like American words than those produced by an order 2 source greater than, equal to, or less than the amount by which order 1 words look more like American than order 0 words? These questions could not be answered by Shannon with the conceptual tools available to him. This problem involves quantification and empirical operationalization of an intuitive concept. The intuitive concept for which quantification is wanted would appear to be 'the amount by which signs look *more* like those of a given natural language'.

The second problem concerns just which concept is to be explicated. It turns out to be simpler and more useful to explicate 'the amount by which signs look *less* like those of a given natural language'. This is analogous to the physicist's explication of 'more of a coldness' in terms of 'less of a hotness'.

The proper concept to be explicated empirically is 'by how much does the shape of a given word deviate from the typical shape of words in a given language'. Because the Universal Sign Structure Theory uses the term 'eidontic' to refer to semiotic aspects of the shape level, this concept bears the label 'eidontic deviance'. We may interpret eidontic deviance loosely as "How strange does a word look relative to a given language?". For this reason, the term 'strangeness' is sometimes used in place of 'eidontic deviance', but it is always used in this technical sense, that is, as a synonym for eidontic deviance. Thus eidontic deviance is an empirical explication of our intuitive feeling that LYDRA looks more like a word of the American language than PBXQQZKTW.

It should also be noted that it is not shape itself that is being measured, but only an aspect of shape -- its strangeness, or eidontic deviance. This is consistent with the interpretation of the external components of the sign as information generators. They are not the measurable aspects themselves, but only the generators of those aspects.

Once the proper concept was chosen, it was quantified initially by simply adopting Chomsky's four lists as an instrumental standard with which to compare other words. Every instrument needs a scale. This is simply a set of names and a linguistic convention for assigning the names to the measured properties. For our scale we chose the Markov-order numbers of the lists. The Markov-orders are mathematical entities and the scale values are empirical entities and there is no logical connection between them even tho in this initial case they do have the same numerical values. We simply made a convenient, but arbitrary, choice of names for our scale values. We called this version our Mark Two (Mk. II) Eidometer. We reserved the name Mk. I for a very simple demonstration instrument consisting of only two lists; one list of extremely normal looking words, and one list of extremely strange looking words.

One result of our choosing the Markov orders for our scale values was that the scale appeared to run backward with very strange looking words having a measured value near zero and very normal looking words having measured values of three and above. This has been only a minor inconvenience in explaining the instrument to others and has not hindered the actual development at all. It has persisted thru the design of the Mk. V instrument. Nevertheless, in the Mk. VI, when the next major redefinition of scale takes place, it will be reversed so as to eliminate this minor confusion.

With an instrument and concept available, the further development of the instrument and explication of the concept could take place jointly, each aiding the other. We first set up a standard list of 100 words with which to calibrate the Mk. II and all later instruments. By measuring these words we not only created the ability to compare the Mk. II against later designs, but we could assess the precision, reliability, and validity of our initial eidometer. This gave us confidence in our original system but told us that we had many practical design problems to overcome.

A major problem with the early lists was due to the fact that each list was constructed by using all the words generated by a single source. Since generation is a probabilistic process, this means that even in a zero-order process there is some small probability of generating a normal looking word, and even in a fourth-order process there is a small probability of generating some very strange looking words. This problem was eliminated in the Mk. III eidometer by introducing a classification step. Individual words were assembled and subjects were asked to classify them into categories. If one word was much stranger than another they should go into two different categories, but if it was difficult to tell which was stranger they should go into the came category.* These categories were then assembled into lists and the average Markov-value of all the words on a list assigned as the scale value for that list.** The Mk. III eidometer had eight lists ranging from 0.0 to 2.0 in increments of 0.25 with one list at +0.25 missing.

Several of the lists on the Mk. III looked distinctly out of sequence and gave problems with precision and reliability in these regions of the scale. The high end (less strange) did not have any normal looking words. For this reason the scale was extended to 3.0 and a sorting step was added in the Mk. IV design. Once the lists were assembled, they were sorted by new subjects to assess the objectivity of their ranking. Lists which caused excessive problems could be redesigned or replaced completely. It was also decided at this point to add one blank list at each end to allow for and to remind users that words could in fact fall off either end of the scale. These were assigned the values -0.20 and +3.20. The extension to 3.0 was made possible by the development of a computational algorithm to automatically generate words by Markov information sources from 0 to 3 [2;3]. Words could also be generated using variable masks to tailor their shape. Later an algorithm was developed by Flowers [3] for generating words of all finite Markov orders but this was not working yet at this time. Also at this point (the design of the Mk. IV eidometer) the present scale definition was fixed giving the eidontic deviance (°ED) scale. Thus the Mk. IV had 18 lists ranging from -0.20 °ED to +3.20 °ED in increments of ±0.20 °ED. It was felt at the time that the ±0.25 increments used on the Mk. III were too far apart and that ±0.20 would contribute to better precision. Later experience has not borne this out. In fact serious thought is being given to going back to ±0.25 °ED on future designs to increase reliability and ease of use.

---

* I wish to thank P.J. Siegmann for suggesting this solution.

** The average Markov-order is not a meaningful mathematical concept. This step is legitimate only because we are arbitrarily creating a name for an empirical property.

The performance of the Mk. IV was described in [7] and can be summarized by Table 1.

Table 1: Mk. IV Eidometer Performance

Trait	Performance (bpm)
Precision	4
Reliability	3
Validity	1

This means that on any given measurement the standard deviation is about ±0.20 °ED, altho by averaging over several subjects, the standard error was reduced on some calibration runs to an average of about ±0.05 °ED. The validity measurement will be discussed in section 7.

There are still some problems with the list design technique used in the Mk. IV. Each list is still the result of experiments on a single subject. As a result, each list contains a few idiosyncracies which reduce measurement performance. Several attempts have been made to overcome these deficiencies by changing sorting techniques [9], by using cluster analysis [11], and by developing a technique of classification called the rank-averaging method [4]. Of these, the rank-averaging method is far superior and is being used to develop the Mk. V design. Many other very minor changes are also being incorporated into the Mk. V design. However, the Mk. V will still use the °ED scale introduced with the Mk. IV.

Once the Mk. V has been built and calibrated, a new scale, called the deviometer scale (dev.), will be introduced to incorporate all the knowledge gained in the study of eidontic deviance with the Mk. IV and Mk. V instruments. At this point, the Mk. VI, a radically new design with a vernier scale, will be developed. We are hoping to achieve a precision of 6 bpm, or about 1 part in 64, with the Mk. VI; with the capability of achieving standard errors in the range of 8 bpm, or about 4 parts in a thousand, or about ±1/2 percent with as few as 10 to 20 measurements. The Mk. VI will be discussed further in section VIII.

VII. Law of Word Interpretation:

The relation between word shape and interpretability has been studied using the concept of eidontic deviance and using experiments involving the eidometer and a teescope. Thruout a wide range of word shapes this relation appears to be linear, leading to a new law of semiotics which is a quantitative description of the Miller, Bruner, Postman Effect.

Conceptually, this law is an empirical relation between the semiotic context, the encoding system, the interpretability of natural language, and its interpreters. The philosophic import of this law is that it relates two different information measures together into one empirical generalization — or law of semiotics. Of these measures, one is experimentally related to interpretation errors, and the other to the perceived shape of the sign, which is itself experimentally related to the redundancy of the semiotic context and the Markov order of the information source. The eidontic deviance is also related to measures of pattern and length, possibly Kolmogorov's measure of algorithmic information and the logarithm of the length. Theoretically this law impacts our theories of sign perception, information loss from immediate memory, interpretation of signs, and coding of information.

Of the two instruments used in this study, one — the teescope — is a classical instrument of experimental psychology [5]. The other, the eidometer, is a new invention, which was described in the previous section. The teescope used in our studies to date was a Polymetric model 0959 two-field teescope. A vernier pot yields three-digit precision on exposure setting, while a three-position range setting switch gave us a total exposure range of 10 to 10,000 milliseconds (ms). All experimental runs were carried out with the shortest exposure control setting of 10 to 100 ms. In psychology, the teescope is used with various stimuluses of presumed known structure for the purpose of investigating the structure of the behavior response of the subject. In semiotics, the teescope is used with various subjects with presumed known behavioral response for the purpose of investigating the structure of the semiotic stimulus. This, in fact, is a characteristic difference between psychological experiments (especially psycholinguistic experiments) and semiotics experiments.

The study was conducted by generating a set of 100 standard specimens, each eight letters long, using the 29-letter American alphabet (26 regular letters, plus hyphen, apostrophe, and space), and ten different information generators. Four of these were of fixed Markov orders 0, 1, 2, and 3. The other six were variable mask generators of shape (0,1), (1,0), (1,2), (2,1), (2,3), and (3,2),[3]. Thus a range of artificial word specimens were generated covering the span of the eidometer. Ten adult, native literate S's measured each of the 100 specimen words. Each S measured the strangeness of the words in a different random order using the eidometer. The ten readings for each word were averaged to get the final deviance for each word.

Ten different adult, native literate S's measured the placement errors obtained by viewing the same 100 words in the teescope as per the procedure used by Miller, Bruner, and Postman in [5]. Each S was calibrated on two sets of twenty warmup words in order to get a measure of his perceptual response rate and to adjust the exposure setting of the teescope for maximum sensitivity to that S. The teescope was set to a standard setting slightly less than 100 ms. for each S on his initial warmup and then adjusted after each warmup trial of twenty words to obtain as nearly as possible a 50 percent error rate. The second adjustment was usually minor and typical final exposure rates ranged from 15 ms. to 50 ms.

After the final run for each S on the 100 specimen words, and the errors had been measured, any deviation from 50% for that S was corrected for by dividing the error rate for each word by the ratio of that S's overall error rate to the ideal value of 50%. With 100 eight-letter words, the ideal was 400 errors per S. For each word, these corrected values were then averaged to get the final measurement of the interpretation error rate. A linear correlation between the eidometer measurements and the teescope measurements gave a correlation coefficient of

$$r = 0.66$$

with r^2 about 44% for a validity of about 1 bit. For 97° of freedom this is significant at the $\alpha = .00001$ level. Sample data values are shown in Table 2.

Table 2: Sample Interpretation Data

#	Word	Order	°ED	Error Rate
3	STIONYTH	3	2.02	3.47
70	DENNMSAO	1	1.74	3.71
28	OOKORGED	(3,2)	1.36	4.35
97	PDXPUYJ-	0	0.44	5.21

Miller, Bruner, and Postman originally attributed the interpretation error effect to perceptual semiotic processes [5]; however, Sperling has more recently interpreted it as a semiotic process associated with immediate memory [14; 15]. The eidometer, however, still appears to be measuring our perception of the strangeness of the shape of the sign, and so this law would appear to relate a perceptual semiotic process to a semiotic process associated with immediate memory. More light is thrown on this analysis by the next comment.

With a correlation of only about 66%, there is much scatter in the data. Is this scatter due to instrumentation and measurement effects, or does it represent real differences in the data? With an evaluated precision of 4 bits/measurement (bpm) and an evaluated reliability of 3 bpm the scatter is way too much to be explained by instrument and measurement effects. It apparently represents real semiotic differences in the specimen words. A study of this question is now underway. The differences do not seem to have anything to do with eidontic deviance, interpretation errors, or the Markov generator of the specimen words. One function that appears to be successful in explaining much of this scatter is the weighted mean familiarity of the letters making up the word:

$$F = \left| \sum_1^8 a_i f_i - b \right|$$

where f_i is the relative frequency of the ith letter of the word, a_i is the weighting coefficient of f_i, and b is a constant. These studies are still in process, however, and no final statement can be made.

The conclusions to be derived from this study are: 1) that the strangeness of the shape of artificial words relative to a given natural language may be reliably measured; 2) that this measurement gives us the means of describing quantitatively the Miller, Bruner, Postman Effect; and 3) that in the range of data studied, this description is best given by a simple linear equation:

$$E = a + bS$$

where E is the tachistoscopic error rate; S is the strangeness measured in eidontic deviance units, °ED; and a and b are constants.

Since this is an empirical generalization relating the measurements of two different semiotic processes, I call this a law of semiotics, specifically, the Law of Word Interpretation. This law is pervasive, having shown up dramatically in many kinds of experiments over the last five years of investigation.

Another semiotic measure, the weighted mean familiarity of the letters making up the word, may represent a correction factor which will eventually enable us to make our understanding of this law much more precise and also to investigate more subtle semiotic aspects such as the processing of signs in immediate memory which are not amenable at all to the more intuitive methods of semiotic analysis.

By claiming the Law of Word Interpretation as a law of semiotics, I do not intend to claim that I have anything like a theory of even a conjectured explanation of this law, other than the Universal Sign Structure Theory which predicts the existence of the empirically observable relationship, but only that I know how to objectively and quantitatively reproduce a pervasive regularity in our observation of semiotic processes. However, we may speculate that this law will give us a tool that will eventually help us to decipher the way human processors encode signs in the mind, and/or will lead us to a further understanding of the structure of signs in that any theory of semiotic structure will eventually have to account for this law.

VIII. Semiotic Applications of Eidontics:

Four applications of the quantitative, empirical study of the shape of signs to the further study of semiotics will be discussed in this section. These include 1) the measurement of the redundancy curve for natural language; 2) improvements in the eidometer; 3) experiments between shape and other components of the sign; and 4) further investigations into the nature of shape.

In contrast to the scientific applications to be discussed here, when the word 'applications' is usually mentioned, one typically thinks of technological applications. Several purely technological applications of eidometry exist, for instance, the diagnosis of reading and spelling deficits and the diagnosis of learning disabilities with foreign languages. However, the discussion of these technological applications is outside of the scope of this purely scientific discussion of the semiotics of shape.

In [13] Shannon introduced the concept of redundancy in natural language as a function of shape. Shannon and others have made mathematical estimates of the upper and lower bounds for this curve, but it has never been measured before. Part of the problem was in measuring redundancy itself, but a more fundamental problem was that there was no way to measure shape. They could control it artificially by controlling a mathematical (nonempirical) parameter, but they could not measure the resulting shape.

In their analysis of the Interpretation Error Effect, Miller, Bruner, and Postman used Shannon's estimates of the redundancy curve in an essential way. Thus in solving the problem of shape measurement for the Miller, Bruner, Postman Effect, the eidometer combined with the Law of Word Interpretation gives us the capability for the first time of measuring the redundancy curve for natural language.

This would appear superficially to be a circular use of Shannon's redundancy curve. Actually, it is not. It is just one example of the bootstrap process, so necessary in science, of using one piece of knowledge to develop information about something else and then using this result to gain more precise knowledge in the original area. In addition, the precision which can be achieved with the eidometer will yield precision limits on redundancy measurements much finer than the mathematical bounds now existing.

The second application of eidometry to be discussed concerns the metrology of eidometry or the measurement of the eidometric measurement process itself. The present Mk. IV eidometer is the third in a series of revisions and refinements to the original eidometer concept in an ongoing effort to improve the measurement performance of the instrument. This means increasing the precision, repeatability, and validity of eidometric measurement. These detailed, onerous, steps would seem mundane to the basic considerations of science and of interest only to instrument engineers, metrologists, and technicians. Actually, nothing could be further from the truth. It is the investigation of and improvement of the measurement of a given concept that gives us scientific understanding of the nature of that concept and allows us to further explicate a concept and explore its underlying relationships. It is in the metrological trenches that the battles of science are won. The history of metrology is practically the history of science itself.

From the start the eidometer has been beset by many idiosyncracies. Some of these were discussed in section VI. There are still two notable problems with the present design. First, each list is constructed in an experiment involving a single subject, altho different lists usually involve different subjects. This results in each list having one or two idiosyncratic words which detract from the eidometer's overall performance. Secondly, the precision of the Mk. IV is far below that which could conceivably be achieved due to having to rely on a single cognitive judgment.

