

# THE FINITE STRING

NEWSLETTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS

VOLUME 12 - NUMBER 4

SEPTEMBER 1975

This issue of the Journal contains no bibliography beyond announcements of some new books; a retrospective list from the Fondazione Dalle Mole, and two reviews. The Editor has been away from his office and the former Editorial Assistant is now employed full time elsewhere. Prospects for the next issue look good.

**A**MERICAN JOURNAL OF COMPUTATIONAL LINGUISTICS is published by the Center for Applied Linguistics for the Association for Computational Linguistics.

**EDITOR:** David G. Hays, Professor of Linguistics and of Computer Science, State University of New York, Buffalo

**EDITORIAL SECRETARY:** Jacquin Brendle

**EDITORIAL ADDRESS:** Twin Willows, Wanakah, New York 14075

**MANAGING EDITOR:** A. Hood Roberts, Deputy Director, Center for Applied Linguistics

**PRODUCTION AND SUBSCRIPTION ADDRESS:** 1611 North Kent Street, Washington, Virginia 22209

Copyright © 1975  
Association for Computational Linguistics

## TABLE OF CONTENTS

A C L Nominating Committee . . . . .	3
13th Annual Meeting: Abstracts of Papers . . . . .	4
Fabens, William, 5	Hendrix, Gary G. 20
Hobbs, Jerry R., 6	Sondheimer, Norman K., 21
Miller, Perry L., 7	Cercone, Nick, 22
Martin, William A., 8	Deutsch, Barbara G., 23
Burger, J. F. et al., 9	Bruce, Bertram, 24
Medewa, P., et al., 10	Phillips, Brian, 25
McDonald, David, 11	Cullingford, R. E., 26
Knaus, Rodger, 12	Badler, Norman, 27
Rhyne, J. R., 13	Kegl and Chinchor, 28
Shapiro, Stuart C., 14	Rosenschein, Stan, 29
Slocum, Jonathan, 15	Klappholz and Lockman, 30
Meehan, Jim, 16	Beckles, Carrington, Warner, 31
Bates, Madeline, 17	Brill and Oshika, 32
Paxton & Robinson, 18	Salton, G., 33
Robinson, J. J., 19	Anderson, Bross, Sager, 34
A S I S Annual Meeting . . . . .	35
A C M Annual Conference . . . . .	39
SOLAR Project Terminates--ARPA support ends . . . . .	42
Energy Information Tools: IIA-NFAIS Workshop . . . . .	42
MT: Chinese Journals offered on subscription from Hong Kong	43
Librarian of Congress: ASIS telegram to President Ford . . .	44
Bibliography on Semantics and Cognition at Castagnola . . .	46
Recent Publications . . . . .	48
Conceptual Information Processing, 48	
Automatic Translation at Grenoble, 49	
Frequency and Distribution (French novels), 51	
A L L C Bulletin, 52	
Word Order and Word Order Change, reviewed by James M. Dunn	53
Informal Speech, reviewed by John B. Carroll . . . . .	78
Personal Notes . . . . .	92



NOMINATING COMMITTEE

A slate of officers for 1976 will be presented at the annual meeting in Boston by

William A. Woods, Jr., Chairman

Bolt Beranek and Newman Inc.  
50 Moulton Street  
Cambridge, Massachusetts 02138

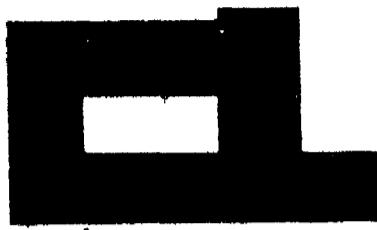
Robert Simmons

University of Texas

Robert Barnes

Lehigh University

Members of the Association can submit suggestions to the chairman or any member of the Committee.



THIRTEENTH ANNUAL MEETING

ASSOCIATION FOR COMPUTATIONAL LINGUISTICS

A B S T R A C T S   O F   P A P E R S   A C C E P T E D

PROGRAM COMMITTEE:    TIM DILLER, CHAIRMAN

*Sperry-Univac*

JON ALLEN

*Massachusetts Institute of Technology*

JOYCE FRIEDMAN

*University of Michigan*

BONNIE NASH-WEBBER

*Bolt Beranek and Newman, Inc.*

CHUCK RIEGER

*University of Maryland*

TIME AND PLACE:    Sheraton Boston Hotel

October 30 - November 1, 1975

SESSIONS:

1. LANGUAGE UNDERSTANDING SYSTEMS
2. LANGUAGE GENERATION SYSTEMS
3. PARSING, SYNTAX, AND SEMANTICS
4. MODELING DISCOURSE AND WORLD KNOWLEDGE
5. TEXT ANALYSIS

BUSINESS MEETING AND ELECTION OF OFFICERS

ABSTRACTS APPEAR ON THE FOLLOWING FRAMES

## PEDAGLOT AND UNDERSTANDING NATURAL LANGUAGE PROCESSING

WILLIAM FABENS  
*Rutgers University*

PEDAGLOT is a language processor that contains a meta-parser. This parser is programmable not only in its syntax, semantics, and their related activities, but also with respect to the modes of operation. The idea of parsing modes differs from other ways of programming parsers in that it is related to the behaviors desired from a parser rather than to its lower level functions. The dozen modes described are roughly independent of one another.

The main activities of parser form a dialectic process of prediction, discovery and construction, all embedded in a higher order control structure. The modes indicate the behavior of each such activity.

This way of programming a parser is seen as a step toward representing theories of natural language in a unifiable way, in the sense that various mode settings might yield parsers, generators, language inferrers, etc. The matter of unification is found to be non-trivial (i.e. a generator is not just an inverse parser), yet the notion of modes makes it at least tractable. Applications of PEDAGLOT to various current theories are discussed

A SYSTEM FOR GENERAL SEMANTIC ANALYSIS  
AND ITS USE IN DRAWING MAPS FROM DIRECTIONS

JERRY R. HOBBS  
*The City College of CUNY*

Our data base is a set of facts involving spatial terms in English; our system takes as input directions of how to get from one place to another and outputs a map. As input we use the output of the Linguistic String Project's transformational program.

The problem of semantic analysis is to find quickly out of a potentially enormous collection the appropriate inferences. The key to selection is Joos' Semantic Axiom Number One (restated) "The important facts in a text will be repeated, explicitly or implicitly." Those inferences which should be drawn are those which are keyed by more than one element in the text.

Semantic interpretations include interpretation of higher predicates, finding antecedents of definite noun phrases, and identifying intersentence relations

We divide the sets of inferences into clusters, this organization of the data base is task dependent, in our application, the top-level cluster concerns the one-dimensional aspects of objects and actions. When a fact in a cluster is accessed it becomes the top-level cluster

We tag our inferences always, normally, or sometimes

The task component makes arbitrary decisions required by the map but not given in the text. A geometry is imposed on the topological natural language information and the map is drawn.

## AN ADAPTIVE NATURAL LANGUAGE SYSTEM THAT LISTENS, ASKS, AND LEARNS

PERRY L. MILLER  
*Massachusetts Institute of Technology*

When a user interacts with a natural language system, he may well use words and expressions which were not anticipated by the system designers. This paper describes a system which can play TIC-TAC-TOE, and discuss the game while it is in progress. If the system encounters new words, new expressions, or inadvertent ungrammaticalities, it attempts to understand what was meant, through contextual inference, and by asking intelligent clarifying questions of the user. The system then records the meaning of any new words or expressions, thus augmenting its linguistic knowledge in the course of user interaction.

## CONCEPTUAL GRAMMAR

WILLIAM A. MARTIN  
*Massachusetts Institute of Technology*

In OWL, an implementation of conceptual grammar, the two types of data items are symbols and concepts and the two basic data composition operations are specialization and restriction

A symbol is an alphanumeric string headed by " ". Symbols correspond to words, suffixes, prefixes, and word stems in English and the programmer can introduce them at will.

OWL concepts correspond to the meanings of English words and phrases. They are constructed using the specialization operation, comparable to CONS in LISP. (A B) is the specialization of A, a concept, by B, a concept or symbol. OWL forms a branching tree under specialization, with SOMETHING at the top.

Concepts are given properties by restriction, which puts a concept on the reference list of another concept (compare property lists and S-expressions in LISP). A/B is the restriction of A by B.

The categories in the specialization tree are semantic, but we use them also for the purposes usually assigned to syntactic categories.

A predication is a double specification of a model such as present tense or can. Examples are

The pool is full of water. ((PRES-TNS (BE (FULL WATER))) POOL/THE)  
The cookie can be in the jar ((CAN (BE (IN JAR/THE))) COOKIE/THE)  
Bob is the father of Sam ((PRES-TNS (BE (FATHER SAM)/THE)) BOB)  
Bob hits the ball. ((PRES-TNS (HIT BALL/THE)) BOB)  
Bob is hitting the ball ((PRES-TNS (BE (-ING (HIT BALL/THE)))) BOB)

Starting from this base we will discuss a number of issues such as nominalization, incorporation, and deep vs surface cases.

## SEMANTIC-BASED PARSING AND A NATURAL-LANGUAGE INTERFACE FOR INTERACTIVE DATA MANAGEMENT

JOHN F BURGER, ANTONIO LEAL, AND ARIE SHOSHANI  
*System Development Corporation*

This paper describes the current state of work-in-progress on a system having both applied and theoretical relevance for computational linguistics. At the applications level, the program we are developing is an interface that will give a natural-language communications facility to users of existing data management systems. The interface can be adjoined to a wide variety of data management systems with relatively little effort. Advantages and disadvantages of this approach to reducing the "DMS communications gap" are discussed.

The theoretical part of the work shows that useful information in a natural-language expression (i.e., its "meaning") can be obtained by a parser controlled by a grammar that uses no description of syntax whatever. The construction of the parsing tree is controlled primarily by semantics in the form of a concept network consisting of an abstraction of the "micro-world" of the data management system's data organizing methods, its functional capabilities; and the semantic relations of the data base content material. Discussion includes a possible method whereby such a parser might be extended to be used with more general "worlds" if "frames" could be used to temporarily restrict world-view information.

## PHILIQA I: MULTILEVEL SEMANTICS IN QUESTION ANSWERING

P. MEDEWA, W. J. BRONNENBERG, H. C. HUNT, S. P. J. LANDSBERGEN,  
R. J. H. SCHA, W. J. SCHOENMAKERS, E. P. C. VAN UTTEREN  
*Philips Research Laboratories, Eindhoven*

This paper outlines a recently implemented question answering system, called PHILIQA 1, which answers English questions about a data base.

Unlike other existing systems, that directly translate a syntactic deep structure into a program to be executed, PHILIQA 1 leads a question through some intermediate stages of semantic analysis thereby providing for a perspicuous treatment of some hitherto largely neglected semantic phenomena. In every stage, a question is represented as an expression in a formal language. The paper sketches the syntax and semantics of the languages designed for this purpose, and points out the distinctions between the semantic representations used at the most important levels of semantic analysis, viz.:

- one at which the constants and language constructs are derived from those of English,
- one that may be characterized as representing the system's assumptions about the "structure of the world", reflecting neither the peculiarities of English nor those of the data base,
- one at which the constants match the ones in the data base and the available arithmetical and logical functions, while the language constructs are well-suited to algorithmic interpretation.

PHILQA 1 was designed in such a way that it can be described abstractly, independent of the implementation, as a series of sets of transformation rules, called "convertors". A convertor translates expressions of one level of semantic interpretation into expressions of the next level. The paper gives a sketch of each of these convertors, and shows how they translate an example question step by step into the data base language.