Several attempts have been made to relieve eidometer lists from the dependence on individual subjects [4; 9; and 11]. That made by Pearson and Hrabe [9] was successful but not practical because of excessive amounts of time required by each subject. The method of list design using cluster analysis reported by Peebles and Pearson [11] was not successful because of inhomogeneity of each list. Rank correlation coefficients were lower than those yielded by Mk. IV design techniques for all methods of cluster analysis examined.

The rank averaging method first suggested by Pearson [11] and investigated by Howell [4] was the first practical breakthru in eliminating individual idiosyncracies from each of the eidometer lists. This method will be discussed in detail in [4]. It involves analyzing the data from a large number of individual sorting experiments. In [4] Howell used forty sets involving seventy such individual experiments. The data from these experiments are then analyzed in such a way as to assign an average rank for each word. The new lists are constructed by taking the 8 words with the lowest average ranks for the first list, etc. Howell has been able to achieve a rank correlation coefficient of better than $r = 0.95$ using this method. This forms the heart of the new design techniques used for the Mk. V which is currently in the development and design (D&D) phase [8].

A new eidometer design, employing the semiotic equivalent of a vernier scale, has been conceived. This design requires for its construction not lists, but a large number (upwards of 1000) of individual words, all measured as precisely as possible by available techniques. These words are displayed in a two-dimensional array in such a way that an individual specimen word may be measured by interpolation both vertically and horizontally at a large number of points (hopefully, exactly ten) and the final result obtained by averaging the independent interpolations, thus eliminating the single cognitive judgment relied on heretofore. This is the design concept for the Mk. VI, which will be the basis for our attack on the redundancy curve for natural language. We will use the Mk. V to measure the eidontic deviance of the individual words required for the Mk. VI design. Our performance goals for the Mk. VI are 6 bpm precision, 5 bpm repeatability, and 2 bpm validity.

Once the Mk. V has been calibrated and its performance evaluated and assessed, the data accumulated in this process as well as all of the knowledge gained in experiments with the Mk. III and IV will be ennucleated into a redefinition of the eidontic deviance concept. The present scale is called the 'deviance' scale (symbolized by °ED), and the new scale will be called the deviometer scale (symbolized by dev). Of course, the new scale will be reversed with respect to the old one with deviometer magnitudes increasing with respect to strangeness. But the tie between Markov generators and scale values will also be broken by abandoning the use of approximate average Markov numbers for nominal scale values. These will be chosen instead on the basis of practical measurement and data recording considerations and also current measurement range, precision, and repeatability limits, all as determined by experience on the Mks. III, IV, and V.

The current deviance scale has an ordinal measurement structure guaranteed by the results of thousands of word sorting and list sorting experiments, but it lacks both an empirically meaningful natural origin and an empirically meaningful interval concept. By improving the validation of the Law of Word Interpretation troth by replicating the Miller, Bruner, Postman results with the Mk. VI eidometer; by multiple, partial, and canonical correlation analysis of the independence of the various empirical phenomena which enter into the law and thereby attempting to understand the nature of any cause and effect relationship present; and by precise measurement of the redundancy curve for natural language, also with the Mk. VI, an empirically meaningful interval concept will be established. This will allow the later development of an interval scale of measurement.

Various clues have already been detected as to where an empirically meaningful natural origin might lie, but these will be discussed under applications to the development of experiments between shape and other semiotic variables. If these clues bear fruit or if by other means an empirically meaningful natural origin can be found, then this would serve as the basis for the development of a ratio, scalar, or similarity measurement scale. If this eventuates, eidontic deviance would then be the first informational or semiotic measurement scale for which the standard tools of dimensional analysis used in the physical sciences are applicable, representing indeed a great breakthru in semiotics and information science.

The next application of eidometry to be discussed concerns the development of experiments between shape and other components of the sign. The concept of shape has played an intuitive, if unexpressed, role in a great number of psychological, especially psycholinguistic, experiments involving a variety of other semiotic concepts [6]. It has even reached the point of overt expression at the semiformulated level of conception in such experiments as Miller, Bruner, and Postman [5], and others. This has served to establish a great potential for designing experiments to explicitly examine the relation between shape and the other components of the sign. A list of experiments in which eidontic deviance enters as a principle parameter and preliminary design for each such experiment now numbers over a hundred. Only one of these, which could serve to examine the set of important links between syntactic sign components, will be mentioned here.

Terwilliger has examined the relationship between the associative meaningfulness and the perceptibility of the medium of a sign [16]. He confounded shape with meaningfulness, and because he failed to recognize this explicitly didn't analyze for the effects of confounding. The eidometer gives us the ability to unconfound this relationship and to design experiments to examine it explicitly. The associative meaningfulness is not a true measure of meaning but measures the tagmatic associative strength of a word which is therefore related to its semiotic context. Thus a redesign of this experiment using Terwilliger's Paradigm will give us the ability to examine the relationships between the medium, shape, and semiotic context of signs.

Terwilliger caused the medium of the sign, the visual carrier, to flicker on and off, and measured a parameter called CFF which is indicative of ability to perceive the flicker. He also measured the number of signs each S could associate to the specimen sign in a fixed time span, called associative meaningfulness. He found that CFF varied as a simple and regular function of associative meaningfulness. His specimen signs were real words of the American language which also varied in their familiarity, frequency, extensionality, tagmatic word category (part of speech), concreteness, abstractness, imageability, emotive responsiveness, and shape. Hence each of these variables, and very likely many others, were confounded with CFF and associative meaningfulness; also many of these variables themselves vary with shape, so that the confounding is leveraged.

Does meaningfulness cause CFF to vary or does strangeness cause it and meaningfulness both to vary? The way toward answering this question lies in eliminating as many of these other parameters as possible in the design of the experiment. By using artificial words as the experimental specimens familiarity, frequency, extensionality, tagmatic word category, concreteness, and abstractness, are eliminated totally; and imageability and emotive responsiveness are greatly reduced as variables. By measuring eidontic deviance as well as CFF and associative meaningfulness and following a multiple regression analysis with a partial regression analysis we can then use an inferential statistical technique called path analysis to analyze the cause and effect relationships between shape and the other two syntactic components.

In order to carry out such an experiment three instruments are required: 1) eidometer; 2) chronometer; and 3) teescope. This is essentially the same equipment as is required for the Miller, Bruner, Postman Experiment, except that the required teescope is somewhat more complex. It must be able to operate in repetitive mode at cycles as short as 5 ms. and must possess the requisite programming controls to operate in this fashion. The same measurements may also be made using a flicker photometer, but not as accurately, nor as flexibly. The teescope for the Miller, Bruner, Postman Experiment is not required to operate in repetitive mode and the shortest cycle time so far used has been 12 ms. (the lower limit of the Polymetric 0590). However, in investigations of the relation between length and strangeness in which shorter specimen words would have to be used, this lower limit would have to be reduced.

As my final example of semiotic applications of eidontics, I will discuss attempts to explore the basic nature of shape itself.

No direct experiments have been made as yet to explore the basic nature of shape itself; however, many clues have already been observed and will lead to the design of specific experiments to understand shape. Analyses of every calibration experiment show that when eidontic deviance measurements are averaged by word length, the highest values (least strangeness) are assigned to words of four and five letters in length. Both shorter and longer words are stranger on the average. Both $C|l-a|^{-1}$ and $C(\lg|l-a|)^{-1}$ are candidate functions for describing this effect, where C, and a are constants and l is word length. Notice that the average word length for written American is 4.56 lpw, which suggests that the constant a may eventually be identified with the average word length.

Miller, Bruner, and Postman originally ran their experiment only for eight-letter words and we have replicated it only for this one length. We want to explore the effect of length on interpretation error rate in the belief that this will lead to a factoring of the eidontic deviance concept into the product of two factors; one concerned with length, and one concerned with pattern. I suspect, but have no proof at present, that the pattern effect will be of such a nature as to be modeled by one of the many versions of Kolmogorov's algorithmic information measure, and the length effect will be found to be modeled by a logarithmic function.

In a large number of experiments of various kinds using the eidometer, an unusual but pervasive phenomena has been observed. This phenomena has taken on many different manifestations depending on the experimental paradigm and the individual subject, but it always occurs at about the same value of eidontic deviance (about 0.80 on the ED scale as meansured by a Mk. IV eidometer). This point has come to be called "the onset of eidension" by those of us actively using the eidometer. It has been characterized by subjects as that "point where pattern begins to make a difference", "where meaning begins to set in", "where funny things happen", and "where the words dance". It would be very interesting to design an experiment to systematically determine if this is indeed the same point for all subjects, whether it varies from subject to subject, or whether it even exists for all subjects. Such an experiment could also be designed to determine if this is the point where length and pattern separate out as aspects of shape.

One paradigm that appears promising for investigating the onset of eidension is an eye dwell-time, or novelty choice experiment. Eye dwell-time is considered in many paradigms of experimental psychology to be a measure of novelty. It has been hypothesized that if two artificial words are both only slightly strange but one is slightly more strange than the other, the stranger one will be regarded as the more novel — i.e., will receive the larger amount of eye dwell-time; whereas if they are both extremely strange and one is slightly more strange than the other, the less strange one will be regarded as the more novel and will receive the larger amount of eye-dwell time. If this hypothesis holds up, an experiment can be designed, using an eye dwell-time meter and an eidometer, which will find the eidontic deviance value at which these two effects cross over.

Because of the pervasiveness of the "onset of eidension" phenomena, there are likely to be a great number of experimental paradigms by which it can be isolated. It may be possible to use something like the onset of eidension for the empirically meaningful natural origin for the eidontic deviance scale, thus leading to another breakthru in experimental technology for the semiotic measurement of shape. And finally the onset of eidension may provide just the tool we need for developing theories of how our minds process the shape of the sign.

IX. References

1. Cattell, J. McK. *"Ueber die Zeit der Erkennung und Benennung von Schriftzeichen, Bildern und Farben,"* Philos. Stud., 2(1885), p635-650.

2. Flowers, J. "Progress in Semiotics Software at Georgia Tech's SemLab". The Semiotic Scene, 1(1976).

3. Flowers, J. "Software Development Progress at Georgia Tech's SemLab". Foundations, FIS-2(1979), p29. A SIG Technical Publication Series of ASIS.

4. Howell, D.P; and Pearson, C. "A New Technique for Eidometer Construction", Presented to the Fourth Annual Conference of the Semiotic Society of America; Bloomington, Indiana; October, 1979.

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

6. Pearson, C. "Towards an Empirical Foundation of Meaning". Ph.D. Thesis, Georgia Institute of Technology, 1977. University Microfilms.

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

8. Pearson, C. "Performance Evaluation of the Mk. V Eidometer", Presented to the Fourth Annual Conference of the Semiotic Society of America; Bloomington, Indiana; October, 1979.

9. Pearson & Hrabe, "Classification of Artificial Words by Strangeness of Shape", unpublished notes.

10. Pearson, C; and Slamecka, V. "A Theory of Sign Structure". Semiotic Scene, 1(1977), p1-22.

11. Peebles, J; and Pearson, C. "A Cluster Analysis Method for Constructing Eidometer Lists", Presented to the Second Annual Conference of the Semiotic Society of America; Denver, October 1977.

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

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

14. Sperling, G. Psychol. Monogr., 74(1960), p1.

15. Sperling, G. Acta Psychol., 27(1977), p285.

16. Terwilliger, R.F. Meaning and Mind. Oxford U.P., 1968.

X. Acknowledgements:

The research reported herein was supported by grant no. GN 40952 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.

A New Technique for Eidometer Construction

by David P. Howell
Columbia Systems Division
National Cash Register Corporation

ABSTRACT

This paper deals with a new technique for improving the individual word lists used in constructing an eidometer, an instrument used for measuring eidontic deviance, or the strangeness of the shape of work tokens [1].

The development of a practical version of any instrument requires many generations of refinements to eliminate idiosyncracies and improve precision and reliability. One of the last remaining idiosyncracies of the eidometer stemmed from the fact that in previous versions (Mk I thru Mk IV), each standard list of words of a given strangeness was constructed by a single subject even tho different lists were usually constructed by different subjects. This resulted in most lists having one or two words which looked obviously out of place and which in turn tended to cause difficulty in using the lists for measurement, thereby decreasing both the precision and reliability. The problem of eliminating this idiosyncracy was first addressed by Peebles and Pearson in [3]. The method employed by them was not successful, however, and later Pearson suggested the procedure adopted in this paper.

This new technique, called 'Rank Averaging', was a significant improvement. A measure of list performance used in preliminary design of eidometers is the rank coefficient between the average Markov value of the information generator of each word on the list and the visual ranking of the lists by naive subjects. Average values of r obtained by previous design methods was 0.87; Hrabe and Pearson suggested a more complex method of sorting, which yielded 0.92; the method of Peebles and Pearson yielded only 0.88; the rank averaging method used in this paper yielded 0.95. The improvement was significant at the $\alpha = .05$ level. The performance of the Mk V eidometer using this new design technique is reported by Pearson in [2].

The rank averaging technique involves sorting 128 artificial words into sixteen groups based on strangeness of shape. Independent trials by twenty subjects produced twenty sets of lists. Forty additional subjects were used to sort the resulting lists, two for each set. Each of the 128 words was thus associated with forty ranks which were then averaged. The words were then sorted by these average ranks and sixteen new lists were produced.

To evaluate the adequacy of the new method of list construction, ten new subjects were asked to visually rank the final sixteen lists by strangeness of word shape, giving $r = 0.95$. Analysis of the list composition was made to determine the words which were most difficult for subjects to rank and suggestions for possible further improvements to the word lists are made.

References

1. Pearson, C. "A Quantitative Concept of Word Shape". Semiotic Society of America's Second Annual Conference, Denver; October, 1977.

2. Pearson, C. "Calibration Performance of the Mk V EIdometer". Semiotic Society of America's Fourth Annual Conference, Bloomington; October, 1979.

3. Peebles, J; and Pearson, C. "A New Method of List Sorting to Eliminate Eidometer Idiosyncracies". Semiotic Society of America's Second Annual Conference, Denver; October, 1977.

PERFORMANCE EVALUATION OF THE MK. V EIDOMETER

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

ABSTRACT

The invention of the eidometer and its use in measuring characteristics of word shape was reported in 1977 [2]. It played a key role in the investigation of the relation between sign shape and interpretability which led to the announcement of a new law of information: an empirical regularity between word shapes and their interpretation [3]. The author called this the Law of Word Interpretation and described it as an empirical explication of the Miller-Bruner-Postman Effect [1].

Finally, in investigating deviations from the Law of Interpretation, a systematic second order correction term was found that involved an information-like function $F = \frac{1}{8}\sum_{i=1}^{8} \lg f_i$, that describes a semiotic process of immediate memory, where the f_i are the occurrence frequencies of the letters composing the shape of the sign [4]. The experiments required for isolating this function reached the limits of reliability and precision of the Mk. IV eidometer.

Therefore, in order to apply the eidometer to its intended purpose of measuring Shannon's Redundancy Function for Natural Language [6], it was necessary to improve the eidometer design, eliminating several second order idiosyncracies of the Mk. IV, and to increase its reliability and precision. Experiments to find a way to eliminate the idiosyncracies were reported by Peebles and Pearson [5]. The method adopted, however, was developed by Pearson & Howell and is reproted in this paper for the first time.

The Mk. V eidometer was built and its performance calibrated against a standard set of 100 artificial words spanning the entire deviance range from 0.0 to 3.0°ED. The reliability and precision results of these calibration experiments are reported here along with the design techniques used in construction of the Mk. V.

The definition of a new eidometer scale is proposed to take advantage of all the improvements in eidometry techniques and the increase in measurement structure that has been achieved since the present scale was defined several years ago. The new unit is called the 'deviometer' (dev.) in contrast to the old unit which was called the 'eidontic deviance unit' (°ED).

References

1. Miller, G.A.; Bruner, J.S.; and Postman, L. "Familiarity of Letter Sequences and Tachistoscopic Identification". Jour. General Psychol., 50(1954), p129-139.

2. Pearson, C. "An Objective Concept of Word Shape for Natural Language". Presented to the Second Annual Conference of the Semiotic Society of America; Denver; October 1977.

3. 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, in New York; November, 1978.

4. Pearson, C. "Information Decay in Immediate Memory: A Second Order Correction to the Law of Interpretation". Presented to the Second International Congress of Semiotics, in Vienna; July, 1979.

5. Peebles, J; and Pearson, C. "A Cluster Analysis Method of Constructing Eidometer Lists". Presented to the Second Annual Conference of the Semiotic Society of America, in Denver; October 1977.

6. Shannon, C.E. "Prediction and Entropy of Printed English". Bell Syst. Tech. Jour., 30(1951), p50-64.

THE ATTRIBUTES OF INFORMATION

by Charls Pearson

ABSTRACT

In order to intelligently investigate the attributes of information, we must first ask what kind of entity is information, that it can have attributes, and what are the allowable logical modes for its attributes. And in order to make this question precise in its turn, we have focused on the atomic units of information and/or its carriers, the sign. For the one invariable principle of all information science is that information always accompanies and is accompanied by a sign process.

We therefore turn to semiotics for a study of the basic structure of information and information processes. At the Georgia Tech SemLab, our efforts have concentrated on empirical investigations (both theoretical and experimental) of the basic structure of signs.

The paper presents a summary of the Universal Sign Structure Theory and uses this to motivate a discussion of the attributes of information, especially those information attributes associated with the shape of the sign and the information measurements being made in Georgia Tech's SemLab.

THE ECHELON COUNTER: A NEW INSTRUMENT FOR MEASURING THE VOCABULARY GROWTH RATE AND THE TYPE-TOKEN RELATIONSHIP

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

Abstract. All type-token measurements require the use of random samples of text. But current measurement methods ignore the statistical structure of these samples and hence most of the type-token information present is lost in the measurement. This paper announces a new instrument which uses this statistical structure to yield increased precision and decreased measurement noise. The Echelon Counter uses the fact that for truly random language samples it makes no difference where the type-token measurement begins as long as the counting proceeds sequentially and the sample is counted completely one time. All the type-token information in the sample can be obtained by making n independent measurements of the relationship starting with e ch of the n word-tokens in the sample in turn.

The advantage of the new instrument lies in an algorithm which makes all n measurements simultaneously, with no more work than required for a single measurement. The calculation time increases only as n rather than the expected n^2. The precision is increased from the present ±1 wt. to $\pm 1/n$ wt. Noise suppression is great enough to make the previously unobservable vocabulary growth rate curve easily measurable.

INTRODUCTION

A constellation of different relationships and information processes are 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 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 such as Lotka's Law and Bradford's Law are closely related to the Type-Token Constellation.

In (1) Pearson recommended several advantages in measuring the type-token relationship rather than the rank-frequency relationship, when attempting to investigate the Type-Token Constellation empirically, and noted the need for a new measuring instrument with finer precision and less measurement noise. This paper reports on the invention of such an instrument.

THEORETICAL DEVELOPMENT

What is the likelihood of the first two word-tokens in any subsample of a given sample being two different word-types? In other words, how many adjacent word-token pairs instantiate the same word-type? Present methods of measuring the type-token relationship use the information available from only one pair of adjacent word pairs in the entire sample, namely the first two. In actuality, information is available to the measurement process from n word pairs where n is the number of word-tokens in the entire sample. Thus most of the type-token information in the sample is simply wasted or ignored by present measurement methods. Similarly, the measurement of the type-token relationship for three word-tokens involves measuring the likelihood of the third word token instantiating a new word-type. Again there are n independent measurements of this available on samples of size n wk. and present methods ignore all but one of these measurements.

By carrying the above reasoning to its logical conclusion we may infer that if a sample of language text is truly random then it should make no difference where in the sample the type-token measurement is begun as long as the counting proceeds in contiguous (grammatical order) sequence and the whole sample is counted completely one time. If the counting is initiated at any place other than the first word of the sample, this requires counting to the end of the sample and then transferring to the first word of the sample and continuing on to the point where counting started. This conclusion was already hinted at in (1).

The new instrument, the Echelon Counter, uses this conclusion to take advantage of all the type-token information available in the sample by carrying out n independent measurements of the type-token relationship starting with each of the n word-tokens in the sample in turn and averaging to get the final measurement result. The advantage of the new instrument lies in a mathematical process called the 'Echelon Transform' which allows all n measurements to be made simultaneously with no more work than a single measurement and systematizes the final averaging calculations. The computational algorithm involves a 'next-token-of-the-same-type' vector (NEXT) of size 2 by n and an Echelon Matrix of size n by n. The Echelon Matrix is calculated by a trivial process involving the next-token-of-the-same type vector and the time of calculation increases only as n rather than the expected n^2.

The precision of present type-token measurement methods is ±1 wt. The new method yields a precision of $\pm 1/n$ wt. Thus for samples even as small as 1000 wk. the precision is increased sufficiently to reduce almost all the noise that contaminates statistical analyses. In fact, noise suppression is great enough to make the vocabulary growth rate curve, which was not measurable at all with previous methods, easily measurable.

All equations will be shown in the next section in their explicit form as developed for

Pearson – 1

the prototype instrument.

PROTOTYPE INSTRUMENT

A manual version of the Echelon Counter was implemented in order to test out the feasibility of a computer based version. Because of the effort involved in manual counting procedures, the test sample was limited to 50 words. Since this was predicted to yield a 50-fold increase in precision, however, it was sufficient to test out the applicability of the instrument. The test sample is given in Table 1. The next token-of-the-same type vector is implemented

Table 1: Test Sample for Echelon Counter
The little dog chased the big cat up a blind alley. They were playing tag and the dog was "it". At the end of the alley the dog tagged the cat and they immediately reversed roles. The big cat chased the little dog all the way back to their home.

in the prototype as a previous-token-of-the-same type vector, called PREV and given in Table 2.

Table 2: PREV: The previous-token-of-the-same-type vector.

i	PREV(i)	WORD	i	PREV(i)	WORD
1	45	THE	26	11	ALLEY
2	42	LITTLE	27	25	THE
3	43	DOG	28	18	DOG
4	40	CHASED	29	29	TAGGED
5	1	THE	30	27	THE
6	38	BIG	31	7	CAT
7	39	CAT	32	16	AND
8	8	UP	33	12	THEY
9	9	A	34	34	IMMEDIATELY
10	10	BLIND	35	35	REVERSED
11	26	ALLEY	36	36	ROLES
12	33	THEY	37	30	THE
13	13	WERE	38	6	BIG
14	14	PLAYING	39	31	CAT
15	15	TAG	40	4	CHASED
16	32	AND	41	37	THE
17	5	THE	42	2	LITTLE
18	3	DOG	43	28	DOG
19	19	WAS	44	44	ALL
20	20	"IT"	45	41	THE
21	21	AT	46	46	WAY
22	17	THE	47	47	BARK
23	23	END	48	48	TO
24	24	OF	49	49	THEIR
25	22	THE	50	50	HOME

PREV(i) lists the previous token of the same type; for instance token 5 is THE and the previous token of this same type is token 1, and we therefore find a 1 shown for PREV(5). All integers are to be interpreted modulo 50. Therefore we may interpret i=1 as i=51 in order to determine the previous occurrence of THE, which is i=45, and we therefore find 45 shown for PREV(1). Once the PREV vector is constructed, all measurements have been completed. Note, in this example we have just made 50 independent measurements of the complete type-token relationship for the 50 word sample. The calculations are now performed most effectively by expanding the vector to a Boolean matrix, called the Echelon Matrix, in which the one's stand for words which are to be counted as new types and the zero's stand for words which are to be counted as repeat tokens of previous types. The matrix and the vector are completely equivalent to each other and each can be fully generated from knowledge of the other. The u.l.h. 15 by 15 quadrant of the 50 by 50 Echelon Matrix for the manual example is shown in Table 3.

Table 3: Upper Left Hand 15 by 15 Quadrant of the 50 by 50 Echelon Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
5	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0
6	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
7	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
11	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
12	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0
13	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

For vertical column 1, $M_{1,1} = 1$, $M_{i,1} = 0$ iff $1 \leq \text{PREV}(i) \leq i-1$; otherwise, $M_{i,1} = 1$. for j > 1, set $M_{i,j} = M_{i,j-1}$, then set $M_{j-1,j} = 0$, and $M_{\text{NEXT}(j-1),j} = 1$. Repeat for each value of j. It is thus seen that the matrix is filled in by a trivial copying process except for the two values $M_{j-1,j}$, and $M_{\text{NEXT}(j-1),j}$, only the latter of which depends on PREV. When the matrix is full, the final measured averages can be calculated by

$$\text{VGR}(i) = \frac{1}{50} \sum_{j=1}^{i-1} M_{j,j-i+1}$$

$$T_i = \sum_{k-1}^{i} \text{VGR}(k)$$

$$\text{TTR}(i) = T_i / i$$

RESULTS

The type-token relation for this 50 word sample as measured by the Echelon Counter is shown in fig. 1.

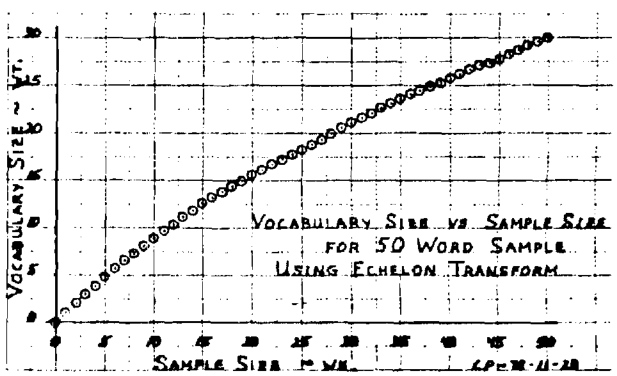


Figure 1. Type-Token Relation for 50 word sample as measured by the Echelon counter.

For comparison, the type-token relation for the same sample as measured by classical counting methods is shown in fig. 2.

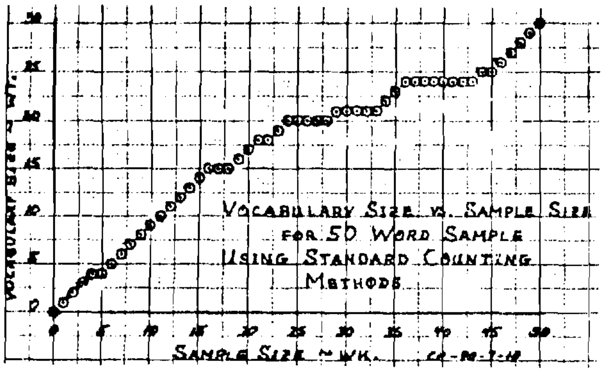


Figure 2. Type-token Relation for 50 word sample as measured with standard counting methods.

The type-token ratio and vocabulary growth rate relations as measured by the Echelon Counter are shown in fig. 3.

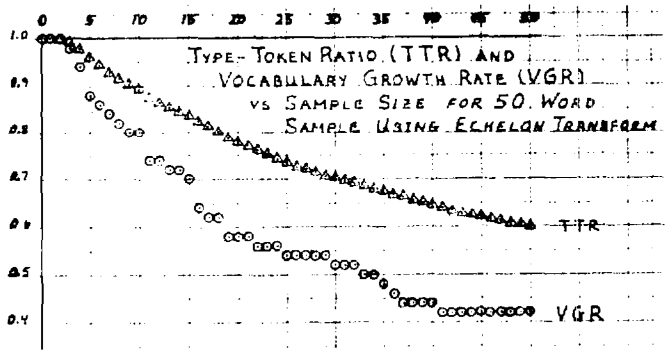


Figure 3. The Type-Token Ratio and Vocabulary Growth Rate Relations for a 50 Word Sample as Measured by the Echelon Counter.

For comparison, the type-token ratio and vocabulary growth rate relations as measured with classical counting methods are shown in fig. 4.

From the above results it can be seen that the 50-fold increase in measurement precision and reduction of measurement noise has been achieved. The case of the Vocabulary Growth Rate curve is especially dramatic. Measuring it with classical counting methods the underlying curve cannot be seen at all because of the measurement noise (fig. 4). Measuring it with the Echelon Counter, however, the curve is readily seen (fig. 3). It can be seen that a sample of size 500 should suffice for producing highly precise, low noise measurements with the Echelon Counter, suitable for most statistical analysis techniques.

DISCUSSION

The Echelon Counter solves a long-standing problem in developing a phenomenological theory of the Type-Token Constellation. A phenomenological theory is a systematization of the known relationships from which each of

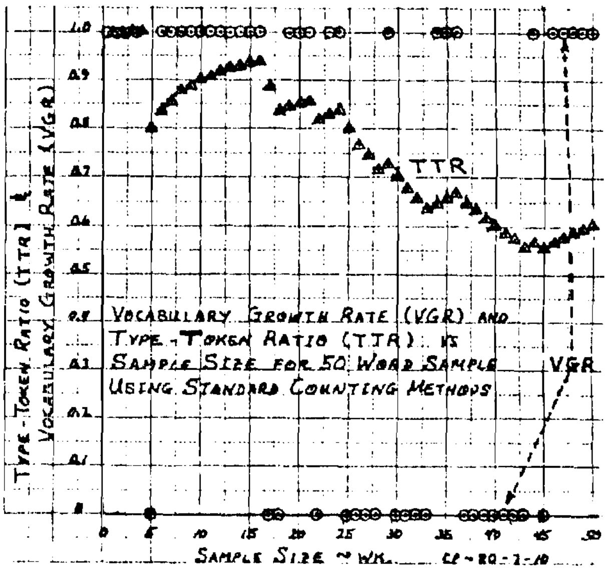


Figure 4. The Type-Token Ratio and Vocabulary Growth Rate Relations for a 50 Word Sample as Measured by Standard Counting Methods.

the original relationships can be recaptured. Such a theory must be based on the Vocabulary Growth Rate relation for reasons of statistical rigor.

CONCLUSIONS

The Echelon Counter is a practical and feasible instrument for measuring type-token data. It has extremely high precision and low measurement noise, on the order of $\pm 1/n$ wt. compared to the ± 1 wt. for classical counting methods. This breakthru allows the solution of two other problems which together with the measurement problem prevented the development of a phenomenological theory, or general law, of the Type-Token Constellation. The general law, when available, should relate practical information engineering relations such as Bradford's Law and Lotka's Law to the Type-Token Constellation, and by throwing more insight on these relationships, make them ultimately more useful to the information engineering community. This phenomenological theory should now be developed.

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. John Gehl, Larry Heilprin, Phil Siegmann, Bill Underwood, and Pranas Zunde have all listened to various attempts to explain how the Echelon Counter works.

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.

THE BASIC CONCEPT OF THE SIGN

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

Abstract. Information Science is a basic natural science which can be identified with empirical semiotics, the paradigm for the trinary natural sciences. The Universal Sign Structure Theory provides a first theoretical basis upon which much future experimental, mathematical, and theoretical development can be founded. The basic concept for the foundation of the entire field of Information Science is that of the 'sign'. Future research in Information Science should be concentrated on experimental, mathematical, and theoretical investigations of the structure of signs, systems of signs, and the processing of signs.

The paper provides a sketch of the metaphysical foundations of Information Science based on an atomistic approach to the carriers of information. Next the Universal Sign Structure Theory is sketched, based on the concept of the 'sign'. Finally the definition of all concepts of Information Science in terms of the concept of 'sign' and the other concepts of the Universal Sign Structure Theory is adumbrated.

INTRODUCTION

Information is a semiotic phenomena rather than a physical phenomena. To understand this statement requires a brief excursion into the metaphysics of science which I carry out in this introductory section.