## A FRAMEWORK FOR WRITING GENERATION GRAMMARS FOR INTERACTIVE COMPUTER PROGRAMS

DAVID McDONALD

*Massachusetts Institute of Technology*

Interactive programs which wish to employ fluent natural language will be communicating specific messages to their users in specific pragmatic and discourse situations. Accordingly, a generation grammar must be so structured that the particulars of the message and the context themselves direct what linguistic devices--words, phrases, syntax--are used in translating the message to English.

A program's message describes what objects and relations are to be mentioned and what the program's intentions are (e.g. to make a prediction, to answer a query, etc.). In the framework described, a grammar consists of a body of procedures which perform the translation in two phases. The first phase uses procedures associated with the intentions to construct a plan for the utterance as a whole--a surface level model incorporating the message's elements largely unanalyzed, but in their intended positions (subject, main verb, etc.).

During the second phase, the plan is refined, from left to right, following its constituent structure. Each message element is analyzed and translated in turn as it is encountered in the plan. The translation is done by discourse-sensitive procedures taken from a lexicon of the possible elements which the program might use. By working from left to right, possible combinatorial interactions between descriptive processes are avoided. Also, a convenient formalism for encoding surface structure dependant phenomena such as pronominalization and quantifier scope becomes possible.

The framework is described with examples drawn from the implementation in progress for an appointment scheduling program. Particular attention is given to how the characteristics of natural language dictate or influence the framework's design.

## INCREMENTAL SENTENCE PROCESSING

RODGER KNAUS  
*University of California, Irvine*

Human short term memory is bounded while the speaking rate is nearly constant. Therefore a listener must use a parsing algorithm for which there is a uniform bound over all sentences for the time needed to add a new word onto existing partially completed parses. A context free parsing algorithm, the incremental parser, is presented which has a bounded parsing time per word on a class of context free sentences containing the syntactic phrase markers of natural language sentences. Psycholinguistic evidence supports the breadth first strategy used by the incremental parser.

The incremental parser is extended to a class of generation grammars with function-like rules more suitable for natural language sentence generation than context free grammars. Instead of grammar nonterminals each generation rule expands a syntactic form consisting of a grammatical category C (noun phrase, sentence, etc.) followed by a list of rule arguments, including a semantic network node N and an association list about the purpose of the sentence. The generation rule expands the syntactic form into a list of words and syntactic subforms each of which generates a syntactic substructure of C from a semantic substructure of N.

A heuristic function FIND is defined which when given a syntactic form F and a list L of syntactic forms generated by F finds a set of association lists A such that with the variable bindings of A, evaluating F produces L. By a bottom-up application of FIND to the syntactic structures built by a regular expression incremental parser, the generation grammar parser finds the semantic nodes and sentence purpose alists which might have produced a given sentence from a generation grammar.

## A LEXICAL PROCESS MODEL OF NOMINAL COMPOUNDING IN ENGLISH

J. R. RHYNE  
*University of Houston*

"Lexical processes" are rather idiosyncratic. Some examples are nominalization, lexical incorporation, nominal compounding, and lexical substitution. These processes are partly syntactic, partly semantic, and partly controlled by information associated with each lexical item. These processes have been largely ignored by computational linguists; recent models with a substantial knowledge base suggest that modeling lexical processes will not be quite as large a task as might have previously been thought.

I have constructed a model which accepts relative clauses and produces nominal compounds. The syntax is simple, if compounding changes a relative clause into a noun-noun pair, it has no semantic aspect. However, only a few relative clauses can be changed into compounds; this process must be controlled by the lexical items

The computer model uses lexical rules associated with items in the lexicon. These rules consist of left part and right part structures representing relative clauses and nominal compounds in a case structure system.

One form of English compound is made by deleting the main verb and several noun phrases and placing the noun which remains in front of the head noun of the original relative clause. This process can be allowed only when the deleted information is "lexically recoverable".

The model has been used to generate several hundred nominal compounds and is very efficient, even with a lexicon containing nearly a hundred rules and capable of generating several thousand different nominal compounds. The kinds of rules used can account for other lexical processes. The same rules can be used for recognition of nominal compounds; this has not been done and would probably require a substantial knowledge base to disambiguate some compounds.

## GENERATION AS PARSING FROM A NETWORK INTO A LINEAR STRING

STUART C. SHAPIRO  
*Indiana University*

In this paper, we discuss the approach we are taking to the generation of English surface strings from a semantic network. The semantic network representation does not contain surface features of the original sentences such as tense or voice. Instead of tense, temporal information is stored relative to a growing time line. Voice is considered to be information about the original speech act rather than essential information conveyed by the sentence. As a result, the generated sentences are not necessarily the same as the original sentence.

We view generation as the creation of a linear surface string that describes a node of the semantic network. The form of the surface string is controlled by a recursive augmented transition network grammar which is capable of examining the form and content of the semantic network connected to the semantic node being described. A single node of the grammar network may result in different forms of surface strings depending on the semantic node it is given, and a single semantic node may be described by different surface strings depending on the grammar node it is given to. For example, a semantic node may be described as an independent sentence in one instance, as a relative clause in another, and as a nominalized sentence in another.

Our approach is to generate surface strings left to right. Rather than start with several deep phrase markers which must be connected appropriately and transformed, we are starting with a network where deep structures are already properly connected. These can be examined by the grammar network which can then build the final surface string directly.

## SPEECH GENERATION FROM SEMANTIC NETS

JONATHAN SLOCUM  
*Stanford Research Institute*

Natural language output can be generated from semantic nets by processing rules that are associated with verbs which correspond to concepts in the net. A rule is essentially a sequence of case names. The set of rules is being derived from a study of the surface syntax of some 3000 English verbs. The active forms of the verbs have been classified according to subject, object(s), and complements--ignoring adverbials of manner, time, distance, etc., which are inserted by heuristic rules. Passives are similarly derived. These major argument patterns are in the process of being converted semi-automatically into sequences of case names which will be used as templates in the generation of text. The text will be in a form that can be entered into a speech synthesis program. Some initial experiments with a VOTRAX speech synthesizer are being conducted.

## USING PLANNING STRUCTURES TO GENERATE STORIES

JIM MEEHAN  
*Yale University*

This paper discusses a computer program which makes up stories from its knowledge of the world, including the characters planning structures (goals and plans for achieving those goals). Goals can be constant (such as preserving one's health) or recurring (eating whenever you're hungry), and can produce subgoals which are more immediate in nature (eating a hamburger now). A plan for achieving a goal can include several subplans, each with a set of applicability tests and enabling preconditions. There must also be a decision algorithm to use in deciding which of several subplans to try first. The criteria for a subplan can relate to the character who is doing the planning, to the other characters in the story, or to general information about the world. After the initial state of the world is established, the program can change details, present obstacles to goals, and introduce unusual events to make the story interesting. A causal chain of states and acts, representing the story in Conceptual Dependency notation, is generated by programs which model combinations of planning structures. The structures are combined by nesting (when a precondition for a particular subplan requires other plans), in series (when several plans are required), or in parallel (when goals are concurrent). The characters in the present system are talking bears, birds, and other animals. This simplifies somewhat the problem of handling enormous quantities of world knowledge, but without changing the basic problems in generating coherent, interesting stories.

## SYNTACTIC PROCESSING IN THE BBN SPEECH UNDERSTANDING SYSTEM

MADELINE BATES

*Bolt Beranek and Newman Inc.*

The syntactic analysis system presented here is composed of two parts, a modified augmented transition network grammar and a parser which is designed for a speech understanding environment.

The parser operates on partial utterances called theories. A theory may be thought of as a set of words which are hypothesized to be in the utterance. The parser processes the words in a theory by building partial syntactic paths using the words of the theory. These paths do not depend on left context, which will be missing if there are gaps in the theory. Syntactic constituents are built where possible and, whenever a constituent is built, the parser can interface with the semantic component of the total speech understanding system for guidance and verification.

The parser tries to predict words and/or syntactic categories to fill or reduce gaps in the theory, particularly small function words which are difficult to detect reliably on acoustic grounds alone. The parser does not follow all possible parse paths, but attempts to select the most likely ones for extension. It uses a judicious mixture of top down, bottom up, depth first, and breadth first parsing strategies to take advantage of local, reliable information. It saves all the information gained while following alternative parse paths, so that several parse paths which share a common part, even if the paths are in different theories, can share that portion without reparsing. This is true even if the parse paths split before and or after the common part and even if the common section analyzes only part of a syntactic constituent.

## SYSTEM INTEGRATION AND CONTROL FOR SPEECH UNDERSTANDING

WILLIAM H. PAXTON AND ANN E. ROBINSON  
*Stanford Research Institute*

Acoustics, syntax, semantics, discourse, and pragmatics play roles in speech understanding and can be integrated into a system that allows the interactions to be easily visible.

The language definition is the focal point for the integration. Basic to the language definition are phrases built from individual words and from other phrases. Integration occurs at the level of each individual phrase. For each phrase type, a statement in the language definition specifies (1) which kinds of knowledge to use and (2) how much weight to give to each source in computing the likelihood that an instantiation of the phrase type is a correct interpretation.

The executive uses a complex heuristic control strategy to control sources of knowledge, establishing priorities for alternative tasks. The processing of an utterance is factored into tasks that make incremental changes to a global data structure and spawn other tasks. Priorities reflect both the expected values of interpretations and the relation of the task to the executive's current focus of activity. The expected values take into account the context established by prior tasks and are based on phrase scores that combine non-Boolean evaluation factors from a variety of knowledge sources. The focus mechanism allows tentatively accepted phrases to inhibit the search for others that would replace them. The tasks and the global data structure are structured in a way that brings together related activities to eliminate duplication of effort and makes it possible to coordinate processing driven by acoustic data with processing driven by goals based on predictions made by higher level linguistic components.

## A TUNEABLE PERFORMANCE GRAMMAR

JANE J ROBINSON  
Stanford Research Institute

A performance grammar (PG) aims to define the form and meaning of intelligible speech uttered during the course of spontaneous dialog. Its definitions are tuneable to particular utterances in particular dialogs. That is, given the problem of determining the applicability of a definition to the understanding of some portion of an utterance, the definition itself specifies how the attributes of that portion and the properties of the discourse and the speaker affect the likelihood that it should be applied.

This paper presents a tuneable PG being developed for a computer-based speech understanding system. Two different discourse contexts from which its definitions are derived are compared and contrasted, and ways of tuning the word and phrase definitions to them by means of 'factor' statements are described. Sequences of definitions involved in parsing and interpreting utterances are examined in detail, emphasizing the interaction of selected factors for evaluating the likelihood of their application. The selected factors are called 'syntactic', but the attributes that are evaluated may be semantic, pragmatic, acoustic-phonetic, or discourse based. It is shown that superficial syntactic factors are useful for disconfirming a wrong parsing path or confirming a correct one, in ways that reduce the need to call on the semantic, discourse, and acoustic components for in-depth evaluations. It is also shown that factors evaluating number agreement, which is traditionally a syntactic matter, need to refer to semantic attributes. This demonstration points to the conclusion that integrating information from different kinds of analyses or sources of knowledge is well-motivated on both linguistic and heuristic grounds.

## SEMANTIC PROCESSING FOR SPEECH UNDERSTANDING

GARY G. HENDRIX  
*Stanford Research Institute*

The semantic component of the speech understanding system being developed jointly by SRI and SDC performs two functions. It rules out those word combinations that are not meaningful, and it produces a semantic interpretation for those combinations that are. The semantic system described in this paper consists of a semantic model embodied in a network and a number of routines that interact with it. The model may be characterized as a description of objects, actions, and relations in the world. The semantic network encoding the model is partitioned into a set of hierarchically ordered sub-nets. This partitioning facilitates the categorization of objects, the encoding of quantification, and the maintenance of multiple interpretation hypotheses during parsing.