Physics is the paradigm binary science. This means that of all the sciences which study the binary properties of nature, physics is the most fundamental and provides the basic understanding for all the others. Binary properties of nature are those whose mathematical models require binary relations such as 'heavier than', 'longer than', 'faster than', etc. for their systematic rigorous study. Thus the binary sciences are all those which study those properties of nature which can be modeled mathematically by binary relations. Similarly, semiotics is the paradigm trinary science. It is the most fundamental of all the sciences which study the trinary properties of nature and provides the basic understanding for all of them. Trinary properties of nature are those whose mathematical models require trinary relations such as 'means — to', 'informs — that', 'more abstract than — for', 'more beautiful — for', etc. for their systematic rigorous study. Thus the trinary sciences are all those, such as linguistics, psychology, information science, etc. which study those properties of nature which can be modeled mathematically by trinary relations.

There are two questions that immediately arise in regard to such a metaphysics as this. Is the distinction between binary and trinary sciences redundant? and is it sufficient? The distinction would be redundant if all trinary relations could be factored into products of binary relations and thus their content studied by physical means. The distinction would be insufficient on the other hand if nature required mathematical relations of quarternary or higher order and at least one of the required higher order relations could not be factored into products of binary and trinary relations. Charles Peirce, the American logician and father of Information Science, answered both of these questions simultaneously in the 1870's with his logic of n-ary relations. There do exist some trinary relations which can not be factored into any number of products of binary relations, but all quarternary and higher order relations can be factored into products of binary and trinary relations. He also found a relation that plays a fundamental role for relational logic similar to that played by the imaginary unit i in algebra, and he called this relation a 'sign'. The sign relation can be expressed, altho somewhat inadequately, in natural language by "x interprets y as standing for z".

We may now return from the metaphysical detour and attempt to understand the import of the first sentence of the paper. I am claiming that information, as a basic concept, involves the notion of a sign and requires trinary relations for its mathematical models. Information is best understood as being conveyed by some particular originator via some particular message about some particular object. In other words, x uses y to inform about z. This has the logical structure which prevents factoring into products of binary relations and is already very close to the structure of the sign concept itself.

In fact I have many times identified information science with semiotics as constituting the same basic empirical science. However, since present-day semiotics is best known for its speculative, non-empirical adherents, while Information Science concentrates almost exclusively on its technological nonbasic engineering aspects, I have most often framed this identification by means of the aphorism: IS^3 which simply means:

Information Science IS Instrumentation + Semiotics

In semiotics, a process is involved either actually or at least potentially and this process can be called 'communication' in the broadest possible sense. Information is the structural aspects of this semiotic process and as such it provides us with the fundamental structural tool of investigation not only for basic semiotics, but for all the semiotic sciences as well.

Information is conveyed by a carrier which has a physical component. This carrier is called a 'MESSAGE' in a very broad sense. However, the information carrier, or message, can be a very complex kind of entity and to study it effectively, it must be conceptionally divided down into its indivisible atomic constituents. The atomic carriers of information are called 'SIGNS'. Signs thus form the most basic concept of information science.

Information is thus carried in messages which are systems of one or more signs. In written alphabetic languages the system is a string; therefore a message is a string of one or more sign. However, in a painting, the message is a two-dimensional structure of one or more signs and in a piece of sculpture or a work of architecture the message is a three-dimensional structure of one or more signs.

This metaphysics of structural atomism allows the development of a very powerful theoretical basis for information science called the Universal Sign Structure Theory.

UNIVERSAL SIGN STRUCTURE THEORY

The Universal Sign Structure Theory is composed of several primitive concepts, a relational model, three principles, and several obvious rules of interpretation or translation between the theoretical vocabulary and the observational (or less theoretical) vocabulary. The most important primitive concept is that of 'sign', which is not defined in the theory. Other important primitive concepts are 'consists of', 'stands for', and 'is interpreted (by) (within)'. The theory is tied to the empirical reality of classical semiotics and Information Science via nine representation theorems. By means of the terms, principles, model, and theorems of my Universal Sign Structure Theory, all concepts of Information Science may be defined.

The relational model is called the Universal Sign Structure Model. Altho it is not shown here because of space restrictions, it is readily available, for instance in (1,2, and 3). Since one of the main purposes of the theory is to explain the interrelationships between various concepts of meaning, various theoretical and observable aspects of signs, and various information concepts, these concepts and aspects are treated by the model as components and appear in the model as the nodes of a directed graph. Some of the relationships between the concepts and aspects then appear as the arrows of the model. These will be discussed in more detail later.

The three principles of the theory are stated as follows:

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'.

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.

The study of the physical, or binary, sciences can be neatly divided between theory and observation with the theoretical components of these sciences containing only theoretical concepts. However, the semiotic sciences cannot be approached in such a simple manner, perhaps because of the trinary structure of the sign, and we find the basic theoretical concept of the sign itself composed of both theoretical and observational components. The theoretical concepts are the internal components and the observational concepts are the external components. The totality of the internal components with their relationships makes up the INTERNAL STRUCTURE and the totality of the external components with their relationships makes up the EXTERNAL STRUCTURE.

We thus see that concepts of meaning are treated by the theory as theoretical concepts and generators of information are associated with observational concepts, as are the various information measures themselves. Note that the term 'INFORMATION MEASURE' is used here in the empirical sense and not in the mathematical sense. A mathematical information function can be, but need not necessarily be, a model of an empirical information measure.

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 the Universal Sign Structure Model), and for each dimension independently.*

The Peircean Taxonomy of signs can be explained by the theory by means of nine representation theorems which are derived from the three principles and the Universal Sign Structure Model.

For instance, from (3,p8) we have theorem 3: A sign is a token iff it has all three levels of syntactic structure. It therefore has three components of syntactic meaning (tagmension, eidension, and ontosion) and three external syntactic components (the semiotic context, the shape of the sign, and the medium in which it is embodied). The other eight theorems and all proofs are given in (3).

Other terms which are useful in using the theory, and which may be defined in terms of the primitive concepts of the theory are: dimension, meaning, information, and levels. A DIMENSION is a maximal, independent, substructure of the sign. The three dimensions of the sign are 1) syntactic; 2) semantic; and 3) pragmatic. MEANING is a theoretical subconcept of the sign and thus every possible sign has at least three components of meaning while dolemic symbolic tokens have all nine components. Meaning includes such components as

TAGMENSION, EXTENSION, INTENSION, CONTENSION, and PURPOSION.

INFORMATION is any observable, or measurable, aspect of the sign and is generated by the information generators or external components of the sign. Each of the nine information generators generates an infinite number of information measures. For instance shape, an external component or information generator, generates LENGTH, PATTERN, COMPLEXITY, STRANGENESS, etc. -- all information measures. In some situations length can be modeled mathematically by Shannon's selective information function, and complexity can be modeled by Kolmogorov's algorithmic information function.

LEVEL refers to a combination of one meaning component and its associated information generator. Thus a token has three levels of syntactic structure while a type has only two levels of syntactic structure.

ADDITIONAL DEFINITIONS

The above basic theory allows the definition of all remaining terms of information science. Lack of space precludes even an attempt to do this here; however, a few of these will be given to show the flavor of how this would go.

The terms 'type' and 'token' are rather important for the topic of syntactic information and occur in such studies as the Rank-Frequency Law of Zipf and Estoup, the Type-Token Relation, the Vocabulary Growth Rate, Lotka's Law, Bradford's Law, etc. 'Type' and 'Token' are two of a set of nine terms used in the Peircean taxonomy of signs which include 'tone', 'type', 'token', 'index', 'icon', 'symbol', 'rheme', 'pheme', and 'doleme'. These terms are defined in (3) from which the following two definitions are taken. A sign which exists as a general kind, both in itself and distinguishable from other signs is called a 'TYPE'. A sign which exists as an actual, single, physically existing individual is called a 'TOKEN'. Since 'symbol' is a pervasive term of information it will also be defined here. A sign which is related to its object by an arbitrary convention, agreement, or general law, is called a 'SYMBOL'. In order to understand this last definition we need the following: Let R be a sign represented mathematically as a trinary relation

$$R = \{<x,y,z> \mid <x,y,z> \in R\}$$

then any individual y is called an 'OBJECT' of R and the set of all $y \in R$ is called the 'EXTENSION' of R. The extension of R is one component of its meaning. In order to classify computers we need the following two additional definitions: A sign which is related to its object by an actual, single, existential cause and effect relation is called an 'INDEX'. A sign which is related to its object by a concrete similarity between the shape of the sign and its object is called an 'ICON'. For the definition of 'shape' see below. We may now define the three categories of computers as follows: A SIMULATOR is an index processor. An ANALOG COMPUTER is an icon processor. A DIGITAL COMPUTER is a symbol processor.

Shape and eidension make up the eidontic level of syntactic structure. In the Universal Sign Structure Theory the technical concept of shape is an explication of our nontechnical or intuitive concept of shape. SHAPE is that collection of attributes by which we distinguish one kind of sign from another kind of sign when both are embodied in the same medium within similar semiotic contexts. For instance, in the case of the letters X and O embodied in printer's ink on white paper, it is our nontechnical, intuitive concept of shape by which we distinguish them. Here the technical concept of shape matches our nontechnical concept of shape. In the case of printed words in an alphabetic language shape refers to orthographic shape and is determined by spelling. In the case of spoken words shape refers to phonemic shape and is determined by the phonemic patterns.

Letters and other CODE ELEMENTS are not signs, however; they are building blocks for constructing the shapes of signs. The technical term for code element is 'EIDEME'. Various kinds of eidemes include phonemes, graphemes, etc. The MEDIUM of a sign is the physical component from which it is embodied; printer's ink on paper, for instance, for printed text. The term 'SIGNAL' is used two different ways in Information Science, sometimes meaning the medium plus the shape of a sign, and sometimes meaning the medium plus the shape plus the semiotic context of a sign (in other words, the complete external syntactic structure of the sign). The vague term 'context' must be distinguished into three separate terms, 'SEMIOTIC CONTEXT', 'PHYSICAL CONTEXT', and 'SOCIAL/BEHAVIORAL CONTEXT'. Attributes of information are the various objectively observable attributes of the information generators, or external sign components, that enter into general empirical regularities; that is, they are the various measurable parameters that enter into the laws of information.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation under grant #IST-7827002. I would also like to thank my colleague Vladimir Slamecka for untiring support and encouragement. and my thesis advisor Pranas Zunde for his intellectual stimulation and challenge.

REFERENCES

(1) PEARSON, C. "Towards an Empirical Foundation of Meaning". Ph.D. Dissertation submitted to Georgia Institute of Technology, June 1977. University Microfilms, Ann Arbor, Michigan.

(2) PEARSON, C. "Attributes of Information". Paper presented at ASIS-79; Minneapolis; October, 1979.

(3) PEARSON, C; and SLAMECKA, V. "A Theory of Sign Structure". Semiotic Scene, 1(1977), #2, p1-22.

INFORMATION SCIENCE: THE CHALLENGE OF A BASIC SCIENCE

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

September 1980

Preliminary draft for comments only. Not for publication or quotation. All rights reserved by the author. Intended for presentation as an invited paper to the SIG/FIS session on "The Challenge of Inforamtion Science" at the 1980 Annual Conference of the ASIS in Anaheim; October 7, 1980.

INFORMATION SCIENCE: THE CHALLENGE OF A BASIC SCIENCE

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

ABSTRACT

The challenge of going basic is the challenge of meeting, solving, or enduring the many mundane problems engendered by the challenge to one's personal life, professional life, and career caused by deciding to concentrate on the search for a basic science. These challenges are engendered when one turns from the practical and accepted path of improving technology onto the path of searching for a basic science and can form the greatest impediment to achieving this goal. Most of these problems are nonsubstantive scientifically but all are extremely important from a practical standpoint. Due to the collective force of all of these mundane, but very practical, problems, there are several implications about how we fund research, reward workers, and select new candidates for entry into the field. These implications are generally contrary to present practices.

The paper lists many of these challenges in a very rudimentary systematization and discusses some of their interrelations. It discusses the challenge that each of these problems presents to us, with a very elementary analysis of each. Finally, the overall challenges and their implications are discussed and the conclusions presented.

INFORMATION SCIENCE: THE CHALLENGE OF A BASIC SCIENCE

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

1. INTRODUCTION:

The theme of this session is "The Challenge of Developing the Basic Science of Information Science". This assumes that there is a basic science to be developed and that information science is not just an unsystematic collection of practical techniques and nothing more. We hope by presenting this session to stimulate some of you to concentrate your efforts towards developing the foundations of a basic science of information science.

That it is not easy to discern, or distill, a basic science from among all the interests in practical techniques, or to turn from an interest in information technology towards concentrating on developing a basic science, is the message of this paper. The other authors in the session will discuss some of the specific and substantive challenges in doing this.

Professor Zunde will discuss the challenge of the foundations. Is the basic science we seek to develop an empirical science, or a formal science? Professor Barnes will discuss the basic challenge of information science. Is there a basic science to be developed or is there nothing but applications? Professor Studer will discuss the logical challenge of information science. Is the basic science a multidisciplinary study or a unidisciplinary science? A fifth paper was originally planned to discuss the methodological challenge of information science. Is the basic science experimental or theoretical? Unfortunately we could not find a willing author who felt competent to discuss the issues involved in this challenge.

With these substantive challenges being discussed by the other authors, my own paper will concentrate on a host of more mundane problems: those generated by the challenge to one's personal life, professional life, and career engendered by turning from the practical and accepted path of improving technology onto the path of searching for a basic science. These problems are much more mundane than the substantive issues discussed by the other authors. Scientifically, each of these problems is trivial in and of itself. However, the weight of these problems as they accumulate, like the weight of each of the grains of sand that together form a mighty mountain, can form the greatest impediment to achieving this goal.

Due to the collective force of all of these mundane, but very practical, problems, there are several implications about how we fund research, reward workers, and select new candidates for entry into the field. These implications are generally contrary to present practices.

In the next section, I will list many of these problems and attempt a very rudimentary systemization. These are neither exhaustive nor exclusive categories and so I discuss some of their interrelations. In section 3 I discuss the challenge that each of these problems presents us, with a very elementary analysis of each. Section 4 is a discussion of the overall challenges and their implications and section 5 is a summary of conclusions.

2. THE CHALLENGE OF GOING BASIC:

The challenge of going basic is the challenge of meeting, solving, or enduring the many mundane problems engendered by the challenge to one's personal life, professional life, and career caused by deciding to concentrate on the search for a basic science. Most of these problms are non-substantive scientifically but all are extremely important from a practical standpoint.

After thinking of more than 20 challenges to the basic, or pure, scientist as opposed to the applied scientist, technician, or engineer, I concluded that there may be no finite list of such problems and so I turned my thoughts in another direction, and attempted to systematize these problems or develop a schema that would characterize them. Only one observation worth mentioning came to mind on this subject. There appear to be two main kinds of challenges to going basic: 1) temptations; and 2) pressures.

What I call 'temptations' are challenges that originate from within the investigator himself, such as the temptation to relax and take it easy rather than exert himself performing some very difficult job which no one will ever notice or appreciate anyway. Pressures are those challenges that originate outside of the investigator from others that he must associate or interact with, such as the pressure from university administrators to publish voluminously, even when one has nothing important to say. I can offer one speculation to the effect that all of these challenges may arise in pairs, one temptation and one corresponding pressure. In fact, whether a challenge is regarded as a temptation or a pressure may be only a function of how the investigator views himself in relation to his research and in relation to others he must interact with. I make no attempt to analyze or answer this question. A good example of this is the challenge to earn a living. Whether this is a temptation on one's part to earn enough to live comfortably, or a pressure on the part of society to conform to its standards is a question of viewpoint. The challenge is real; the classification is debatable. In fact, many, if not all, of both the pressures and the temptations, originate as part of, or because of, the overall challenge to make a living. Whether or not all could be analyzed so as to fall under this one category goes beyond the scope of my present analysis. Since my challenges are overlapping and non-exhaustive I do no more than list them under the headings of temptations, and pressures. I do place 'earning a living' as first on each list.

Table 1. Challenges to the Basic Information Scientist

A.	Temptations (Internal, or self-induced challenges)
	1. Temptation to earn a living. 2. Temptation to turn out quick and easy results. 3. Temptation to be lazy, to relax, and to succumb to inertia. 4. Temptation to play it safe. 5. Temptation to act normal and "fit in". 6. Temptation to succumb to diversions. 7. Temptation to be distracted.
B.	Pressures (External, or socially induced challenges)
	1. Pressure to make a living. 2. Pressure exerted by prestigious and/or granting agencies. 3. Pressure exerted by academic peers, departmental pressures, etc. 4. Pressure exerted by professional peers, i.e., ASIS, etc. 5. Pressure to publish voluminously and often. 6. Pressure to be "practical", i.e., produce applied results rather than basic or fundamental results. 7. Pressure to be concrete/and or specific, i.e., to produce case studies rather than general inductive analyses or theoretical formulations. 8. Budgetary pressures. 9. Pressure to conform methodologically. 10. Pressure to research only "planable" areas. 11. Pressure to be friendly and sociable with students (and other faculty, too). 12. Pressures to do administrative chores. 13. Pressure to do safe research. 14. Pressure to be normal. 15. Pressure to justify results either before they have been achieved, or before their justification is evident. 16. Pressure to be understood by one's peers. 17. Pressure to do menial work. 18. Pressure to be non-esoteric. 19. Pressure to be a good spouse and/or parent. 20. Pressure of distractions.

3. <u>DISCUSSION</u>:

In order to eat, sleep in a warm shelter, and have clothes to wear, it is often necessary to earn a living. The better one likes to eat, dress, and shelter oneself, the more temptation this generates. One way to avoid this temptation is to be born rich or inherit wealth. Another way is Lawrence of Arabia's solution of simply "not caring". Another approach that has been tried by many is to postpone one's life goal of doing basic research and concentrate initially on getting rich so that one can then retire and devote himself exclusively to basic science. For a few this works, but for most who attempt this method it becomes a trap as either success breeds a desire or need for more success, or what is more likely in these days of inflation and economic maladjustment, one is never able to get ahead of his daily economic necessities. However, the approach most often tried involves time-sharing: earning a living <u>AND</u> doing basic research. As usually results, this approach does justice to neither. There is never enough time to earn a good living and never time enough left over to tackle basic research in a fundamental way. This, by the way, is the solution recommended by Einstein, a notable exception to my above generalization. He worked at the Swiss patent office while he developed his early theories and until he developed enough prestige to qualify for a faculty position on "his own terms". He advocated an even more dramatic separation of work and research. One should be a plumber or ditch digger by day and do basic research at night. Perhaps Einstein could have made his discoveries even this way, but most people simply get too tired after a day's labor to be productive in a second job that requires intellectual creativity. A final way, and one that cannot work in today's atmosphere generated by all of the pressures I will discuss, is to earn one's livelihood <u>BY</u> doing basic research.

The temptation to turn out quick and easy results is an internal solution to many of the external pressures, such as the pressure to publish voluminously, etc. which is made worse by the temptation to be lazy, etc. Yielding to this temptation precludes of course achieving any basic or fundamental results, which are never quick or easy.

One is often not even aware of the temptation to be lazy, relax, and/or succumb to inertia, since it is a passive temptation. One often yields to this temptation by exerting effort, but by failing to exert (or even recognizing the need for) the extra effort required to overcome some of the pressures and other temptations. Inertia is the insidious temptation in that it is always present, never conspicuous, and has drastic consequences. In conjunction with any other pressure, inertia can totally prevent any basic results. An example was given in the last paragraph in conjunction with pressures to publish voluminously. It can often result in the attitude: "don't start anything new when one can always fiddle a little more with some practical development". The only solution to such an insidious temptation is simply to demand success of oneself in every single research endeavor and never to compromise. This failure to compromise, however, greatly exacerbates almost all of the pressures to be discussed shortly.

But the challenge to demand success of oneself in every single endeavor increases one's susceptibility to another temptation, the temptation to play it safe. To play it safe is to work on a problem that is known to have an answer and whose answer is known to lie within the capabilities of the investigator, no matter how inconsequential that problem or its solution is to basic science. This challenge is to always be honest with oneself in determining one's own research goals and in setting one's research priorities in relation to one's own view of the strategic requirements of developing a basic science of information science.

The temptation to act normal and "fit in" may only be a temptation to yield to society's demand to be normal and conform. No one likes to be regarded as queer.* No matter how normal one regards himself, the fact that he knows other people consider him queer exerts a pressure and a concomitant temptation. It is also a fact that many people automatically regard basic researchers as being somehow queer or eccentric or social misfits. Again one must challenge oneself to be honest with oneself and to be himself in the face of his dedication to basic information science.

Diversions are a way of life for modern society. Most good researchers have a wide range of interests. One is readily tempted to go to the symphony tonight because Toscannini is making what will probably be the last guest appearance here of his life. One says, "I'll make just this one last exception and then settle down to work tomorrow", but the exceptions keep coming to the exclusion of *mañana*. This is related to the temptation to inertia and requires challenge to dedicate oneself very narrowly to a life of constant work. Society presents the diversions but doesn't pressure one to succumb to them. This is purely a temptation within oneself.

Distractions are also a way of modern life. The phone rings while one is thinking; the children are yelling while you're trying to carry out a complex calculation, then when you just get settled down a delivery man comes to the door. These interruptions are imposed by society, but one can be more or less tempted to be distracted by them. Concentration is helpful to overcome this temptation and self-hypnosis is useful in developing concentration.

We come next to a discussion of pressures that challenge the basic scientist. I have already discussed the temptation to earn a living, but society thru its pressures to be normal, pressure to be a good spouse, good parent, etc. also presssures one to earn a living. As discussed above, this can be a very definite challenge to doing basic research.

Many pressures are exerted by prestigious and/or granting agencies that challenge one's ability to do research that will lead towards the establishment of a basic science of information science. The effects of these pressures have serious implications about how prestigeous and/or granting

---

*Thruout this paper the word 'queer' is used with its original meaning of strange, odd, or singular, and never with its more recently acquired sexual connotations.

agencies should establish their guidelines for future prizes and/or grants. The specific pressures such as that to do only planable research, budgetary pressures, etc. will be discussed separately while the overall implications of these collective pressures will be discussed in section 4.

Many pressures are also exerted by one's academic peers, such as departmental pressures, pressures by one's daily associates, etc. and by one's professional peers, such as professional societies like ASIS, or one's colleagues in such professional societies, or simply the peers who read one's work in the journals or judge one's grant proposals for the granting agencies. These pressures will also be discussed separately under each specific pressure.

One of the most obnoxious pressures is the pressure to publish voluminously and often, even when one has nothing significant to contribute. The time and effort required to write and publish reduces the time available to do productive research. This pressure is obnoxious because it is artificial and is only imposed arbitrarily because of the difficulty of evaluating academic performance, and the desire of modern administrators to "evaluate" performance. This pressure is exerted most obviously by one's own administration, either one's own department, college, or university. But it is also exerted by prestigious and granting agencies in terms of the criteria they set for recognizing results, awarding prizes, or giving grants. This thus reinforces the pressure exerted by the administration.

The pressure most directly antagonistic to the development of a basic science of information science is the pressure to be practical. This means the pressure to produce applied or useful results rather than basic or fundamental results. A proposal to study the basic nature of information always fares badly in comparison to a proposal to facilitate the retrieval of some particular "information". This pressure is exerted most directly by the prestigious and granting agencies and also by the academic tenure and performance review committees. It is also generated by our own peers, both professional and academic, in a field which by and large sees itself as a field of practical technology rather than a basic science. ASIS and NSF could both be of tremendous help in overcoming this pressure, but I do not expect much from either one of them. SIG/FIS could address itself to this issue, but I don't know how much effect its results would have. The individual scientist will still have to meet the challenge of doing what he knows is right in the face of fewer or smaller grants, and smaller raises and less frequent promotions. He must challenge the technological bias of the entire discipline.

In a field where there exist few, if any, empirically valid abstractions and/or generalizations there is significant pressure to be concrete and/or specific. This results in an overabundance of case studies and/or specific designs or implementations and an intolerance for general inductive analyses, rigorously controlled experiments, or theoretical formulations. We face a challenge here to produce a first few valid general laws which inductively generalize from rigorously controlled experimental results and a first explanatorily useful theoretical paradigm that explains these laws of information. Once we meet this challenge we can hope that this particular pressure may decrease or disappear entirely.

Budgetary pressures are exerted both at the departmental and the granting agency level. They can be exerted because of the lack of precedent for high budget investigations in information science, the feeling that information science is not a basic science in which such expenditures are warranted, and the feeling that even where successful such researches would not have practical benefits that would pay back the investment of such expenditures. These pressures have their largest effect on research involving controlled experiments which may require expensive equipment and apparatus. The challenge here again is to achieve the first few successful experiments that result in insightful laws of information science, and thus learn how to ask experimentally answerable questions about information and how to design successful experiments. This should alleviate such pressures considerably. If, in addition, these results are useful for information technology, this particular pressure may disappear entirely.

By the pressure to conform, I mean the pressure to ask the same questions all other investigators are asking, to answer them in the same kinds of ways, etc. Don't deviate from the "traditional" methodology. This pressure is generated by one's peers who sit on tenure and performance review committees, evaluate grant proposals and referee publications. The traditional methodology has not been successful yet in establishing a basic science for information science and never will be. The successful one cannot be developed by calculation or speculative analysis. It can only be established empirically by trial and error. Therefore individual investigators must dare to be different and must meet the challenge of doing what they know is right. Granting agencies, journals, and review committees must challenge themselves to tolerate methodological nonconformity, even to the point of supporting what some might call "crackpots".

The pressure to research only "planable" areas selects only those areas where we can schedule and guarantee results. It is thus related to the temptation to do only "safe" research. A discovery that cannot be scheduled on a project plan is not worthwhile. This pressure is exerted by present granting agency guidelines. It is most effectively overcome by lying. An investigator may propose work on a topic in which he secretly already has results. He then publishes these results to satisfy the grantor, but spends his time investigating another topic where he may or may not make a breakthru. Obviously having to lie in the name of basic science is a sad comment on the state of information science and it would be more desirable for the funding agencies to establish different guidelines more sympathetic to basic research.

The pressure to be friendly and sociable with students (and other faculty too) is exerted by student evaluations which the administration may use to reward performance. If you don't spend time bull-shitting with the students and don't make the course easy for them, they will get even at evaluation time. The same goes for other faculty members who sit on review committees, etc. This pressure is the same for all investigators whether basic or applied, whether physical science, or semiotic science, and there may be little one can do to overcome it except steel oneself to accept it and learn to live with it.

Pressures to do administrative chores like handle registration, march in processions, counsel students, etc. are exerted by administrations and by other colleagues who want as little to do with such things themselves as possible. Some of it is necessary, but all of it steals time away from basic research. Some colleagues are favorably disposed towards this kind of activity and they should be encouraged to carry as big a load as possible in order to free up as much time for others as possible to do basic research. The challenge for the basic investigator is to learn to say "no" to his administrators.

The pressure to do safe research is related to the need to schedule results and also to the pressure to justify results before they are achieved. It is the opposite side of the coin of the temptation to play it safe. This pressure results in gaining assured but trivial results rather than risk being wrong on a significant topic or gaining no result at all.

Since most true geniuses are misfits, and since most results in basic science are achieved by daring to be different, the pressure to be normal, i.e., not peculiar, funny, unusual, or queer, tends to weed out the most eccentric and hence quite possibly the most likely to develop a basic science of information science.

The pressure to justify results either before they have been achieved, or before their justification is evident is applied directly by the guidelines of the granting agencies, and results in the selection of proposals for applied or technological research and mitigates against the type of research that would promote the establishment of a basic science of information science. Here is a direct challenge for the granting agencies to change their guidelines in a way that would further promote the growth of the basic science of information science. The challenge to the individual investigator is to violate such guidelines when he sits on grant application review committees, which is itself a sad commentary on both the status of the basic science of information science and on the procedures of those agencies charged with promoting basic scientific research. This pressure is related to the pressure to be practical.

The pressure to be understood by one's peers (who judge grant applications, sign recommendations for promotions, etc.) is related to publishing, but also includes talks, visits, travel, etc., all of which take time away from research. But this is one pressure that can never be relieved completely, and should never be, because it is related to a fundamental sociological aspect of all science, the requirement for a consensus of informed opinion. One must be understood by his peers in order to form a consensus of opinion.

The pressure to do menial work is exerted by one's administration. The reason given is usually "lack of funds for clerical support" and hence is also related to budgetary pressures. This includes pressures to do one's own filing, typing, editing, phone answering, etc. but also includes shopping, paying bills, appearing for traffic tickets, having medical checkups, etc. since basic research is a 24-hour activity. The challenge here is to generate enough funds to support the clerical help needed and also to find an optimum between letting all menial chores slide and hence being completely disorganized, and spending all of one's time on menial chores and hence having no time left for basic research.

The pressure to be nonesoteric can by typified by the quotation: "don't use funny sounding words, or Proximire will get you with a golden fleece award". Yielding to this pressure destroys precision in stating one's ideas. Science is an activity for self-selected technically trained experts, and the idea that all proposals, requests, reports, etc. should be understood by any layman with no training in the particular science is self-defeating for the cumulative nature of science.

I had almost forgotten about the pressure to be a good spouse and/or a good parent until a married colleague of mine reminded me of it. Whether this comes under the heading of a diversion or a distraction is debatable, but the penalty to basic research is real. Single investigators have very much the same kinds of pressure since dates place much the same demands on one's time and attention as mates.

The pressure of distractions has already been discussed under the temptations.

4. IMPLICATIONS FOR THE ESTABLISHMENT OF A BASIC SCIENCE OF INFORMATION SCIENCE:

A short analysis of each individual challenge was given in the last section. In this section I will present the overall implications of these challenges for the establishment of a basic science of information science. These implications fall into six categories as shown in Table 2.

Table 2. Categories affected by the Challenges for Developing a Basic Science of Information Science

1.	The discipline	
2.	Granting agencies	
3.	Professional societies	
4.	Journals	
5.	Academic Institutions:	department level college level university level
6.	Individual investigators	

Information science must find, select, and sponsor some Copernicuses, Brahes, Keplers, Gallileos, Newtons, Einsteins, and Peirces. Those selected must be provided with secretarial and clerical support, funded adequately, and all pressures removed from them including the pressure of time. Sponsorship should be for a minimum period of 10 years with the possibility of life sponsorship for successful performance.

It makes little difference where this responsibility is actually carried out, whether by ASIS, by DIST, or by some other institution. The main thing is that it be accomplished by and for the discipline of information science.

The discipline needs to develop approaches to creativity and technical aids to creative productivity such as self-hypnosis. Again this is a need by the discipline and it matters not what institution actually takes the responsibility for effecting it as long as it is accomplished on behalf of information science.

The discipline must begin to see itself as a basic science rather than a practical technology. Every institution in the discipline, including DIST, ASIS, academic institutions, the research journals, and all individual investigators, must share in this transformed outlook. Perhaps SIG/FIS could take the lead in the promotion of this, but it must reach to the entire discipline and must be effective.

The discipline must be tolerant of methodological nonconformity even to the point of supporting what some might call "crackpots". To some extent this is already the case, especially within ASIS, but this tolerance must be extended especially to the granting agencies, the journal editors, and the review committees. They must become willing to tolerate crackpots because they cannot be separated from the true geniuses.

The discipline should provide practical help "where it counts", i.e. teaching investigators how to manage their work environment, how to manage their time (24 hours a day), how to organize their files, how to manage their research assistants, their students, how to be creative, etc. This could be provided by any institution within the discipline, but would perhaps best fit within the present scope of DIST or ASIS.