Interacting with the semantic network is a set of routines, associated with a set of language definition rules, that combine utterance components into larger phrases. In the course of their operations, these routines consult network descriptions of prototype situations and events and reference data describing how surface cases may be mapped into network representations. The output from these routines is a semantic network fragment consisting of several sub-nets. At the utterance level, the composite of these sub-nets is a complete network description of the semantics of the utterance, while the hierarchical ordering of the sub-nets reflects the syntactic composition of the input.

SPS: A FORMALISM FOR SEMANTIC INTERPRETATION  
AND ITS USE IN PROCESSING PREPOSITIONS THAT REFERENCE SPACE

NORMAN K. SONDHEIMER  
*The Ohio State University*

This paper presents a formalism, called SPS, for writing semantic processors for natural language understanding systems. SPS is intended for use in turning underlying syntactic structures in the form of constituent structure trees into underlying semantic structures in the form of nets composed of PLANNER-like assertions. The formalism is based on Woods-style "pattern→action" rules. The pattern element specifies tree fragments and various types of selectional restrictions. On the action side a variety of devices, including the use of registers, allow common reference to entities in the assertions produced. The registers used for reference can also be used to specify selectional restrictions across rules and for establishing default conditions for handling semantic ellipsis. Finally, SPS provides a control structure for the ordering of the application of the rules that interpret constituents and to control, in part, where the tree fragments are matched.

The power of SPS is seen in its unique ability to allow for the development of Case structures, especially the structures connected with the English prepositions that reference location, orientation and motion in space. These forms have always been troublesome for Case systems. Particularly difficult are the facts that 1) more than one of these prepositions can appear in a sentence in the same role, 2) their appearance can correspond to the need for multiple predicate semantic structures, and 3) they exhibit complex distributional and semantic relations among themselves and with respect to other sentential elements. SPS can allow for each of these phenomena.

An interpretation system for SPS has been implemented in LISP 1.6.

## THE NATURE AND COMPUTATIONAL USE OF A MEANING REPRESENTATION FOR WORD CONCEPTS

NICK CERCOME  
*University of Alberta*

The proposition-based semantic network notation of Schubert is especially well suited for including pragmatic and semantic information as part of the meaning representation of individual word concepts. Representations are networks based on propositions that consist of an n-ary predicate with a finite number of arguments. Terms are used to represent a given word concept can also be represented by semantic networks; there is no insistence that a given set of "primitives" comprise the meaning of a word.

Implication templates stored with the network help identify arguments that we expect to find in the surface utterance, they make the most commonly used inferences part of the meaning representation of a word concept.

Whenever the meaning representations of several word concepts are similar, we can extract the similarity and use that part of the network as a higher or more general level concept. More efficient memory storage utilization is obtained. Implication templates that involve only the extracted propositions are stored with the higher level concepts.

The use of networks has several consequences. A natural hierarchy of levels of analysis is suggested, the terms used in the network representation for another term are themselves represented and explicated through networks. Heuristic algorithms and other organizations can be superimposed to take advantage of special situations. The representation is suggestive of the meaning of a term independently of the routines used to process it. Binary decomposition trees are neither suggestive of the type of processing required nor of how they are internally consistent.

## ESTABLISHING CONTEXT IN TASK-ORIENTED DIALOGS

BARBARA G. DEUTSCH  
*Stanford Research Institute*

Task-oriented dialogs comprise conversation directed toward the completion of some task. For these dialogs, context is supplied both by the surrounding task (the task context) and by the surrounding dialog (the dialog context). This paper describes the discourse component of a speech understanding system for task-oriented dialogs. This component evaluates a proposed interpretation for an utterance in terms of how well it fits the context surrounding the utterance. In particular, the discourse component identifies the referents of noun phrases and fills in missing information in elliptical expressions. Task context is supplied by a model of the subtasks constituting a task and their relationships to one another. The dialog context is supplied by a history of the preceding utterances. In building a representation of the dialog context, the discourse processor takes advantage of the fact that task-oriented dialogs have a structure that closely parallels the structure of the task. A semantic network is partitioned into focus spaces with each focus space containing those concepts pertinent to the dialog relating to a subtask. The focus spaces are linked to their corresponding subtasks, they are ordered in a hierarchy determined by the relations of the subtasks to one another. Hence, this mechanism both supplies a dialog context and coordinates it with a task context.

## DISCOURSE MODELS AND LANGUAGE COMPREHENSION

BERTRAM C. BRUCE  
*Bolt Beranek and Newman Inc.*

Work of Goffman, Sacks, Labov, Schegloff, Searle, Schmidt and others has demonstrated the prevalence of higher order structures in communication. These structures are not just undifferentiated means for organizing discourse, but rather, essential carriers of information. As such, they must be included in any complete theory of language understanding.

This paper compares two approaches to modeling discourse. The first centers on the concept of a "discourse grammar" which defines the set of likely (i.e. easily understood) discourse structures. Participants in a discourse are then, in effect, following a path through the discourse grammar. A major advantage of the grammar approach is that it provides a relatively easy way to give an economical characterization of a wide range of discourse types. On the other hand, it sometimes makes faulty predictions from which it is difficult to recover.

The second approach is a "demand processing" model in which utterances create demands on both the speaker and the hearer. Responses to these demands are based on their relative "importance" the length of time they have been around, and conditions attached to each demand. The flow of responses provides another level of explanation for discourse structure.

These two approaches are discussed in terms of flexibility and efficiency. Finally, the paper considers their role in a more complete theory of discourse understanding.

JUDGING THE COHERENCY OF DISCOURSE  
(AND SOME OBSERVATIONS ABOUT FRAMES SCRIPTS)

BRIAN PHILLIPS  
*University of Illinois at Chicago Circle*

The surface form of any discourse is logically incomplete. The prime task for a reader is to fill in the gaps, if he is unable to do this, the discourse is incoherent.

Chomsky notes certain syntactically recoverable items, but the structure of discourse cannot be described by an extension of syntactic devices, the most perspicuous description can be made in terms of cognitive concepts. In such a model we can give canonical representation to facts which enable discourse with gaps to be understood. Omissions will be inferred from world knowledge.

The hypothesis is that a coherent discourse must be (a) connected and (b) have a single topic. My system has been applied to a conceptual analysis of some half-dozen examples of stories of accidental drowning written by students. The examples conform to a single abstract prescribed topical pattern that can only be recognized after inference of logical connections omitted by writers.

A conceptual encyclopedia in the system has frame/script-like structures. There are several possible objections to the systems of Minsky and Schank.

1. A lack of flexibility. The structures do not seem well suited to handling novel scenes.

2. There should be some bottom-up means of selecting a suitable frame which may then be used predictively.

3. Frame recognition must be recursive

In the present system, 1 is overcome by having groups of concepts of various degrees of abstraction. Abstract structures are located during analysis, this improves on 2. The relations between a complex concept and constituent concepts are explicit and can be used as in 3.

## AN APPROACH TO THE ORGANIZATION OF MUNDANE WORLD KNOWLEDGE: THE GENERATION AND MANAGEMENT OF SCRIPTS

R. E. CULLINGFORD  
*Yale University*

In understanding stories or natural language discourse, human hearers draw upon an enormous base of shared world knowledge about specific situations to help establish the needed background or context. Much of this knowledge appears to be episodic in nature, distilled from many experiences in common situations. Thus, for example, ordinary members of middle-class American culture have in common much information about mundane activities like going to birthday parties, restaurants, or supermarkets, simply because these things are done in much the same way all over the country. This paper presents an approach to the representation and management of this type of low-level world knowledge, based upon the concept of a situational script (Schank and Abelson). The application of scripts in story understanding will be illustrated via a working computer model called SAM (Script Applier Mechanism).

As implemented in SAM, a situation script is a data base comprised of interlinked causal chains describing the widely understood paths and turning points encountered in stories about mundane activities like eating in a restaurant or riding a subway. The process of story understanding begins with the construction of a 'trace' through this structure which contains the explicit input, other events not mentioned but commonly known to have happened, the more important enabling or resultative inferences to be drawn from the events, and the causal links connecting them. This trace or scenario is then examined by programs which generate summary, paraphrase or question-answering output. In complicated stories requiring several scripts, SAM handles the invocation and disabling of sequential, parallel and nested scripts.

## THE CONCEPTUAL DESCRIPTION OF PHYSICAL ACTIVITIES

NORMAN BADLER  
*University of Pennsylvania*

In proposing conceptual theories of language it is easy to overlook a primary motivation for language: The description of sensory information. The visual system provides much of this input and we easily generate descriptions of the activities we perceive happening around us.

In the first part of this paper we outline a representation for objects and events which enables us to describe certain classes of changes in object attributes and relations in conceptual terms. The visual aspects of this problem, as well as the linguistic, are described in the thesis from which this paper is condensed. The methodology for the description of the motion of rigid or jointed objects in a simulated man-made environment utilizes simple "demon" procedures with a straightforward control structure to watch for semantically-significant situations or changes at several different levels. The organization of these levels is hierachic, depending on spatial trajectory and rotations, observer movement, and spatial context at the low levels, directional prepositions and adverbs (adverbials) at the intermediate level, and motion verbs at the highest level.

In the second part of this paper we describe the event building algorithm which essentially fills in "deep" cases of a generic motion verb, such as Schank's "PTRANS" or Miller's "TRAVELS". In this sense the resultant descriptions are compatible with current paradigms of natural language understanding systems (for example, Rumelhart et al.). Moreover, the event construction semantics are based on physical activities so that conceptual events can be related to actual occurrences. Examples are given for descriptions obtained for a few scenes.

## A FRAME ANALYSIS OF AMERICAN SIGN LANGUAGE

JUDY KEGL

*Massachusetts Institute of Technology*

NANCY CHINCHOR

*University of Massachusetts*

Due to the fact that frames were first used explicitly in terms of visual imagery, the clearest examples of frames are those dealing with visual information. In order to see how a frame analysis of language would work, we thought it would be instructive to examine American Sign Language, a multidimensional spatial language. Our major aim is to use the computational formalism of frames in order to reach a more sophisticated understanding of American Sign Language. In so doing, we have been forced into certain definitions of computational concepts which we hope will be helpful in the general computational work on natural language.

Definitions of such computational concepts as perspective, scenario, frame, and prototype are given in the spirit of Minsky and, more primarily, Winograd. These definitions lead to a better understanding of how signs are abstracted from the perception of an event in order to communicate that event. From the linguistic standpoint, we focus mainly on the concept of verb and the processes of indexing and referring to objects in ASL. A by-product of these findings is a further clarification of the distinction between mime, statement, and discourse in ASL. The data which we will present as an illustration for our work is a telling by a native speaker of ASL of the story of "The Three Little Pigs" which we have on voice-over videotape.

## HOW DOES A SYSTEM KNOW WHEN TO STOP INFERENCE?

STAN ROSENSCHEIN  
*University of Pennsylvania*

Natural language processing systems that are sensitive to semantics and pragmatics generally draw 'inferences'; the presence of certain 'thoughts' or 'beliefs' triggers the retrieval and/or construction (usually pattern-matched) of other related thoughts or beliefs (the inferences). The problem we attack is this How can this process be controlled?

Various approaches are possible. One might use external criteria, such as ordering the inferences associated with a given antecedent, attaching 'probabilities' to the inferences, having the control program set an arbitrary limit on the length of inference chains, etc. The drawback of this approach is its arbitrariness. Alternatively, one could choose a strongly goal-oriented approach, however it is not clear how this approach might be reconciled with the data-driven (bottom-up) nature of free inferencing.

We have been looking at inferencing as an operation depending on the whole set of beliefs. The purpose of the operation is to find the least extension of the set which causes the beliefs to cohere, that is, to satisfy some pre-defined pattern. The set of patterns has to have some well-defined structure, it is then possible to define an inference operation in which there is no sharp distinction between antecedent and consequent, a pattern is a collection of subpatterns each of which may serve as an antecedent on one occasion and as a consequent on another

We have viewed the problem of restraining the process of inference as essentially one of making precise the idea of minimal unifying structure for a set of beliefs in such a way that internal (rather than external) criteria are established for inference cut-off. The ideas will be illustrated by various examples of inferring event descriptions, including descriptions of speech acts.

## CROSS-SENTENTIAL REFERENCE RESOLUTION

DAVID KLAPPHOLZ AND ABE LOCKMAN  
*Columbia University*

The problem of cross-sentential reference resolution involves the determination of the normally selected referents of pronouns across sentence boundaries, both in the absence and presence of "referent-forcing" context (sentences 1 and 2 vs. 3 and 4)

1. Yesterday a group of boys ran after a pack of dogs, the largest one broke a leg.
2. Yesterday John chased Bill half a block, he was soon out of breath.
3. The wild dogs outside our village all seem to suffer from a bone-weakening disease. Yesterday a group of boys ran after a pack of dogs, the largest one broke a leg
4. My friend Bill has a severe case of asthma. Yesterday John chased Bill half a block, he was soon out of breath

Note that in the former cases the preferred referents seem to be the surface subjects of the first sentences of each pair.

A further ramification of the problem is the determination of the relationship which the reference bears to the referent when the former is other than a pronoun (sentences 5 and 6)

5. John went for a long walk yesterday, the park was all abloom.
6. I met a lovely family yesterday, the father is a computer scientist.

Here the problem is to determine that "the park" is an area through which John took a long walk, and that "the father" is the father of the lovely family.

An attempt is made to formalize the notion of cross-sentential "focus", this notion is incorporated into an attempt to devise a general algorithm for establishing cross-sentential referents and their relationships to their references in the context of a primitive-based, inference driven model of natural language conversation. The question of properly directing an inference mechanism through a large base of world knowledge in the solution of the reference problem is discussed, and partial solutions presented

DEVELOPING A COMPUTER SYSTEM  
FOR HANDLING INHERENTLY VARIABLE LINGUISTIC DATA

DAVID BECKLES, LAWRENCE CARRINGTON, AND GEMMA WARNER  
*The University of the West Indies*

Linguistic communication in Trinidad and Tobago is characterized by use of varieties of English and varieties of Creole English in a sociolinguistic complex that appears similar to what has been described in Jamaica and Guyana as a post-creole dialect continuum. A host of pedagogical problems result from the absence of adequate description of the language system and the mismatch between the socio-linguistic facts and instructional methodology. The tape-recorded speech of a sample of children (aged 5-11+) is being analyzed to determine

- (a) the structure of their language,
- (b) the correlation of socio-linguistic factors with structures,
- (c) their progress in the acquisition of English

Given the inherently variable nature and volume of the data, manual counting of features and correlation with factors is not feasible. This paper is concerned with the development of a computer system for handling such data. Because of the difficulty of performing linguistic analyses by computer, the system is designed to deal with manually codified data, the results of such coding being among other things derivational trees with associated grammatical and semantic information. Since the communication complex does not have readily identifiable norms, the analytical method and matching computer system have to effect recognition of stable sub-systems (regardless of which set of external criteria constitute the determinants of these sub-systems) as well as state the evolution of the children's language. The computer system takes as input the derivational trees with associated grammatical information and semantics and classifies them in a fashion that allows the output stated above.

## A NATURAL LANGUAGE PROCESSING PACKAGE

DAVID BRILL AND BEATRICE T. OSHIKA  
*Speech Communications Research Laboratory*

A set of SAIL programs has been implemented for analyzing large bodies of natural language data in which associations may exist between strings and sets of strings. (A file containing parallel orthographic, phonemic, and phonetic transcriptions of a discourse would be an example of this type of data.) These programs include facilities for compiling information such as frequency of occurrence of strings (e.g. word frequencies) or substrings (e.g. consonant cluster frequencies), and describing relationships among strings (e.g. various phonological realizations of a given word)

Also, an associative data base may be interactively accessed on the basis of keys corresponding to the different types of data elements, and a pattern matcher allows retrieval of incompletely specified elements. For example, a pattern specifying the sequence

voiceless stop - /i/ - voiceless stop

can be used to retrieve the strings

'keep'	/kip/
'peeking'	/pikɪŋ/
'repeated'	/ripitəd/

from orthographic and phonemic transcriptions

Applications of this natural language processing package will be demonstrated. These include

- a) analysis of systematic phonological variation which serves as the basis for specifying and testing phonological rules,
- b) interactive testing of word recognition error rates associated with indeterminacy in the phonological or orthographic string,
- c) analysis of phonotactic patterns which can be used as the basis for specifying and testing syllabification algorithms,
- d) comparison across languages or dialects, to discover systematic sound correspondences, and to aid in the study of historical reconstruction or dialect relationships

## ON THE ROLE OF WORDS AND PHRASES IN AUTOMATIC TEXT ANALYSIS

G. SALTON  
Cornell University

Automatic indexing normally consists in assigning to documents either single terms, or more specific entities such as phrases, or more general entities such as term classes. Discrimination value analysis assigns an appropriate role in the indexing operation to the single terms, term phrases, and thesaurus categories. To enhance precision it is useful to form phrases from high-frequency single term components. To improve recall, low-frequency terms should be grouped into affinity classes, assigned as content identifiers instead of the single terms.

Collections in different subject areas are used in experiments to characterize the type of phrase and word class most effective for content representation.

The following typical conclusions can be reached:

- a) the addition of phrases improves performance considerably;
- b) use of phrases is better with corresponding deletion of single terms in practically all cases;
- c) the use of both high-frequency and medium-frequency phrases is generally more effective than the use of either phrase-type alone;
- d) the most effective thesaurus categories are those which include a large number of low-frequency terms;
- e) the least effective classes either consist of only one or two terms, or else they include terms with unequal frequency characteristics permitting the high-frequency terms to overcome the others.

The discrimination value theory is developed and appropriate experimental output is supplied.

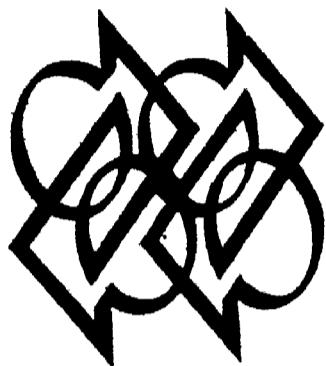
## GRAMMATICAL COMPRESSION IN NOTES AND RECORDS: ANALYSIS AND COMPUTATION

BARBARA B. ANDERSON            IRWIN D. J. BROSS            NAOMI SAGER  
*University of New Brunswick    Roswell Park Memorial Institute    New York University*

All languages have mechanisms of compression. One situation where these are used is when people are making notes. Usually, such note taking occurs within a practical context where the objects and meanings are known; the linguistic forms are degenerate but the message is unambiguous. This paper describes the linguistic mechanisms of compression which achieve this result, as they appeared in a study of the notes used in medical records for collaborative study of breast cancer. The syntactic devices were found to be mainly deletions of fixed types. the deletion of words having a special status in the grammar of the whole language (e.g. the verb be); and the deletion in particular positions of words having a special status in the particular subject matter (e.g. in the medical sublanguage, the word physician). A linguistic description of the forms with deletion was made and sublanguage word classes were defined. To test the description, a subcorpus of the medical records (357 sentences and sentence fragments on Xray findings) was parsed by an existing computer parsing system, using an English grammar to which a small component covering the deletion-forms was added. The paper concludes with a discussion of the modifications required in the computer grammar to parse the deletion forms and a summary of the parsing results.

**AMERICAN SOCIETY  
for  
INFORMATION SCIENCE**  
**38<sup>th</sup> ANNUAL MEETING**

**SHERATON-BOSTON  
OCTOBER 26-30, 1975**



SUNDAY, OCTOBER 26, 9 AM TO 5 PM: WORKSHOPS  
(Separate registration required)

A. MANAGEMENT OF THE LIBRARY IN TRANSITION: COSTING  
ANALYSIS AND MANAGEMENT TECHNIQUES FOR THE ADMINISTRATOR  
OF AUTOMATED SYSTEMS (BEGINS OCTOBER 25)

B. AUTOMATED TEXT PROCESSING AND PHOTOCOMPOSITION

C. MICROGRAPHICS IN INFORMATION SYSTEMS: NEW APPLICATIONS

10-4:30 PSYCHOLOGICAL RESEARCH ON USER ON-LINE INTERACTION

10-5 THE FUTURE OF SCIENTIFIC COMMUNICATION--WORK IN PROGRESS

1:30-4:30 NUMERICAL DATA

1:30-4:30 ORGANIZATION AND DISTRIBUTION OF PRIMARY SOCIAL SCIENCE  
DATA IN A COMPUTERIZED SETTING

MONDAY, OCTOBER 27

9:15 OPENING SESSION Ruth Tighe, Chairman

9:45 1965-1975--A DECADE OF INNOVATION Carlos Cuadra, Chmn

1:15 INTERDISCIPLINARY ISSUES CONCERNING THE USER/COMPUTER  
INTERFACE

## ASIS 1975 MEETING PROGRAM

36

- 1:15 STATE LEGISLATIVE INFORMATION SYSTEMS
- 1:15 COPYRIGHT 1975: A SEMI-STRUCTURED FREE-FOR-ALL
- 3:30 IMPROVEMENT OF METHODS FOR FORECASTING INFORMATION REQUIREMENTS AND SERVICES
- 3:30 TRENDS IN CLASSIFICATION
- 3:30 CONTRIBUTIONS TO INFORMATION IN THE BEHAVIORAL AND SOCIAL SCIENCES
- 5:30 OPPORTUNITIES IN INFORMATION SCIENCE TODAY

## TUESDAY, OCTOBER 28

- 9:15 KEYNOTE SESSION Daniel Bell and panel
- 1:15 PLANS AND FUNDING FOR NATIONWIDE INFORMATION PROGRAMS
- 1:15 SPECIAL PROBLEMS IN STORING AND RETRIEVING HUMANITIES MATERIALS
- 1:15 NATIONAL INSTRUCTIONAL MATERIALS INFORMATION SYSTEM
- 1:15 INFORMATION ANALYSIS CENTERS
- 3:30 FEDERAL LEGISLATIVE INFORMATION SYSTEMS
- 3:30 COMPARISON OF SYSTEM REQUIREMENTS FOR ON-LINE AND BATCH RETRIEVAL
- 3:30 EDUCATION INFORMATION RESOURCES--AFTER ERIC, WHAT?
- 3:30 MICROFORM CATALOGS--NOW MORE THAN EVER
- 8:00 REPORT OF ASIS LONG-RANGE PLANNING COMMISSION

## WEDNESDAY OCTOBER 29

- 9:15 MICROPROCESSORS AND RECENT TECHNOLOGICAL INNOVATIONS
- 9:15 A LOOK AT SATELLITE-MEDIATED MEDICAL COMMUNICATION EXPERIMENTS

ASIS 1975-MEETING PROGRAM

37

- 9:15 TOWARD A UNIFIED THEORY OF INFORMATION
- 9:15 INTERACTIVE SYSTEM EVALUATION: METHODOLOGY AND RESULTS
- 1:15 INTERACTIVE SYSTEM EVALUATION. PSYCHOLOGY RESEARCH  
METHODS AND RESULTS
- 1:15 LIBRARY NETWORKS: ORGANIZATION AND GOVERNANCE
- 1:15 REPORTS OF NATIONAL AND INTERNATIONAL EDUCATION PROGRAM  
STUDIES
- 1:15 REPORT ON ASIS PROJECT TO INVESTIGATE THE PLANNING REQUIRE-  
MENTS OF THE SCIENTIFIC AND TECHNICAL INFORMATION COMMUNITY
- 1:30 ERIC DATA BASE USERS CONFERENCE
- 3:30 BUSINESS MEETING Dale Baker, President
- 7:00 AWARDS BANQUET Jules Bergman, ABC News