Granting agencies must change proposal, review, and granting procedures and guidelines in order to make them more compatible with long-range basic scientific research.

Granting agencies must stop giving merely lip service to the need for a basic science, with experimental results based on assayed measurements, with conceptually explicated and empirically based explanatorily adequate theoretical paradigms, etc. They must start providing adequate funding for the required instrumentation, support for metrological studies, etc.

ASIS must begin to see itself as a scientific society rather than a professional society.

Academic institutions need to change their guidelines for performance evaluation to remove the artificial pressure resulting in the "publish or perish" syndrome. This change needs to take place at all levels, including departmental, college, and university.

Those individual investigators selected for sponsorship in the Copernicus, etc. programs mentioned earlier must be willing to devote their whole lives to the pursuit of the basic science of information science. But until such a program is initiated, those devoting themselves to basic information science must be willing to steel themselves against all diverting pressures and temptations. They must devote themselves to a life of dedication.

The individual investigators should be immune from the vicissitudes of life; i.e., do not be afraid to go hungry, do not be afraid to lose friends, do not be afraid to lose your job, do not be afraid to be called a crackpot, etc. They must be honest with themselves; i.e., never be afraid to do what is right even in the face of adverse consequences, even to the point of being bullheaded.

The dedicated individuals must dare to be different. They must also discipline themselves to use precise thinking, using precise terminology with rigorously defined definitions, etc. They must learn to say "no" to the administrators. We must bring back academic freedom.

But most importantly, we may criticize everyone else for all of the many things wrong with the state of information science today as a basic science, but none of this has any effect, unless we as individual investigators are willing to buckle down and start producing. We must begin to produce results in terms of empirically valid general laws inductively generalizing the results of rigorously controlled experiments and in terms of explanatorily adequate theories to explain these laws, rather than in terms of talk, speculation, practical designs and implemented systems. In addition, the granting agencies must begin to demand such basic results, journals must begin to publish them, and administrations to evaluate performance on the basis of how fundamental, how basic, and how important are the results achieved by us as investigators.

5. SUMMARY AND CONCLUSIONS:

In summary, the discipline of information science is challenged to see itself as a basic science rather than a collection of technological skills. Until this basic scientific methodology is established, the discipline should be tolerant of methodological nonconformity even to the point of tolerating crackpots. The discipline can facilitate the transformation to a basic science by funding a pioneering program of selecting individual investigators who are dedicated to the development of basic science and providing them with long-term support. It also needs to develop approaches to creativity and technical aids to creative productivity. It should also provide practical help teaching investigators how to manage their work environment and their time.

Granting agencies are challenged to change their proposal guidelines in order to make them more compatible with long-range basic scientific research. They must stop giving only lip service to the need for developing a basic science and begin to actually support the required instrumentation, metrological studies, etc.

The professional societies, and especially ASIS, are challenged to see themselves as scientific societies rather than as professional societies.

Journals are challenged to be more tolerant in their selection of contributions directed toward the development of a basic science, especially wherė those contributors may involve unusual or peculiar methodology.

Academic institutions are challenged to change their guidelines for performance evaluation to remove the artificial pressure resulting in the "publish or perish" syndrome and to establish criteria based on substantive and fundamental contributions to the basic science of information science.

It is the individual investigator, however, who faces the most challenges in developing a basic science of information science, and properly so. In order to develop a basic science of information science in today's milieu of a strictly technological, applied, discipline, requires individual investigators who are willing to devote their whole self to a life of dedication in pursuit of the basic science of information science. They must be willing to steel themselves against all diverting temptations and pressures, they should be immune from the vicissitudes of life, honest with themselves, must dare to be different, discipline themselves to use precise thinking, must learn to say "no" to administrators, but most importantly, we must start to achieve basic scientific results. Rather than talk about results, speculate, design or implement systems we must begin to actually produce empirically valid general laws inductively generalizing the results of rigorously controlled experiments conducted with specially designed and rigorously assayed instrumentation and to actually produce explanatorily adequate theories to explain those laws. Granting agencies must begin to demand such basic results, journals must begin to publish them, and administrators to evaluate performance on their accomplishment.

6. ACKNOWLEDGEMENTS:

This work has been supported in part by grant #DIST-7827002 from the National Science Foundation. This support is gratefully acknowledged and appreciated. I would also like to thank my colleague Vladimir Slamecka for his help and encouragement. I take full responsibility, of course, for the views expressed herein.

The Measurement of Comentropy Transfer Rates

Richard Lo

SemLab
Georgia Institute of Technology
Atlanta, Georgia 30332

Most people have heard of a game that children play called "Hangman." Each player tries to guess serially at the letters in the word that the other player has chosen, and every letter guessed incorrectly contributes to a penalty. But most of the time, the letters are correctly guessed. Given a good familiarity with a language, say American, most people would guess that the letter "n" would usually follow, if "a" were the preceding letter. Not only that, but the more letters that we know in a word, the easier it becomes to guess at the succeeding letters. Thus, if we are shown the following, "Experi," most people would easily guess that what followed would be "ment," since there are only a few logical choices. So it seems that there is a lot of redundancy built into a language such as American, or for that matter, any other language. But the problem remains in how to obtain a good measurement of this level of redundancy. How can it be measured?

A high redundancy is indicated when there is a high probability of being able to correctly predict a sign occurrence; there is little uncertainty associated with it. Therefore, by analyzing the relative frequency of sign occurrences, we should be able to arrive at an index of redundancy for any given language. There are various ways to arrive at the relative frequencies; they can either be calculated and arrived at mathematically, or they can be empirically obtained. But a look at the size of the problem indicates the level of difficulty involved in arriving at a moderately precise result. It is quite easy to obtain the relative frequencies for the occurrence of the various single letters in an alphabet, and even for digrams and trigrams, but as we go on to tetragrams and higher sequences, excessive demand for storage space for the various combinations of the letters in the alphabet poses almost insurmountable difficulties. Here, the alphabet for American is defined to be all the letters from A to Z, the apostrophe, the hyphen and the blank.

The Comentropy[1] for a given language is directly related to the concept of the average level of redundancy in use of the alphabet of a language. The amount of redundancy in the alphabet of a language can be thought of as how well we can predict what the next letter in a word will be, given a string of letters from the alphabet to start with. A high redundancy indicates a high probability of being able to correctly predict the next letter in such a string of letters, and hence, a low comentropy since the comentropy is the inverse of the redundancy.

An important instrument for the measurement of Comentropy Transfer Rates is the tachistoscope (teescope). An early experiment involving the teescope was one conducted by Cattell[2] before the turn of the century. He found that groups of four to five letters which formed a word were perceived as well as single letters, when all these were exposed to S's using a teescope. In other words, groups of letters that formed familiar words were perceived wholistically and this is called the Cattell Phenomena.

While many experiments incorporating the use of a teescope were carried out in the period between Cattell's experiment and the time Shannon's paper[3] was published, they were mostly in the field of psychology, and then, in the study of visual phenomena. Shannon laid the groundwork for the measurement of Redundancy of the alphabet for any language, the results of which were published in 1948. He showed how successive improvements in resemblance to American could be effected if discrete Markov processes were used to generate letters according to their frequency of occurrence in actual American text. A zero-order Markov process would generate letters equiprobably, a first-order process would generate letters according to their frequency of occurrence in the language, and a second-order process would generate letters according to the digram structure of the language. Thus the higher the order, the better the approximation to real American words. However, while it was apparent that improvements occurred with the higher orders, there was no method available for quantifying the levels of improvements [4]. Was a second-order generated word half as "realistic" as a third-order word? Shannon left this question open as he went on to study the matter of Comentropy, as previously defined. Another of Shannon's papers on this topic[5] delved into the question of Comentropy more deeply. Using mathematical methods based on individual letter frequencies, Shannon came up with upper and lower bounds for the Comentropy of American.

Miller, Bruner and Postman[6] proposed a hypothesis that the total amount of Comentropy that was available in a brief teescopic exposure was more or less a constant, for a given exposure time. This was based on their use of Comentropy as a mathematical model for redundancy, which was suggested by the results from Cattell's experiment and Shannon's theoretical work. Miller, Bruner and Postman used various Markov order letter generators to generate the artificial words for their experiment. Then using teescopic exposures ranging from

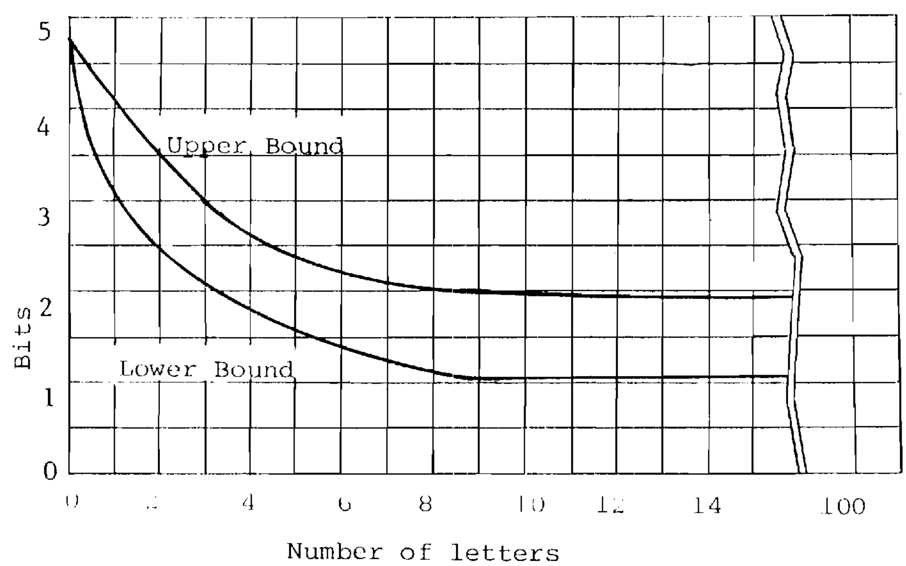


FIG. 1. SHANNON'S REDUNDANCY GRAPH

ten milliseconds to five hundred milliseconds, the artificial words were exposed to naive S's. Using the upper limit of the redundancy curves plotted by Shannon, it was found that the Comentropy in a given teescopic exposure was a constant, and therefore, the amount of Comentropy transferred was a function of the exposure time, the field contrast and brightness, but not dependent on the order of the Markov generator used, thus confirming the hypothesis. If a plot is now drawn with the X-axis representing the Markov order used to produce the words, and the Y-axis the Comentropy, a set of lines roughly parallel to the X-axis will be obtained. This indicates that given a certain teescope exposure duration, the Comentropy remains constant for words with different word shapes. This result is called the Miller-Bruner-Postman Effect.

Sperling's experiments[7] involving the use of a teescope have also been of interest, due to the results obtained in the study of information decay in memory, and the influence of pre- and post-exposure fields in teescopic exposures. These experiments were carried out in 1960 by exposing let-

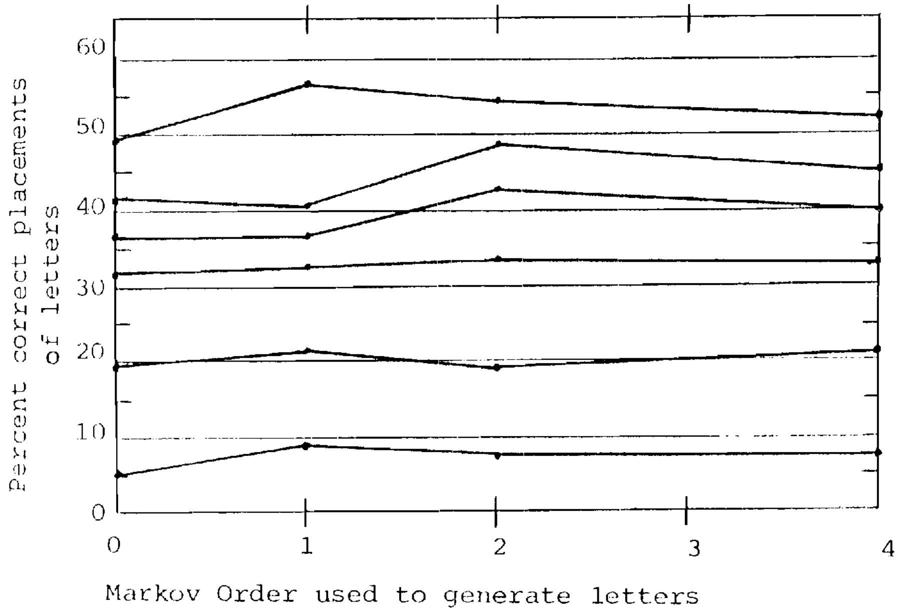


FIG. 2. GRAPH ILLUSTRATING M-B-P EFFECT

ters in the American alphabet teescopically, and then measuring the recall under various conditions. However, he was more interested in the study of immediate memory than in Comentropy transfer rates. Sperling's experiments were also more or less oriented towards individual letters rather than words. Nevertheless, some of his results prove to be rather useful in the conduct of any experiments involving the teescope. One important result was that humans have a limited 'channel capacity' of about 4.5 "chunks" of data per exposure. The concept of "chunks" has posed to be an open problem in Semiotics as it is not known at this juncture just what characterizes a "chunk" - whether they be eidemes, symbols, etc. Other experiments were concerned with the study of how exposed material was stored immediately following the exposure. For example, the study of Comentropy decay in immediate memory would help to correct for the secondary effects in the Law of Word Interpretation. It has been postulated that spelling errors measured in the M-B-P experiment, called Placement Errors, were mainly due to perceptual processes. Sperling cast doubt on that view, and results of his work point to semiotic processes in immediate memory to be the major cause of the Placement Error Effect.

While M-B-P's experiment provided valid results on Comentropy transfer rates, the relationship between the Markov order of the artificial words and the Placement Error Rate is one between a mathematical concept and an empirical one, and hence, cannot be considered a true effect. M-B-P could control the shape of the words used in the experiment by choosing an appropriate generator, but they could not measure the shape of the words thus produced.

An instrument had to be devised to be able to measure the shape of a word. Pearson[8] devised such an instrument called the Eidometer that enables the strangeness of the word shape to be measured precisely. (Latest results obtained with the Mark V, Sn.#2 eidometer runs have indicated a reliability of 3.8 bits/measurement, much better than the 3.0 bpm for the Mk. V, Sn.#1, and the 2.5 bpm for the Mk. IV, Sn.#1. Using this device, Pearson was able to carry out a similar experiment as that carried out by M-B-P. From this, Pearson discovered a relationship in 1978, which he called the Law of Word Interpretation[9]. This is a quantitative statement of M-B-P's work:

$$E = a + bS$$

where E is the teescopic error rate measured in letters per word, (lpw); S is the strangeness measured in units of Eidontic Deviance, (^{o}ED); a is a constant equal 4.61 ± 0.10 lpw; b is another constant equal -0.107 ± 0.08 lpw/^{o}ED.

However, the results showed the existence of excessive scatter in the data. A linear correlation between the Eidontic Deviance and teescopic measurements resulted in a coefficient of $r^2 = 0.44$, yielding a validity of 0.82 bits/measurement[10]. However, the resultant scatter in the final data cannot be explained by the lack of instrument precision or measurement since these had been assayed earlier, and found to be incapable of producing the scatter in the magnitude seen. Pearson went on to carry out an investigation of these secondary deviations, to determine and try to correct the measurement to account for the causes of these deviations and provide an explanation for them. Sperling's experiments had suggested that placement errors were caused by Comentropy decay in immediate memory which would manifest itself as regularities in the frequencies and positions of the placement errors in each word. Further analysis revealed that the correlation could be improved to $r^2 = 0.61$ by using a function of the frequencies and positions of the individual letters in the words to correct for the measurement of the placement errors. However, there still leaves a statistically significant gap in the correlation that cannot be explained by semiotic processes in Immediate Memory.