THURSDAY, OCTOBER 30

- 9:15 TELEPROCESSING AND INFORMATION NETWORKS
- 9:15 INTERACTION BETWEEN PUBLIC AND PRIVATE SECTORS IN INFORMA-  
TION POLICY FORMATION
- 10:00 OBTAINING QUICK, INEXPENSIVE HARD COPY FROM MICROFILM AND  
COMPUTER SYSTEMS
- 10:00 CLINICAL DATA, MEDICAL RECORDS, AND THE NEW TECHNOLOGY
- 1:15 PRICING--A RATIONAL MECHANISM FOR ALLOCATING INFORMATION
- 1:15 TRENDS AND CURRENT RESEARCH IN AUTOMATED LANGUAGE PROCESSING
- 1:15 FEEDBACK AND CONTROL: IS THE USE OF HUMAN BEINGS MORE  
HUMAN AFTER 25 YEARS?
- 3:30 DOCTORAL FORUM

3:30 AVAILABILITY AND ACCESSIBILITY PROBLEMS FACING THE  
INFORMATION USER

3:30 THE INFORMATION REVOLUTION IN MEDICINE

FEES AT CONFERENCE

ASIS MEMBER, FULL CONFERENCE	\$60.00
NON MEMBER, FULL CONFERENCE	\$90.00
STUDENT, FULL TIME, FULL CONFERENCE	\$10.00
NON MEMBER, ONE DAY	\$25.00

REGISTRATION BEGINS AT 2 PM ON SATURDAY, OCTOBER 25

ASSOCIATION FOR COMPUTING MACHINERY  
ANNUAL CONFERENCE 1975  
RADISSON HOTEL, MINNEAPOLIS  
OCTOBER 20-22

CONFERENCE SESSIONS

MONDAY 20 OCTOBER 9:00

Opening session Alan F. Westin, The next decade of the computer revolution. Privacy, participation, and power

MONDAY 20 OCTOBER 1:15

SIGCAS-1 Panel Information and public policy

SIGBDP-1 Panel Researcher/user dialogue on file design

SIGOPS Panel Analysis of memory management and operating systems

SICSOFT-1 Panel Software management

MONDAY 20 OCTOBER 3:30

Panel. Computer user and vendor legal issues

SIGMICRO-1 Tutorial Microprocessors

SIGPLAN-2 Panel Programming and its implication on programming languages

MONDAY 20 OCTOBER 8:00

Panel Dollars and sense of D P

SIGMICRO-3 Panel Microprocessors--chips to working systems

SIGPLAN-1 Debate Should high level languages be used to write systems software?

TUESDAY 21 OCTOBER 8:30

SIGGRAPH Panel Evaluating computer graphics systems organization

SIGBDP-2 Panel GUIDE efforts in data base management

SIGMICRO-2 Panel Microprocessors and their architectural implications

SIGPLAN-3 Papers: New ideas in programming language theory

SIGSAM-1 Tutorial: Presentation of symbol manipulation systems

SIGCUE-3 Panel: The role of instructional simulations

TUESDAY 21 OCTOBER 10:15

SIGCAS-3 Panel: Computers in the electoral process

SIGBDP-3 Panel: Performance measurement and data base design

SIGPLAN-4 Papers: New methods and techniques in programming  
languages

SIGSAM-2 Papers: Symbolic and algebraic manipulation

SIGCUE-4 Panel: Computer programming in mathematics

TUESDAY 21 OCTOBER 1:15

SIGCAS-2 Panel: Computers and public policy

SIGBDP-5 Panel: Human resource requirements for business  
application development

SIGCOMM-2 Panel: Providing user services in a computer network  
environment

SICSOFT-2 Panel: Programming environments

SIGMAP-2 Panel: Math programming data structures

SIGCUE-1 (CUE/CSE/CAS) Panel: Computers in teaching environment

STUDENTS: Special tutorial-seminar: Computer science and the  
future

TUESDAY 21 OCTOBER 3:30

SIGDOC Tutorial and Panel: Experience with HIPO

SIGBDP-4 Panel: EPOS systems in the distributing industry

SIGIR-1 Panel: Information networks

SIGMINI Panel: The challenge of minicomputer software

SIGMAP-1 Panel: Computational practice in mixed-integer programming

SIGCUE-2 (CUE/CAS) Panel: Certification of CS teachers for  
secondary schools

TUESDAY 21 OCTOBER 7:

SIGSAM-3 Demonstration: Symbol manipulation systems

WEDNESDAY 22 OCTOBER 8:30

SIGARCH Papers

SIGBDP-6 Panel: Innovations on computer charging mechanisms

SIGCOMM-1 Panel: Applications of distributed computing

SIGSIM Tutorial: Simulation: State of the art

SIGMOD Panel: Data base administration

SIGCUE-5 Panel: Computers and society

WEDNESDAY 22 OCTOBER 10:15

PANEL: Summary of CODASYL report on selection and acquisition  
of data base management systems

SIGBDP-7 Panel: Developing user-oriented business systems

SIGDA Panel: Human factors engineering issues in design automa-  
tion systems

SICSOFT-3 Panel: Software product assurance

SIGNUM Papers and Talks: Applications of numerical analysis

SIGCUE-6 Panel: Tutorial computing: the state of the art

WEDNESDAY 22 OCTOBER 1:15

SIGART-1 Papers

SIGBDP-8 Panel: Performance measurement

SIGIR-2 (IR/ARCH) Panel: Non-numeric processing and computer  
architecture

SIGBIO Panel: Computer systems and the quality of health care

SIGMETRICS Papers: Computer system performance analysis

SIGCUE-7 Panel: Computer managed instruction and guidance

WEDNESDAY 22 OCTOBER 3:30

SIGART-2 Panel: Artificial intelligence and perception

SIGBDP-9 Panel: Structure of future systems

PANEL: ACM self assessment

PANEL: Computers outside the fishbowl

SIGCUE-8 Panel: Time-sharing instructional systems in Minnesota

**SOLAR PROJECT TERMINATES**

ARPA support for the semantic bibliography and lexicography project at System Development Corporation is ending. Online access to the five SOLAR files through ARPANET ends September 15. Arrangements for printed distribution of data are being explored. Expressions of interest in receiving such listings can be addressed to Dr. Tim Diller, 3244 Butler Avenue, Los Angeles, California 90066.

**ENERGY INFORMATION TOOLS**

The Information Industry Association and the National Federation of Abstracting and Indexing Services will hold a Workshop November 10-11 at the Quality Inn Capitol Hill, Washington.

Lectures, discussions, and workshops with access to online services will treat the tools presently available. The fee is \$55. For information, call Paul Zurkowski, IAI, 4720 Montgomery Lane, Bethesda, Maryland 20014; 301-654-4150.

**COMPUTER TRANSLATION  
OF CHINESE JOURNALS**

Shiu-Chang Loh, Professor of Computer Science, began research on machine translation at the Chinese University of Hong Kong in 1969. The system presently in use requires pre-editing by high-school graduates with two weeks of special training; the output is direct from a lineprinter.

The project is accepting subscriptions to machine-translated editions of *Acta Mathematica Sinica* and *Mathematics in Practice and Theory*, both quarterlies, and to the bimonthly *Acta Physica Sinica*. The price for the quarterlies is \$40 surface, \$48 air; for the bimonthly, \$60 surface and \$72 air.

Professor Loh is considering publication of translated editions of several other Chinese journals and designing a program for English-to-Chinese translation.



AMERICAN SOCIETY FOR INFORMATION SCIENCE

1155 SIXTEENTH STREET NW WASHINGTON D C 20036 • Telephone 202/659-3644

L I B R A R I A N   O F   C O N G R E S S

ON MAY 23, A S I S SENT THE FOLLOWING TELEGRAM TO THE  
PRESIDENT OF THE UNITED STATES:

This is in regard to the position of Librarian of Congress as related to today's information age.

The management of the Library of Congress has become an extremely complex task. With more than 4,000 employees and an annual budget of nearly \$100 million, the administrative skills required to direct the operation and growth of our de facto national library are comparable to those required to manage a company that is roughly the size of the New York Times, or about twice that of the Washington Post.

When the Library of Congress was created in 1800, America was just entering the industrial age, and information was relatively unimportant. Less than half of the American people could read. The few libraries that existed were owned and used primarily by the aristocracy.

Today, virtually every community in the United States has a public library. News is transmitted instantly via satellite. Computer-based data bases have become the new storehouses of information, and information itself has become the key to power.

As the importance of information has changed, the functions and responsibilities of the Library of Congress have changed as well. Originally established as a facility for Congress, the

Library has accepted many additional responsibilities including

1. It serves many common services in the areas of technical processing and reference for libraries throughout the country.
2. It provides many user services that supplement local efforts of libraries.
3. It performs many services in the areas of technical processing and reference for other libraries.
4. It serves as the hub of the recorded knowledge system of society.

Many professionals who are familiar with the present functions and responsibilities of the Library of Congress believe that the organization's next head must be more than a scholar. He or she must be a skilled and accomplished administrator; familiar with the implications of major national issues in information science, able to evaluate, and take advantage of, emerging technologies; effectively deal with political and personnel problems; and able to satisfy the information requirements of educators, researchers, professionals, legislators, local libraries, and the public at large.

Members of the American Society for Information Science believe that, while academic scholarship should be considered for our next Librarian of Congress, it certainly is not totally sufficient.

Sincerely,

JOSHUA I. SMITH  
Executive Director  
A S I S

B I B L I O G R A P H Y :

WORKING PAPERS PUBLISHED IN 1974 - 1975

FONDAZIONE DALLE MOLLE

ISTITUTO PER GLI STUDI SEMANTICI E COGNITIVI

VILLA HELENEUM, CASTAGNOLA 6976, SWITZERLAND

Note: A = hard copies available

B = microfiches available

S = Computer Science Department, Stanford

\* = in preparation

1. Causality and reasoning. Roger C. Schank. A, B.
2. Computer generation of natural language from a deep conceptual base. Neil Murray Goldman. A, S.
3. Is there a semantic memory? Roger C. Schank. A, B, \*.
4. Computational understanding: Analysis of sentences and context. Christopher Riesbeck. A, B.
5. "He will make you take it back": A study in the pragmatics of language. Eugene Charniak. A, B.
6. Understanding paragraphs. Roger C. Schank. A, B.
7. Selezione di parole per l'estrazione di unità foniche atte alla sintesi delle lingue tedesca e italiana. G. B. Debiasi and A. M. Mioni. A, B.
8. Fonetica e fonologia autonoma della lingua Hindi. Romeo Galassi. A, B.
9. Ottimizzazione delle caratteristiche delle unità normalizzate per la sintesi del parlato mediante mini-computer. Mildonian Offeli. A, B.
10. Ueber die Struktur des semantischen Langzeitgedächtnisses. Manfred Wettler. A, B.

11. "Information Storage and Retrieval" modello di un sistema integrato. B. Treusch, F. Lestuzzi, S. Rova. A, B.
11. Konzepttheorie--Ein praktischer Beitrag zur Textverarbeitung und Textrezeption. Wolfgang Samlowski. A.
13. A partial taxonomy of knowledge about actions. Eugene Charniak. A.
14. Organization and inference in a frame-like system of common sense knowledge. Eugene Charniak. A.
15. Linguistischer Thesaurus. B. Treusch. A\*.
16. Semantics, preference and inference--A full description of a system and a program. Margaret King and Yorick Wilks. A\*
17. Seven theses on artificial intelligence and natural language Yorick Wilks. A\*.  
Proceedings of the Tutorial on Computational Semantics given by the members of the Institute of Semantic and Cognitive Studies.

C O N C E P T U A L   I N F O R M A T I O N   P R O C E S S I N G

ROGER C. SCHANK

*with contributions by*

NEIL M. GOLDMAN, CHARLES J. RIEGER III, & CHRISTOPHER K. RIESBECK

*Fundamental Studies in Computer Science, Volume 3*

NORTH-HOLLAND PUBLISHING COMPANY  
1975

1. MARGIE
2. The conceptual approach to language processing
3. Conceptual dependency theory
4. The conceptual analyzer (Riesbeck)
5. Conceptual memory and inference (Rieger)
6. Conceptual generation (Goldman)

viii + 384 pages

ISBN 0-7204-2507-7

Price: \$27.50

## AUTOMATIC TRANSLATION AT GRENOBLE

(La Traduction Automatique à Grenoble)

BERNARD VAUQUOIS

*Documents de Linguistique, 24*Dunod  
1975

Foreword . . . . .	9
FIRST PART: The CETA MT Experiment (1961-71) . . . . .	13
CHAPTER I: THE STATE OF MT IN 1961 . . . . .	14
1. The birth and spectacular growth of MT . . . . .	14
2. Diverse approaches to MT in 1961 . . . . .	16
3. Design of the Grenoble project . . . . .	31
CHAPTER II: MORPHOLOGICAL ANALYSIS . . . . .	35
1. Purpose of morphological analysis . . . . .	35
2. Organization of the morphological model . . . . .	37
3. Limits and peculiarities of the morphological model . . . . .	49
4. Sample of the results of morphological analysis . . . . .	50
CHAPTER III: THE MODEL OF SYNTACTIC ANALYSIS . . . . .	52
1. Purpose of syntactic analysis . . . . .	52
2. Syntactic analyzers . . . . .	61
3. The CETA model of syntactic analysis . . . . .	82
4. Examples . . . . .	97
CHAPTER IV: THE PIVOT LANGUAGE . . . . .	105
1. Choice of a level of transfer . . . . .	105
2. Definition of the pivot language . . . . .	106
3. Informatic methods . . . . .	112
4. Linguistic information . . . . .	116
5. Examples of representation in pivot language . . . . .	118

CHAPTER V: THE PROCESS OF GENERATION . . . . .	123
1. Generation of superficial syntactic structure . . . . .	123
2. Morphological generation . . . . .	128
3. Examples. . . . .	128
SECOND PART: The work of GETA since 1972 . . . . .	141
CHAPTER VI: REFLECTIONS ON AUTOMATIC TRANSLATION . . . . .	142
1. MT at the start of the '70s . . . . .	142
2. Critique of the CETA MT system . . . . .	143
3. Perspectives on MT . . . . .	148
CHAPTER VII: THE WORK OF GETA SINCE 1972 . . . . .	157
1. The new conceptions of informatic systems . . . . .	157
2. The ATEF system . . . . .	160
3. The CETA system . . . . .	163
4. Algogrammars and interactive methods . . . . .	168
5. The LEIBNIZ group . . . . .	171
<i>Bibliography</i> . . . . .	173

## FREQUENCY AND DISTRIBUTION

*Fréquence et distribution du vocabulaire dans un choix de romans français*

GUNNEL ENGWALL

Språkförlaget Skriptor AB  
Stockholm 1974

1. Introduction . . . . .	9
2. Constitution of the corpus . . . . .	14
3. Units to be examined in a frequency study . .	24
4. Results on the level of occurrences and types	30
5. Methods for classifying words by order of importance . . . . .	43
6. Definition of the inflected form and lemma . .	56
7. Results on the level of inflected forms and lemmas . . . . .	69
8. Comparisons among five earlier studies and the corpus of tex novels . . . . .	85
9. Conclusion .	101
Appendices . . . . .	108
Abbreviations and symbols . . . . .	176
Index . . . . .	178
Bibliography . . . . .	181

BULLETIN

Volume 3 Number 2

Summer Term 1975

EDITOR:

Mrs. Joan M. Smith  
6 Sevenoaks Avenue  
Heaton Moor, Stockport  
Cheshire, England

CONTENTS

GUEST EDITORIAL . . . . .	J. R. Allen	95
THE ACTIVITIES OF THE INSTITUT FÜR DEUTSCHE SPRACHE'(IDS) MANNHEIM IN THE FIELD OF LITERARY AND LINGUISTIC COMPUTING . . . . .	Godelieve Berry-Rogghe	97
CARLYLE AND THE MACHINE: A Quantitative Analysis of Syntax in Prose Style . . . . .	R. L. Oakman	100
HOW TO USE COCOA TO PRODUCE INDEXES (TO BOTH BOOK AND SUB- ROUTINE LIBRARIES) . . . . .	Kathleen M. Crennell	115
COMPUTATIONAL LINGUISTICS OR WHAT'S IN A NAME? . . . . .	W. Martin	124
THE BILINGUAL ANALYTICAL LITERARY AND LINGUISTIC CONCORDANCE - BALCON . . . . .	Susan M. Hockey and V. Shibayev	133
TOWARDS AN ALGORITHMIC METHODOLOGY OF LEMMATIZATION . . . . .	M. L. Hann	140
THE BODLEIAN PROJECT: A COMPUTERIZED INDEX TO THE VISUAL CONTENT OF MANUSCRIPT ILLUMINATIONS . . . . .	Alice F. Worsley	151
A COMPUTER-ASSISTED PROJECT AT THE HEBREW UNIVERSITY, ISRAEL: THE COMPILATION OF AN INDEX TO THE WRITINGS OF AGNON U. Sakhm		156
THE USE OF A COMPUTER IN DEVISING A BEGINNERS' LATIN COURSE . . . . .	C. W. E. Peckett	158
TEACHING COMPUTATIONAL LINGUISTICS: A continuation of the Discussion . . . . .	Margaret King	161
LITERARY STATISTICS VI: ON THE FUTURE OF LITERARY STATISTICS . . . . .	N. D. Thomson	166
COMPTE RENDU D'UNE TABLE-RONDE DU C.N.R.S. SUR LE THEME: 'PROCÉDURES D'ANALYSE ET MÉTHODES DE VALIDATION DANS L'ÉTUDE DES DONNÉES TEXTUELLES' . . . . .	J. Virbel	172

WORD ORDER AND WORD ORDER CHANGE

CHARLES N. LI, EDITOR

*The University of Texas Press  
Austin  
1975*

REVIEWED BY JAMES M. DUNN  
*Princeton University*

This is a collection of twelve of the thirteen papers presented at the Conference on Word Order and Word Order Change that was held at the University of California, Santa Barbara, on January 26 - 27, 1974. The first eight deal with the diachronic aspect of word order, while the other four represent a synchronic treatment of the subject.

In the preface the editor acknowledges the influence of Joseph Greenberg on these proceedings. His 1961 paper, 'Some universals of grammar with particular reference to the order of meaningful elements', is seen as 'the starting point' for most of the papers in this volume.

The papers in this collection appeal to a great diversity of interests: sign language, languages of the Niger-Congo group, Chinese, Indo-European, drift, discourse grammar, metatheory, the evaluation metric, and, of course, language typology. Obviously, their common purpose is to move toward a clearer explanation of the causal relationships between the surface constituents of a sentence both synchronically and diachronically.

But many of the papers actually share more than the common denominator of interest in word order. At several points where other mutual interests overlap, the discussions assume the nature of a dialog (or, more often, a debate), and the reader finds transition from paper to paper relatively smooth.

I shall withhold further comment on the merits of this book as a whole until the conclusion of this review. To help the reader make his own evaluation and to guide him to topics of special interest I will present a summary of some of the essential points of each paper (with apologies to each author for any unintentional misrepresentation).

1. Influences on word order change in American Sign Language', by Susan Fischer (1-25). In American Sign Language (ASL) today the basic word order is SVO. Just one hundred years ago it exhibited a predominantly SOV word order. Fischer illustrates this with two texts relating to the story of the Prodigal Son (9):

(1871) Days few after, son younger money all take,  
country far go ...

(1970's) Later-on second-of-two young son decide, gather,  
pack, leave home, gone.

In ASL it is still possible, either in the case of a few idioms, in topicalization, or when the interpretation of a sentence would be unambiguous (e.g., 'the boy likes ice cream'), regardless of the order constraints on base forms, to find SOV -even OVS- arrangements.

The pressure that caused the shift from SOV to SVO Fischer attributes to factors of prestige and contact with English.

Evidently some critics regard sign language as a second-class language. Recognition of this status supposedly leads (in some vague way) to an imitation of the patterns of the dominant English language. More persuasively it is argued that deaf children learning to sign receive a mixed input of ASL and a signed version of English.

Fischer explains that the interpretation of a NNV-sequence today would be OSV rather than SOV. Hence, in representing a sentence such as, 'The girl kicked the boy', the sequence of gestures in the context of a discourse would be (19-20):

boy (here)	girl (here)	she-kick-him
right hand	left hand	left kicks right from direction of location of girl to location of boy
(patient)	(agent)	
(1) ←----- (2)	direction of movement	

This is the preferred (unmarked, and evidently more efficient) order.

Since sign language is a visual medium, the use of the space around the signer is important in indicating grammatical mechanisms. This function of space represents a countervailing force to the pressure from English word order patterns. Fischer suggests that since location is available to disambiguate the grammatical relations, non-reversible subject and object sequences may continue to occur in SOV and OSV orders (21).

This article gives the reader an immediate view of the broad spectrum of topics presented in this volume. Fischer's investigation is intriguing and very informative. One would

only wish for greater elaboration with more data on the interesting discussion of how ASL got (or is getting) to SVO.

2. 'Dynamic aspects of word order in the numeral classifier' by Joseph Greenberg (27-45), begins by presenting eight synchronic hypotheses about the numeral classifier construction that have been extracted from one of his earlier papers (1973). From these synchronic observations he suggests three diachronic hypotheses, briefly sketched here.

First, the classifier phrase is originally a Quantifier-Noun phrase 'with a particular syntactic use' (31). The Quantifier (Q) - Classifier (Cl) array as a favored sequence reflects characteristics of the Quantifier  $\leftrightarrow$  Noun (N) relationship in non-classifier languages.

Second, the order of the N in relation to the classifier phrase is often in the process of undergoing a shift. The Q  $\leftrightarrow$  Cl sequence remains relatively fixed.

Third, it is more likely that in such cases the earlier order is N - (Q  $\leftrightarrow$  Cl) rather than (Q  $\leftrightarrow$  Cl) - N.

The rest of the article is a consideration of evidence that tests the validity of these hypotheses. In Greenberg's words, 'the most cogent [evidence comes from] direct historical documentation' (31).

Evidence for the shift from posposed to preposed classifier is adduced from the history of several languages, including Chinese, Khmer, and Burmese. Gilyak shows the same shift across present generations.

Greenberg observes that while in phonology there is

independent evidence concerning the relative plausibility of historical changes, 'it is precisely the plausibility of [the hypotheses] which is at issue' (36). 'There are...cases...in which the evidence points to a historical shift from postposed to preposed position and no counterevidence. [However] it is not claimed that the construction always arises in the post-positive form. [There may be instances when the preposed form is found] in which there is nothing to show that it was ever otherwise' (38).