M-B-P considered placement errors to be all letters reported in the wrong position by the subject. Pearson mentioned two major defects in the M-B-P concept of placement error: misplaced strings, where a whole string of letters in a word, though correct, is placed in a wrong position relative to other letters in that word, i.e., ACDEFGH for ABCDEFGH and secondly, confusion errors. However, we can now distinguish between two different types of confusion errors - visual confusion errors and verbal confusion errors. Mistaking E for F, O for Q, etc., are visual confusion errors, while verbal confusion errors occur when, for example, C is mistaken for Z, E for B, etc. It is suspected that the correction terms, dependent on the frequencies and positions of the letters are actually approximations to a single, all encompassing correction term which would provide a much better correlation for the Law of Word Interpretation.

Between the S observing the teescopic exposure and then recording what he had observed, any period of time ranging between one second to a few seconds could have elapsed. It is generally agreed that the span of immediate memory lasts anywhere from half a second to a second, after which short-term memory[11] takes over. What the present study into the causes of scatter in the Law of Word Interpretation is concerned with, then, is both immediate and short-term memory. An analysis of the causes of scatter due to short-term memory involves primarily the rehearsal/verbalization process, since this process mainly occurs in short-term memory[11,12]. Hence, verbal confusion errors would occur in the period of short-term memory, during the continuous rehearsal[13]. The simplest method of overcoming this particular error would be to speed up the process between observation and recording the data. At present the subject has to look away from the eyepiece of the teescope, realign his vision and refocus onto the record chart, and then write down the results. Most of the time is taken up by this shift of focus, and the process of writing itself. This problem has a simple solution in that we could just require the S to recite what he had just seen directly onto a verbal recording device. The S would be able to begin recording almost immediately after the exposure field had been turned off, since he would neither have to remove his eyes from the Teescope viewport, nor have to write anything down. One possible problem with this method would be that the E might himself run into the problem of verbal confusion in transcribing from the tape. However, the ability to go back and check the recording, and careful procedures should eliminate this problem. Other benefits accruing to this method would be the reduction of subject fatigue due to reduced work on his part, as well as increased speed in the conduct of the overall experiment. It would also eliminate mistakes due to the S writing down something other than what he intended, very likely to occur during lapses in attention in a long experiment.

Visual confusion cannot be controlled so easily, since it occurs in immediate memory. Procedures that would increase the persistence of vision on the retina would help to reduce this error to a certain extent. Increasing the contrast between the letters and the background, and providing for a dark post-exposure field would enhance the retinal image greatly[12]. Experimenting with various varieties of typefaces used and the letter spacing would also probably yield an improved image less prone to confusion errors. Letter spacing is critical though, and too wide a spacing between letters could deter from the appearance of a single word and cause them to be viewed as individual letters or as several separate words, thus losing the Cattell Phenomena on which the experiment depends.

Errors due to misplaced strings seem to be caused by semiotic processes in short-term memory. During the rehearsal process, a letter might be dropped out, and the whole string of letters following would get shifted over by one location. If this loss were to occur in immediate memory, it would mean that within the retina image, a whole letter would have to fade away with nothing to replace the blank 'hole' left behind. This does not seem too likely as an explanation. If this were true, then it would be a short-term memory process and could be overcome by the precautions mentioned in reducing verbal confusion.

Certain strings of letters within the words might contain familiar initials, abbreviations, antonyms or acronyms or even words in a foreign language familiar to the S. There might be a tendency for the S to concentrate on these letter sequences, and subconsciously mask the other letters within the word, thus giving rise to misplaced strings, or causing the suppression of these other letters during the recall process. This effect could be checked for through the use of appropriately designed experiments. If they indeed pose a significant problem, a correction term should be developed and inserted into the final equation. However, due to the tenuous nature of this problem, it might prove difficult to remove.

The correction terms that had accounted for part of the scatter in the data in the experiment that Pearson carried out might be improved upon so as to yield a single term, instead of the present four correction terms, three of which are dependent on the absolute value functions of the letter frequencies and their positions and the last term on the Markov order of the generator used. At present, studies are being carried out to locate a single term that would give a much better correction than the four terms have achieved so far, or at least to find a similar set of terms that would yield better corrections. Pearson has also determined a set of experiments that would yield this set of improved correction terms as an alternative to the present statistical methods being used. This can be achieved through the control of the maximum and minimum letter frequencies for each letter position, as well as the maximum and minimum differences between the maximum and minimum frequencies in all letter positions within the artificial words. Pearson has also mentioned another possible source of errors as being the weighted mean familiarity of the letters making up the word.

We have now covered all the known problems that exist in the experiment to determine Comentropy transfer rates. An experiment to determine the rate of Comentropy transfer and to determine the redundancy curve for American is at present being carried out at the Semlab. All of the previously mentioned error effects are at present being analyzed and experimental procedures being designed. The basic equipment being used is a Mark VI Eidometer for determining the Eidontic Deviances of the artificial words being used in the teescopic experiments. In addition to this, the Georgia Tech Semlab has acquired a four-field teescope, together with ancillary electronic equipment for the logic and exposure control of the field. Initially, experiments as devised earlier will be carried out to analyze the various causes of errors, before a replication of the M-B-P experiment is carried out. The overall experiment thus will be carried out as follows:

(1) Carry out preliminary experiments needed to study the causes of deviations in the Law of Word Interpretation. Implement experimental controls into the actual teescope experimental procedures, and mathematical corrections into the equations for the data processing phase.

(2) Carry out experiments to assay the precision and reliability of the instruments using these procedures. Ensure that the needs of the experiment in all its phases can be met.

(3) Replicate the M-B-P experiment. This will provide an opportunity to check out the correctness of the experimental procedure and provide a "hands on" picture of any existing shortcomings.

(4) Carry out the experiment to determine the Comentropy Transfer Rates. The artificial words that have been measured previously using the Eidometer will be exposed to S's using the teescope, after compensating for the exposure times to allow for the varying perceptual sensitivities of different S's. This can be achieved by adjusting the exposure times on a set of sample words such that the overall error rate corresponds with the highest experimental sensitivity. This works out to a 50% error rate overall.

(5) The basic data is now arithmetically corrected again to achieve a 50% error rate exactly. Other correction terms which were determined earlier will be included in this data reduction phase. In addition, statistical analyses have to be made at this stage to confirm the validity of these correction factors. From this processed data, the Comentropy Transfer Rate can be determined.

(6) The data from the last phase can now be used to confirm the Law of Word Interpretation. However the statistical studies made in the last step can be used to check if any other significant error residuals remain.

(7) The previous data enables the redundancy curve for American to be drawn up. Since the Comentropy transferred remains constant for a given exposure time on the teescope, we are now in a position to measure the redundancy very easily. We now have a means of measuring the redundancy for American, since we have a scale for measuring the strangeness of word shape in terms of units of Eidontic Deviance, in addition to the teescope error rate. We can now make an equation that will yield the redundancy, given these two measurements. The scale used to measure the strangeness of the words will not have

> a valid comparison with the mathematically based strangeness in terms of the Markov order used to generate the words in the M-B-P experiment. An analysis of the data and procedure shall be made on conclusion to verify the accuracy of the redundancy curve and establish the limits of precision in the results. This concludes the entire experiment.

Many of the subsidiary experiments have been made necessary in an effort to discover the cause of the large scatter in the data from Pearson's initial experiments in this field, which would also have adversely affected the results of this experiment if they were not taken into account.

References

1. C. Pearson, "Lab Manual for Semiotics," Georgia Institute of Techology, Atlanta, Georgia (Jan. 1975). (Publication no. 017-16006-0)
2. J. M. Cattell, Ueber die zeit de Erkennung und Benennung von Schriftzeichen, bildern und farben, Philos. Stud. 2:635-650 (1885).
3. C. E. Shannon, A mathematical theory of communication, Bell System Tech. Jour. 27:379-423, 623-656 (1948).
4. C. Pearson, "Semiotics and the Measurement of Shape," Seminar Presentation at the Technische Universitaet Berlin, July 1979.
5. C. E. Shannon, Prediction and entropy in printed English, Bell System Tech. Jour. 30:50-64 (1951).
6. G. A. Miller, J. S. Bruner, and L. Postman, Familiarity of letter sequences and tachistoscopic identification, Jour. of Gen. Psychol. 50:129-139 (1954).
7. G. Sperling, The information available in brief visual presentations, Psychol. Monographs: General and Applied, 74:1-29 (1960.
8. C. Pearson, "An Empirical Concept of Word Shape," Paper presented at the Second Annual Conference of the Semiotic Society of America; Denver, Co., Oct. 15-16, 1977.
9. C. Pearson, A new law of information: An empirical regularity between word shapes and their interpretation,

10. C. Pearson, "Information Decay in Immediate Memory: A Second-Order Correction to the Law of Word Interpretation," Paper presented at the Second International Semiotics Congress in Vienna, Austria, July 2-6, 1979.

11. D. E. Broadbent, A mechanical model for human attention and immediate memory, Psychol. Review, No. 3, 64:205-215 (1957).

12. G. Sperling, Successive approximations to a model for short term memory, Acta Psychologica, 27:285-292 (1967).

13. M. Glanzer and W. H. Clark, Accuracy of perceptual recall: An analysis of organization, Jour. of Verbal Learning and Verbal Behavior, 1:289-299 (1962).

THE MARK VI: A NEW EIDOMETER DESIGN CONCEPT

Charls Pearson

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

INTRODUCTION

The eidometer is an instrument for measuring the eidontic deviance of words (strangeness of their shapes relative to a given natural language). The present design concept, called the "visual comparative eidometer", has been in use for several years and its precision has peaked with the Mk V design at about 3.8 bpm.

Increased precision is required however to determine Shannon's redundancy curve for natural language and to measure the rate of information decay in immediate memory. The redundancy curve in particular requires strangeness measurements with a precision of about 5.0 bpm.

To achieve this increase in precision a new design concept, called the Mk VI, has been invented. The design was motivated by an analogy with the vernier scale concept used with many physical instruments. The Mk VI design uses lists, each word of which is precisely measured and spaced vertically on the list according to this measurement. The observer then judges not only the list which corresponds to the strangeness being measured, but the vertical location on that list which best matches the specimen word. Since the strangeness of the specimen falls at different vertical locations on each of several different lists, it may be measured on each of these lists independently and the resulting measurements averaged to get the final value. This takes advantage of a mathematical similarity between regression calculations and mean value calculations. The averaging calculation is simplified in turn by using ten different readings in the measurement process.

REVIEW OF MARK V DESIGN

The Mk V eidometer is based on the original design concept for the eidometer, called the "visual comparative eidometer", which has been in use now for several years [4]. As reported in [8] the precision of the original Mk V was 3.5 bpm compared to about 2.7 bpm for the Mk IV. During the past year a second instrument, the Mk V, S/N 2, has been built, which together with improvements in measurement procedures and improved calibration techniques, has resulted in raising the precision to 3.8 bpm.

This improvement from the Mk IV to the Mk V is due to a series of word experiments and a "rank-averaging" technique reported before this body last year by David Howell and myself [1]. The Mk V is composed of eighteen lists, all but two of which contain exactly eight artificial words aligned vertically in semi-random order on the list. The other two lists are completely blank and appear at the extreme ends of the eidometer before and after the other sixteen. The lists are spaced 0.20 °ED apart and range from -0.20 °ED to +3.20 °ED. The first non-blank list is at 0.00 °ED and represents the highest eidontic deviance measured to date. The blank list at -0.20 °ED serves to remind the observer that this limit is arbitrary and that words may be encountered with a stranger shape than 0.00 °ED. The highest non-blank list is at +3.00 °ED and represents the lowest eidontic deviance measured to date. Again the blank list at +3.20 °ED serves to remind observers that this limit is arbitrary and that words may be encountered with a shape less strange than +3.00 °ED. The words on each list were selected for the preliminary design of the Mk V by the rank-averaging method reported by Howell and Pearson [1] using the data from a series of experiments that they carried out on 80 S's. A series of sorting experiments were then carried out by Pearson similar to those described in [5]. The purpose of the sorting experiments was to analyze the performance of the individual lists and show where they could be improved. This preliminary design and sort experiment procedure was iterated a total of eight times until a sort correlation coefficient of 0.986 and a fairly flat error spectrum was obtained. A final design of the Mk V was then built using this final set of lists.

The sorting experiments also served as a means for improving the measurement procedures and calibration techniques. During the design of the Mk V it became obvious that there are two broad categories of observers. At first these were classified as native Americans who learned to read by the "phonics" method, and native Americans who learned to read by the "whole-word" method. Later we learned the classification cannot be described quite so simplistically as this. The categories have two experimental characteristics. The first category, which I will call the "phonics" group for short, scored high on the list sorting experiments and were fairly uniform among themselves. The second category, which I will call the "whole-word" group for short, scored low on the list sorting experiments and were completely idiosyncratic both among themselves and with respect to the phonics group.

Since it was strongly suspected that this uniformity and variability carried over to the operation of the eidometer, the calibration experiments were designed to test this hypothesis. Ten S's are normally used to calibrate an eidometer. In this case we used 20 S's and measured their sort performance as well as their eidometer performance. The S's were divided into the ten highest sort performers and the ten lowest sort performers and the eidometer was calibrated separately for each group. The low sort performers had a precision of 3.01 bpm while the high sort performers had a precision of 3.66 bpm. The difference was significant. In addition the differences in eidometer measuresments among the high sort performers was small and homoscedastic, indicative of a uniform population. The differences in eidometer measurements among the poor sort performers was not only large, but as in the sorting experiments themselves, were idiosyncratic, indicative of a separate population for each individual who scored low in the sorting experiment. Many discussions with S's, hints contained in the written protocols, etc. also indicated that the low sort performers were conscious of the fact they could not distinguish shape or were having a great deal of trouble doing so, while the high sort performers, had very little trouble, were confident of their results, and were less frustrated. Because of all these results it was decided that a person's learning background and reading experience had a great deal of effect on his ability to adequately use the eidometer. At present the list sorting experiment is the only objective, quantitative index of this background. Hence in the future all measurements with the eidometer will be made only by persons who score above 0.98 in the sorting experiment.

Also during early development of the Mk V, and especially while still working with the low sort performers, it was obvious that there were individual but systematic differences among observers. For instance, an observer who measured the strangest word on the calibration list stranger than any of the other S's, was also more likely to measure all of the other strange words stranger than the other observers. Similarly, at the high end of the instrument. At first it was thought this was a paralax effect. Accordingly, a paralax correction was developed for use in data reduction and analysis. Four experiments involving 20 S's were analyzed using this correction term: high sort, low sort, and mixed sort in each of two experimental conditions. In all cases, the paralax correction was not statistically significant. Later it was discovered that this correction actually takes the form of an individual sensitivity curve obtained by regressing the calibration results for an individual against the mean of the calibration results for all S's. For the high sort performers, this correction is small, but significant, yielding a precision of 3.80 bpm for S's that pass the sort performance screening test. This is the highest precision obtained for any eidometer to-date.

Finally, there is a very strong intuitive feeling by all observers who have used the Mk V, that its scale runs backward, and the unit is poorly sized.

REVIEW OF SHANNON'S REDUNDANCY FUNCTION

Redundancy was first analyzed as a function of the information source by Shannon [9]. Shannon isolated a pseudo-empirical curve which purported to relate the amount of redundancy present in language to various Markov approximations to the language. Logically, Shannon's analysis was flawed in three ways. Even tho redundancy is an empirical concept, he could not measure the redundancy at the various Markov approximations, so he therefore analyzed the upper and lower bounds on his curve mathematically. More importantly, Markov approximations are not an empirical concept, they are a mathematical one; so that Shannon's relation was not even an approximation to an empirical relation. It was a pseudo-empirical relation and logically speaking, a bastard. Also 'Markov chain' is a concept from discrete mathematics, that has no continuous generalization, there is no such thing as a Markov chain of order 1.763.