The author proceeds next to considering 'the factors involved in the synchronic favoring of the postposed classifier construction such that even consistent SOV languages with preposed nominal modifiers, such as Japanese, have postposed order as usual or exclusive' (38). In investigating the occurrence of variant orders Greenberg suggests looking for differences in function. Illustrating this point with examples from Standard Malay, Palaung, and Hungarian he shows that some quantifying expressions may be typed as prenominal, while others are adverbial in nature.

In viewing the classifier expression as an original quantifying phrase that 'serves as comment to the head noun functioning as topic' (41), Greenberg proposes that in these instances 'the use of a classifier...can be viewed as a device which avoids the bare predication of numerals which is disfavored in many languages' (41). This would seem to suggest that there is a comparison here to some generative accounts of the derivation of the adjective phrase: '1) predication, 2) relative clause,

3) adjective follows noun, 4) adjective precedes noun (41).

Greenberg does not claim that this should be taken as the model for a diachronic sequence.

3. 'Serial verbs and syntactic change: Niger-Congo', by Talmy Givon (47-112). The aim of this paper is to study two diachronic processes: 1) the demise of SOV syntax and its associated syntactic typologies in Niger-Congo, one of which is a specific type of verb serialization; 2) a process involving two mutually linked changes that combine to affect the lexico-syntactic typology of the language as follows: [a] 'the lexical re-analysis (or, 'grammaticalization') of verbs as prepositional case markers and [b] the correlated change from a serializing to a non-serializing VP typology' (49).

Givon presents evidence from the Mande, West Atlantic, (Voltaic) Gur, Benue-Kwa, and Bantu groups to reconstruct Proto-Niger-Congo as an SOV language. (section 2). In the section following the author discusses verb serialization (found mostly in the Benue-Kwa group) which he asserts is one of the major typologies that resulted from the shift away from SOV syntax. Givon notes that the synchronic analysis of serial verb constructions has been the subject of a long debate over the following issues: are these entities verbs or prepositions; if verbs, do they represent a coordinate or subordinate structure; does serialization arise diachronically from conjunction or subordination?

In section 4 Givon presents arguments to show that the verb-serializing languages of Benue-Kwa may be undergoing a

gradual syntactic-lexical change, from SVO verb-serializing syntax towards a non-serializing verb phrase in which erstwhile verbs are re-analyzed as prepositions' (80), or postpositions in Ijo. His criteria for the reanalysis are semantic (there is a depletion of some semantic material out of the erstwhile verb (82)); morphological (there is a loss of ability to take normal verb affixes (84)), and syntactic (after semantic reanalysis as a preposition or a conjunction a verb quite often remains at its original serial-verb position (84)).

Givon argues that a shift from serialization must be gradual: the morphological and syntactic behavior is likely to lag behind the more progressive semantic re-analysis' (86). Another type of argument Givon characterizes as 'rather futile' (86) is one undertaken by Hyman (1971b) and others for the coordinate diachronic origin of serial verbs. To Givon 'it is quite clear that languages do proceed to reanalyze semantically the relationship between two erstwhile coordinate ("consecutive") clauses so that eventually a non-coordinate semantics prevails' (87). Further on he states, 'the lexical-semantic re-analysis of verbs into prepositions in a serial-verb construction is likely to create semantically more complex verbs in all cases... and is also likely to introduce some SOV syntax into an erstwhile SVO-serializing language. But it is not likely to introduce a complete SOV syntax into the language' (89).

This paper is well written and most laudable for providing the reader with an abundance of data to illustrate the author's contentions. The paper that follows should be read to see

Hyman's response to Givón's claims.

4. On the change from SOV to SVO; evidence from 'Niger-Congo', by Larry Hyman (113 -147). Hyman, like Givón, focuses his attention on the Niger-Congo family of languages in an investigation of 'the various factors which may contribute to the change from an SOV to an SVO word order' (115). Hyman discusses the following four 'explanations' for word order change: 1) contact, 2) disambiguation, 3) grammaticalization, and 4) afterthought.

While acknowledging that contact is often responsible for word order change, Hyman prefers to leave it aside, reasoning that a diachronic search for lost contact languages to explain a change might prove fruitless. On disambiguation as an explanation Hyman cites Vennemann (1973a) in which it is contended that 'word order changes result from the leveling of morphological case markings, which in turn are lost through phonological change' (116). Vennemann's model is rejected because it is not readily adaptable to the facts of Niger-Congo, since Proto-NC was not characterized by case markings on nouns' (123).

Grammaticalization as proposed by Givón (this volume) is also rejected (124)

However, since Givón is correct to point out that Proto-Bantu did not serialize verbs (though an earlier ancestor may very well have involved serial verbs), the grammaticalization of verbs to postpositions could not have caused the change of SOV to SVO in Bantu--where, recall from the preceding section, the whole thing is presumed to have started. We must therefore conclude that grammaticalization plays little if any role in the word order changes discussed in the first part of Givón's paper.

Hyman concludes that afterthought is the best explanation to support the evidence from Niger-Congo.. Afterthought, a cover term for a number of different-though related phenomena' (119), is to be understood as an aspect of the 'conflict between syntax and pragmatics. That is, speakers, in the course of using a language sometimes find it necessary to break the syntax and add grammatical elements in positions where they normally should not appear' (119-120). Hyman takes evidence from Kru and Kpelle to support his claim. He writes (135-136):

The reason why [afterthought] hits the sentence first is because of the magnitude of the problem of afterthought -- i.e., the units which can serve as afterthoughts are simply larger in scope, more likely to be forgotten. Thus, if afterthought is to lead to a rearrangement of the syntactic units, it will take place historically first in the change from SOV to SVO, and then in the change from Mod-N and N-Mod, as was seen in the two separate syntactic waves which hit Kwa territory (section 3.1).

Hyman devotes the fourth section of his paper to a reply to Givón's treatment of serialization. Givón, he writes, attributes the rise of serialization to a response to the loss of case markings on nouns. Citing the replacement of a case marker or a preposition expressing instrumentality with a verb such as 'to take' as in, 'take the knife and cut the meat' (138), Hyman says there is no disruption of the instrumental meaning. This would then show English to be a serializing language and 'the distinction between serializing and not serializing becomes trivial, if not nonexistent' (138). Further on he asserts that serialization does occur in SOV languages, e.g., Laku and Japanese. 'It doesn't occur in too many African languages, because the only SOV language in the serialization belt is Ijó.'

And there it occurs' (141)

Hyman's paper is valuable for his insights into the notion of afterthought - vague though it may be. His refutation of Givon's position is not as convincing.

5. 'A discussion of compound and word order', by Winfred P. Lehmann (149-162). Lehmann offers what might be described as the 'keynote' paper of the conference. While the purpose of his effort is to 'examine the position of nominal elements of verbs' (151), (the data comes mainly from Sanskrit, and focuses on Proto-Indo-European), the reader is impressed by the hortatory ring of the prose. In fact when some of the sentences are extracted from the article and displayed in a list (incomplete) as below, they read like maxims.

1. The time has come to set up universal laws of language development, if cautiously (151).
2. We should state our procedures and abide by them (151).
3. We seek an understanding of syntactic phenomena by practicing comparison to determine universal laws, combining such comparison with philological study and historical comparison... (151).
4. In studies concerning universals of language we generally start from an examination of data and then ask questions regarding the data... In dealing with such questions we must examine the data in accordance with a model of language, and in accordance with specific principles that have been observed regarding linguistic structures. Moreover, we must realize that languages are historical products (151-152).
5. [A]ny hypothesis of syntactic change must be framed in accordance with a strict framework... [T]he question which may be the most pressing in historical linguistics at present [is] identifying the events and structures resulting when a language undergoes syntactic change (154).
6. The processes of syntactic change, and the influences

proposed for it, must be determined by observing what happens to languages in transition (155).

7. When a language is undergoing syntactic change, some of its characteristics must be modified before others (155).

I would urge the reader to turn to this article first because it captures the spirit that has animated the other contributors in their endeavors. This assessment of Lehmann's paper is not intended to diminish interest in his well-articulated factual discussion. To the contrary. But it is outstanding in its general appeal, and must be read for that, if for no other reason.

6. 'The semantic function of word order: a case study in Mandarin', by Charles N. Li and Sandra A. Thompson (163-195). This is a study of the semantic function of word order with respect to definiteness in Mandarin Chinese. The authors present evidence 'to demonstrate that definite nouns, whether subject or object, tend to be placed before the verb, whereas indefinite nouns tend to follow the verb. [They contend that] this function of word order was developed in the past millennium and that, as a relatively new grammatical device, it is in conflict with the shift from SVO to SOV - a diachronic process presently in action. [They claim that their analysis] will indicate that this conflict is most likely to be resolved in favor of the shift to SOV word order' (165-166).

The authors assert that their evidence suggests the following generalizations and associated 'refinements' or modifications:

Tendency A

Nouns preceding the verb tend to be definite, while those following the verb tend to be indefinite (170).

Refinement 1

The noun in postverbal position will be interpreted as indefinite unless it is morphologically or inherently or non-anaphorically definite (173).

Refinement 2

A sentence-initial noun must be interpreted as definite, and may not be interpreted as indefinite even if it is preceded by the numeral yi 'one' (177).

Refinement 3

The noun following bei [an agent marker], although pre-verbal, is immune to Tendency A (179).

Refinement 4

Nouns in prepositional phrases are immune to Tendency A (182).

Tendency B

Mandarin is presently undergoing a word order shift from SVO to SOV (185).

Evidence for the hypothesis stated as Tendency B is:

- 1) the ba- construction that allows SOV word order is becoming more extensive (187-188); 2) in modern Mandarin the demonstrative article, nei- 'that' and the numeral, yi- 'one' may serve as definite and indefinite articles respectively in subordinate clauses, indicating a gradual trend (188).

This paper is exemplary in its orderly presentation and strong empirical orientation. It represents a continuation of similar studies undertaken by Li and Thompson.

7. 'On some factors that affect and effect word order', by Susan Steele (197-268). Steele claims that in her survey of the position of grammatical modal elements in 44 languages she has found them to be 'ordered with respect to the other elements of the sentence in a regular fashion' (199). She classifies

languages into two types with respect to the position of modals. Type A languages, where modals are dependent on the main verb, commonly show the following word orders (218)

S Modal VO  
SOV Modal  
Modal VSO

In Type B languages 'modals tend to occur in the sentential second position' (221). In this group modals seem to be defined solely by their sentential position.

Steele observes that there are two major theories about grammatical modals. One holds that 'modals are generated in the deep structure in the position in which they occur on the surface, dominated by the category symbol, Modal' (222). The other theory derives modals as main verbs. Since neither theory can completely explain the positional tendencies she describes, Steele suggests a third alternative. She maintains that the position of modals in the surface structure is dominated by two factors:

1. There are certain unmarked surface positions for modal elements. In verb-initial languages this is the sentence-initial position; in verb-final languages, sentence final position (223).
2. The unmarked positions are acted upon by two tendencies—the tendency for certain elements (including, but not exclusive to, modal elements) to be attracted to the verb and the tendency for these same elements to be positioned initially (224).

Steele hypothesizes, 'the importance of the sentence-initial position is related to a strategy that psychologists have called "primacy" [by which] the first element in a series is perceived to be the most important' (235).

Further on she states that the assumed tendency of modals

to sentential second position is a function of the importance of first position. Her conjecture is:

1) of all the elements (topic, negative, past tense, quotatives, and modal elements) that may be attracted by sentence initial position, the attraction for topic is the strongest; 2) topic may solidify in sentence initial position, thus forcing all of the other elements to sentential second position (238-239).

Steele concludes that 'the multi-purpose importance of first position will force grammatical elements - and verbs - out of first position and topic (developing to subject) in (243).

Steele's occasional mention of psychological strategies reminds us that exploration into the relationship between word order and cognitive strategies is still tentative but would doubtless add considerable explanatory power to the observations made by linguists.