All three of these problems are solved by recalling Markov's Law (actually only a pseudo-effect) which states that the higher the order of a Markov information source the less strange the text generated by that information source looks relative to the language from which the statistical data for the information source was gathered. Thus Shannon's redundancy curve is really a curve of redundancy versus strangeness of shape of a piece of text, and eidontic deviance is the proper concept for the logical rectification of this relationship. Further, the eidometer gives us for the first time the means for actually measuring this relationship. The methodology was hinted at by Miller-Bruner-Postman [3] and has been discussed often by Pearson. I will not discuss Shannon's redundancy curve further in this paper since Richard Lo has devoted a while paper in this session to a discussion of its measurement [2]. But the Mk V is not precise enough to carry out the required measurements of this curve. An eidometer with 5.0 bpm precision is required. This is one of the design goals of the Mk VI design.

REVIEW OF THE LAW OF WORD INTERPRETATION

Since the scatter in the measurement data for the Law of Word Interpretation is much larger than the assayed precision and reliability tolerances of the Mk IV eidometer and the teescope, Pearson sought an explanation for the semiotic reality of these deviations in the procedures used for the Word Interpretation Experiment [6]. In [7] he suggested that these deviations are due to information decay in immediate memory while the S is writing down the teescope result (which reques several hundred ms.), and gave a partial mathematical formulation of the Law of Information Decay in Immediate Memory. He also suggested several improvements in the explication of Miller-Bruner-Postman's concept of Placement Error, which he called Interpretation Error, in order to improve measurement tolerances.

In order to continue the study of semiotic processes in immediate memory, it is necessary to improve the design of the eidometer beyond the capabilities of the Mk V. Lo has analyzed the information decay process in a way that suggests semiotic processes in both immediate meory and short-term memory are involved, and he discusses this and its implications for eidometry and the measurement of information transfer rates in his paper [2].

THE VERNIER CONCEPT

A vernier scale uses two dials which slide relative to each other and is based on the fact that observers can make much finer visual judgments of relative position differences than they can of the absolute positions themselves. Thus the two scales are adjusted so that the position differences vanish at the mark that represents the last digit of the scale reading.

The vernier concept itself can not be carried over to the eidometer design because all current eidometer designs utilize several different lists, each containing words of uniform strangeness, for the purpose of making comparative judgments. Thus no continuous sliding scale is available for vernierization. However, the two vernier scales may be viewed as a two-dimensional function that serves to bring the independent values in the two separate dimensions into conjunction. It is this generalization of the vernier concept which can be applied to the eidometer.

Using this generalization, it is no longer necessary for the two scales to either slide, or be continuous; in fact it is not even necessary for them to be contiguous. They may in fact lie in separate dimenions orthogonal to each other, such as the x and y coordinates of a graph. It is only necessary that the two scales measure the same values independently and function to bring these independent measurements into conjunction.

THE MATHEMATICS OF DATA REDUCTION: SIMPLIFICATION BY DESIGN

The key to understanding the design of the Mk VI is to imagine the artificial words on each list to be measured very precisely and to be placed on the list with a vertical displacement precisely determined by this measured value. The range of values on each list overlap to the extent that each value may be measured on ten different lists, at ten different vertical locations. The x and y coordinates thus represent measurements of a straight line that intersects the axis of the eidometer at the measured value. This is the vernier function of the design.

This point can be estimated precisely by the statistical method of regression, thus eliminating any requirement for geometrical construction. However, by spacing each list precisely the same distance apart and the same interval of eidontic deviance apart, advantage can be taken of the similarity between the regression equations and the equations for calculating mean values. This reduces the estimation of the actual value of strangeness to the calculation of the mean of several measurements. Further, by allowing exactly ten independent measurements of each value, the calculation of arithmetic means is reduced to simple addition and a decimal point shift. Finally, by suitably labeling each word on each list, no actual point of intersection need ever be found, and no geometry is required. All work can be done by simple arithmetic using either paper and pencil, calculator, or computer. This allows a further simplification of the design. The lists can be folded so that each word serves as the index of both a vertical scale and a horizontal scale.

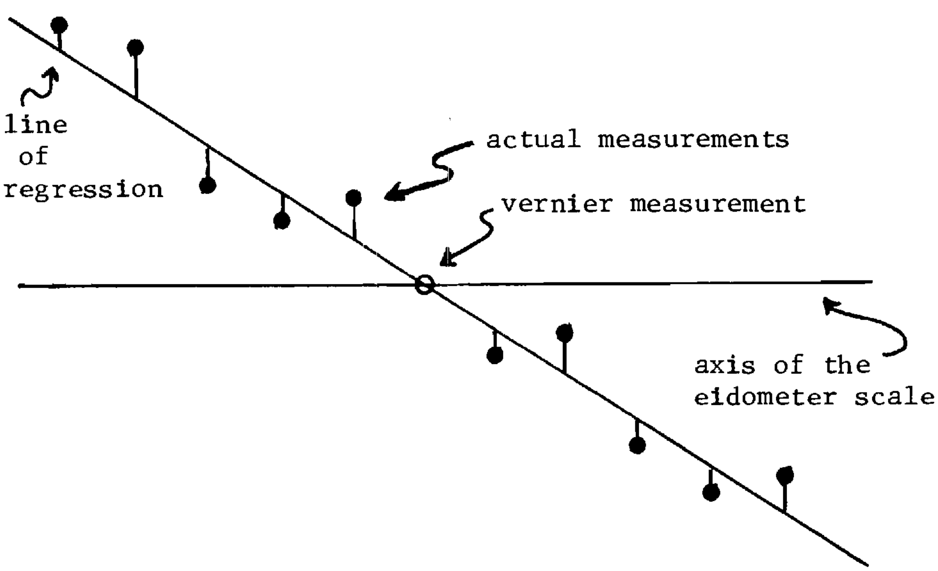


Fig. 1. Vernier Function of the Mk VI Design.

A measurement then consists of five vertical interpolations and five horizontal interpolations. The regression equations are not affected, but the total size of the instrument can be reduced from 632 words to 184 words cutting the development cost by about a factor of four.

In summary we have replaced geometrical construction by statistical regression, replaced the regression with a calculation of mean values, replaced the calculation of the mean with arithmetic addition, and finally eliminated the original vernier scales by folding to reduce the overall size of the instrument.

THE DEVIOMETER SCALE

In order to rectify the intuitive feeling that the eidontic deviance scale is reversed, and to take advantage of the gain in measurement structure obtained with the Mk IV and Mk V designs (the present °ED scale was defined when going from the Mk III to the Mk IV design), I define a new scale called the deviometer scale and abbreviated dev. The deviometer scale runs from 0 dev (completely normal -- no strangeness to 100 dev (very strange). The present 0 point on the °ED scale intersects the deviometer scale at 65 dev. The present 3.0 point on the °ED scale intersects the deviometer scale at 35 dev. One dev. is thus equal to -0.100 °ED which is just a half of one standard deviation of individual measurements made with the Mk V instrument.

For purposes of gross comparison only, then, the scale transformations are given by:

$$(\text{Dev}) = 65 - 10\ (^\circ\text{ED}) \qquad (1)$$

and

$$(^\circ\text{ED}) = \frac{65 - (\text{Dev})}{10} \qquad (2)$$

However, it must be kept in mind that these equations do not define the scales themselves. This is done by means of the operational definitions. The transformation can only be used for gross comparisons within the tolerances established by the measurement structure of the °ED scale which does not have a meaningful origin and has tolerances of ± 0.2 °ED or ± 2 dev.

The deviometer scale is an interval scale, but has no natural origin, while the °ED scale was ordinal only.

DESIGN PARAMETERS OF THE MARK VI

The Mk VI uses 184 words arrayed in 34 lists to obtain 5 vertical and 5 horizontal interpolations. The words are laid out as follows where a value is used to indicate a word whose strangeness has that value within the design tolerance of ± 1/3 dev.

36	37	38	39	40	41	42	64	65				
35	36	37	38	39	40	41	63	64	65			
	35	36	37	38	39	40	62	63	64	65		
		35	36	37	38	39	61	62	63	64	65	
			35	36	37	38	60	61	62	63	64	65
				35	36	37	59	60	61	62	63	64

Fig. 2. Layout of the Mk VI Scale.

In order to assure obtaining 184 words within the required tolerance ± 1/3 dev of the target values, our design effort starts by choosing 750 words equally spaced throut the range from 0.0 to 3.0 °ED. Thirty-six S's will be used to obtain the required precision using the Mk V, giving 27,000 S-words. S's can measure approximately 100 words per hour; we therefore have 270 S-hrs @ $5/S-hr, giving a S cost of $1,350.

With this design we expect achieve a standard deviation for individual measurements with the Mk VI of ± 1/2 dev. leading to a precision of better than 5 bpm.

NEW COMPUTER PROGRAMS

Several new computer programs are required to support this development effort. One fairly simple program is required to manage the 27,000 S-word measurements and calculate the mean measurement values, standard deviations, and standard errors. Because DEVIOS is designed specifically for the Mk IV-V series eidometers, a completely new calibration program will be required for the new design. Several other data reduction and calibration programs are anticipated.

BUDGET ESTIMATES FOR THE MARK VI DEVELOPMENT

The following budget estimates are for purposes of illustration and evaluation only and do not reflect the actual accounting procedures of the SemLab. The budget items are categorized to variable costs which are specifically due to the Mk VI development and would not be spent if the Mk VI were not developed; assignable fixed costs, which can be assigned to the Mk VI development but would be spent on other project work if the Mk VI were not developed; and overhead fixed costs which are not attributable to the Mk VI development per se and would have to come out of other funds if the Mk VI were not developed.

Variable Costs	
$ 1350	subject expenses
500	program development charges
500	computer charges (data reduction)
2350	
Fixed Costs (assignable)	
500	GRA
1000	PI
1500	
Fixed Costs (overhead)	
3080	(80% of $3850)
$ 6930	Total costs
$ 7000	Estimated total costs
$ 2500	Estimated variable costs

SUMMARY

A new eidometer design concept has been introduced that is motivated by the concept of a vernier scale but in which the mathematics of vernierization have been so simplified by specific design that all geometric construction has been eliminated and all calculations reduced to simple addition. This development requires a budget of approximately $7,000 and will yield an instrument with a precision of better than 5 bpm.

The new instrument will enable the measurement of Shannon's redundancy curve as a function of eidontic deviance, an important problem in the semiotics of natural language; and the investigation of information decay in immediate and short-term memory, an important investigation in human semiosis. It is expected that the Mk VI will lead to new laws of information decay, empirical explication of the concept of semiotic interpretation, and several other laws of syntactics, semantics and pragmatics.

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, and my assistants Mr. Richard Lo and Mr. David Howell, for their help in performing the many experiments required by this project.

REFERENCES

1. Howell, D.P.; and Pearson, C. "A New Technique for Eidometer Construction". Presented to the session on "Methods of Experimental Semiotics" sponsored by SIG/ES at the 1979 Annual Conference of the Semiotic Society of America in Bloomington, Ind; October 5, 1979.

2. Lo, R.H. "Measurement of Information Transfer Rates". Presented to the session on "Paradigms for Experimental Semiotics" sponsored by SIG/ES at the 1980 Annual Conference of the Semiotic Society of America in Lubbock, Texas; October 17, 1980. To appear in the published proceedings.

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

4. Pearson, C. "An Objective Concept of 'Word Shape' for Natural Language". Presented to the "Symposium on Empirical Semiotics" sponsored by SIG/ES at the 1977 Annual Conference of the Semiotic Society of America in Denver, Colorado; October 14-15, 1977.

5. Pearson, C. Lab Manual for Semiotics. (Revised January, 1978), Georgia Institute of Technology Publication 017-

16006-0; Atlanta, January, 1975.

6. Pearson, C. "A New Law of Information: An Emprical Regularity Between Word Shapes and Their Interpretation". tion". Presented to the forty-first annual conference of the American Society for Information Science; New York; November, 1978.

7. 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 published proceedings.

8. Pearson, C. "Preliminary Calibration Results with the Mk. V Eidometer". Presented to the session on "Methods of Experimental Semiotics" sponsored by SIG/ES at the 1979 Annual Conference of the Semiotic Society of America in Bloomington, Ind; October 5, 1979.

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

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 Revoluations, 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.

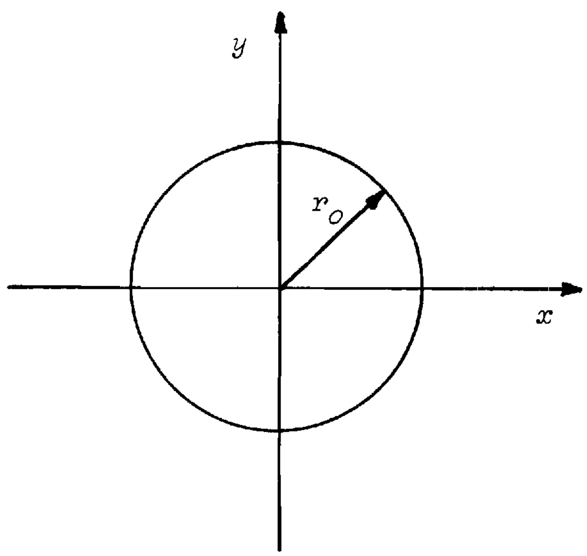
Conceptual Paradigms and the Language of Science

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 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.

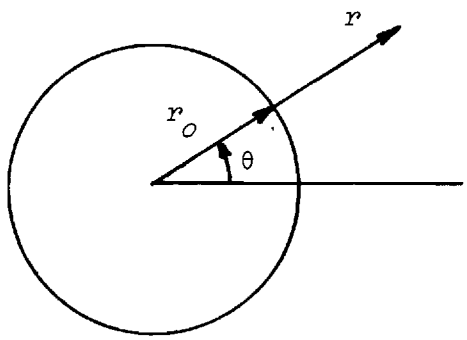
However, their effects may be drastically substantive in that certain empirically substantive questions may be drastically easier to express in one language than another. This is illustrated in figure 1. Figure 1.a shows a circle as described by rectangular coordinates and gives the corresponding algebraic 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



$x^2 + y^2 = r_o^2$

a)



$r = r_o$

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 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 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 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 semotics. 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. Kuhn, T.S. The Structure of Scientific Revolutions. U. of Chicago Press, 1962.

2. Newell & Simon. "Computer Science and Empirical Inquiry".

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

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

5. 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 in Vienna, Austria; July 2-6, 1979. To appear in the published proceedings of the congress.

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

7. Pearson, C. "Semiotics and the Measurement of Shape". Seminar presented at the Technische Universitaet Berlin; July 8, 1979.

8. Pearson, C; and Slamecka, V. "A Theory of Sign Structure". Semiotic Scene, 1(1977), p1-22.

For immediate release.

SEMLAB'S TYPE-TOKEN PROGRAM
NOW AVAILABLE ON THE CYBER-70

by Thanarak Phongphatar

TTKANAL, the SemLab's program for the type-token and rank-frequency analysis of text which was previously available only for the Burroughs B5700 has now been modified and extended for the CDC Cyber-70.

The program keeps running counts of word-types and word-tokens in a sample of text stored on magnetic tape and prints these out for statistical analysis. It also assigns ranks to each word-type on the basis of observed frequencies and sorts these into rank order for print out.

TTKANAL was written in CYBER extended FORTRAN IV. The present version is limited to text samples containing no more than 20,000 wt., but users could modify the program to relax this limitation. In its present version it accepts only text coded in Georgia Tech type C coding, but this can be modified by adopting suitable coding conventions both within the program and in the coded text.

This work was partially supported by NSF-DIST under grant No. IST-7827002.

For further information contact:

Dr. Charls Pearson
SemLab
School of Information and Computer Science
Georgia Institute of Technology
Atlanta, 30332, USA
Tel: 404-894-3988 or 404-881-1100

APPENDIX

Updated Human Subjects
Certification To Be
Forwarded Under Separate
Cover.