8. 'An explanation of drift', by Theo Vennemann (269-305) begins with a review of past discussions of drift in Sapir(1921), Fries (1940), and Lakoff (1972). Sapir's identification of three 'major drifts' (leveling of distinction between subject and object cases, tendency to fixed position in the sentence, and the drift toward 'the invariable word' (272)) shows some shortcomings. It is anglocentric, uses few examples, and appears 'impressionistic' (276). Nevertheless, the studies by Lakoff and Fries represent regressions from the advanced position taken by Sapir. Of Lakoff Vennemann is particularly critical for her 'amazing misrepresentations of Sapir's straightforward and insightful original account of drift' (286).

Greenberg and Lehmann in their numerous publications shun the term 'drift' but Vennemann notes that they investigate

phenomena closely related to it and have contributed much to its explanation.

In the last section entitled, 'The universality of drift: natural generative grammar' Vennemann discusses explanations for individual drifts and the literature associated with them. His discussion leads primarily to this conclusion (301):

Sapir was moving in the right direction when he established causal relationships among his individual drifts and viewed phonological change as the ultimate cause of drift. We are now, half a century after Sapir's exposition of the problem, in a position to make deeper and more comprehensive generalizations about the nature of phonological and syntactic change. This enables us to say that given the inevitability of neutralizing and reductive phonological change, and given the various, often conflicting demands of pragmatics and semantics on grammatical structure, drift is inescapable, and its course predictable.

This article deserves special attention for two reasons. One, drift as a plausible linguistic phenomenon has had a 'bad press' for too long. Vennemann synthesizes the findings of several scholars across a broad chronological spectrum to justify the validity of the concept of drift and to relate it to notions of linguistic universals. The second reason why this article is so commendable is that it is a satisfying reminder of the lasting value of the insights of Edward Sapir.

9. 'Order in base structures', by Emmon Bach (307-3 3).

Bach's paper is yet another example of the comprehensive scope of this collection. The author undertakes to present arguments in favor of an ordered base and to refute the claims made by the proponents of order-free theories of the base. This article is important because in Bach's words, it 'is particularly relevant to hypotheses about universal grammar' (309).

Bach notes in his introduction that because base structures are theoretical constructs they 'cannot be directly observed or intuited' (310). Therefore, the hypotheses on which these constructs are based must be scrutinized, since they are more accessible to empirical justification. He then procedes to outline different theories of the base (section 2); to examine some of the arguments that support the claim that base structures are unordered (section 3); and in the last section to present his refutation of the case for unordered base structures. Although Bach's empirical evidence is predominantly (but not exclusively) from English, the force of his arguments remains strong. Bach's conclusion is that the evidence suggests there is an inherent linearity in language at all levels, a condition which, if true, would weaken the claims of order-free theories.

Although Bach integrates some typological evidence to support his arguments, the discussion remains mostly on a metatheoretical level, which, of course, sets his paper apart somewhat from the general tone of the others. This detachment is desirable because it serves to bracket the studies presented here with a theory of grammar at the most abstract level. All of the contributors to varying degrees relate their studies to a theory of generative grammar. Because of the understandable limitations imposed by the subject matter of the previous papers, Bach's paper, as well as the three that follow it, offsets what might otherwise have represented a noticeable imbalance in this volume.

It is recommended that Sanders' paper be read in conjunction with Bach's. Sanders offers a somewhat negative critique quoted in part here (401n):

In Bach's case all that is shown is that there are certain facts about certain languages that appear to be consistent with the hypothesis of variable ordering. It is not shown that these facts are inconsistent with the hypothesis of invariant ordering; or with any other principle of the theory of Derivational Ordering. It is the latter, of course, that must be demonstrated, and not merely the former, if one wishes to support the claim that invariant order theories are false or inadequate.

10. 'The presentative movement or why the ideal word order is V.S.O.P.', by Robert Hetzron (346-388). Hetzron's introductory argument goes as follows. In a discourse no sentence is uttered in a vacuum. Not only are the preceding discourse and situation important in the context of an utterance, but also he notes that any given sentence may figure prominently in the background of subsequent sentences. When a sentence is constructed so that a certain component of it will be 'given a status of prominence in short-range memory, so that it will dominate the immediate sequel to that particular utterance' (347), the motivation for this promotion to prominence is called the 'presentative function'. In the derivation of a sentence elements marked by the presentative function often end up in a sentence-final position. This 'transfer of presentative elements to the end of the sentence' Hetzron calls the 'presentative movement' (348).

Hetzron collects evidence mainly from English, Hungarian, and Amharic to demonstrate how the presentative movement brings certain elements to 'a sentence-final, or at least to a later

than usual, position' (374). He argues clearly and persuasively in showing the existence and operation of the presentative movement, but leaves the reader uncertain about why, as the title suggests, the ideal word order is V.S.O.P.

The article makes an interesting contribution to this volume because the presentative function belongs to discourse grammar which operates on somewhat less exact, less strict principles than sentence grammar' (376). Hence, when it comes to making claims for the universal status of the presentative movement the prose becomes equally inexact: We can state that the presentative movement is a universal tendency potentially always present in the speech system of humans, applying whenever there is an opportunity' (376). Hetzron says universal tendencies play a weighty role in discourse grammar (376), though he does not say how. On the role of the presentative function in historical change the discussion becomes even more tentative (377):

The presentative...shows up in all cases where it has been given a chance to influence the direction of historical change. Once it has managed to become part of a particular grammar, it tends to persist, notwithstanding the erosive effect of later historical developments, as in Amharic (Section 5.). In other cases it succeeds in sneaking in the back-door, as in cataphoric predictions where the presentative element has to be promoted to the status of predicate to attain the final position.

11. 'On the explanation of constituent order universals', by Gerald Sanders (389-436). The purpose of this article is to show how 'any serious attempt to achieve even the lowest order of explanation [of all significant facts and generalizations about the subject matter of our discipline] requires the

assumption of numerous precise and highly restrictive meta-constraints on natural-language grammar, meta-constraints which have far-reaching implications and interrelations with respect to all aspects of phonology and syntax' (405-406).

Sanders is particularly critical of statements typically found in the literature that he labels 'gross numerically unspecified likelihood assertions' such as (393):

In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object. (Greenberg 1963 1962 :61)

Such statements (he cites other linguists as well) are 'too vague and unnecessarily elaborate to be really useful even as statements of mere description. They have no possible predictive or explanatory uses at all' (394). If linguists are to act on their commitment to hold to the rigorous standards imposed by an empirical science, their metatheoretical and methodological prerequisites must make it possible to establish 'the scientifically indispensable implication relations that must hold between empirical hypotheses and factual observations that would suffice to confirm or disconfirm them' (428).

Sanders sees his objective in this article is 'to exemplify the complexities and ramifications attendant upon any serious attempt to explain constituent ordering in natural language' (429). He develops his arguments 'primarily with respect to the highly restrictive theory of Derivational Ordering...and the meta-constraints that comprise this theory--the principles of Terminal Completeness and Invariant Ordering' (429). Sanders specifically deals with natural language data concerning (for example) the

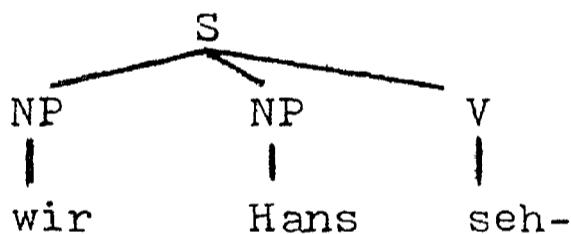
ordering of oblique arguments, adjectives, and nominal modifiers. To generate the most general explanations of all the facts about the ordering of these elements, the author develops his case for 'the grammatical law of Specificity Preposition' Sanders shows that the importance of this 'law' lies in the fact that it can also predict the non-existence of orders that do not occur.

The article, while uniquely dealing with the evaluation metric, interacts nicely with the other papers in this collection. Obviously, it relates to the discussion presented by Bach, but perhaps it is in seeing what amounts to a specific response to Lehmann's exhortations in this volume for the determination of universal laws (151) that the reader might acquire the greatest stimulation and satisfaction.

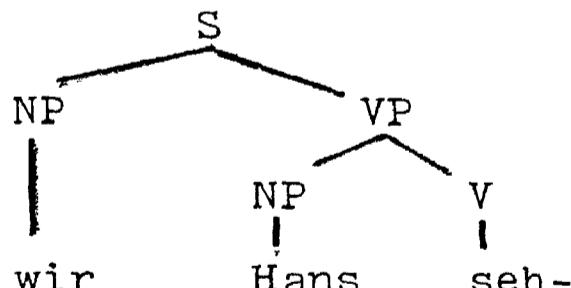
12. 'Verb-anchoring and verb-movement', by Arthur Schwartz (437-462). In mapping deep structural representations onto surface structure Schwartz suggests that constraints on transformations be made in terms of 'nucleus' and 'constituent' which would make 'no reference to lexical categories like N, V, P, etc.' (439). Schwartz claims that the distinction between SVO and VSO orders lies in terms of VP-constituency in that the notion VP is peculiar to SVO organization. VSO and SOV systems 'involve a decision about the position of the verb (predicate, generally) whereas SVO do not' (439-440). Verb-movement of any kind is to be found only in SVO systems. Or, put another way, 'SVO language-learners do not "make a decision" about the position of the verb, and so the verb is "movable"; learners of V-initial and V-final languages view the verb as a fixed point and so do

not "imagine" it as movable' (457). How Schwartz could ever know - let alone prove - all this is beyond me.

The discussion is interesting but omits careful definition of the presumably important constraints mentioned in the introduction such as the 'Unit Movement Constraint' and the 'Fixed Nucleus Constraint', referring the reader instead to the writer's other publications. Moreover, what the author means by '(make a) decision' and 'imagine it as movable' (above) is also left unstated. The reader feels prepared to accept Schwartz's arguments (examples from several languages are provided) but senses a lack of focus due perhaps to the preliminary nature of this investigation. Hence, exactly why the analysis of German subordinate clause structure that shows



is to be preferred over



is not clear.

## Conclusion

Li makes the following observation in his preface (iii)

The empirical facts amassed by Greenberg have made it possible to study the WHY and the HOW questions concerning the synchronic nature of word order and the diachronic process of word order change. However, during the Sixties, the field of syntax in the United States was almost exclusively the domain of those who researched the synchronic structure of English, as if there were an operational synonym between 'theoretical significance' and 'transformational study of English.' Thus, in the years immediately following the publication of Universals of Language, the immense potential for theoretical investigation offered by Greenberg's cross-language study was accorded little attention. Not until the Seventies have attempts been made to understand and explain those WHY and HOW questions which are the obvious consequences of Greenberg's universals.

This book appears to have a place in the contemporary scene somewhere the Labovian part of the spectrum and the part occupied by the generative theoreticians; the former, allowing a dynamic interpretation of empirical linguistic and supralinguistic facts; the latter, strongly theory-oriented, producing as a consequence a more static model. The efforts represented at the Conference on Word Order and Word Order Change exhibit on one hand the dynamic empiricism of the Labovian working method, while on the other, show a response to the general call for greater rigor and a reconciliation of these linguistic facts with the generative model.

It is interesting to go back a few years to the symposium held at the University of Texas at Austin in 1966, the proceedings of which were published in 1968 as Directions for Historical Linguistics. In an essay entitled, 'Empirical foundations for a theory of language change', by Weinreich, Labov, and Herzog we can see an anticipation of the contributions made by the

the participants in the Santa Barbara conference. Greenberg's work, they write, indicates two important modes of investigation (Lehmann and Malkiel 1968:138):

(1) the clarification through empirical means of the abstract claim that synchronic systems have 'dynamic' tendencies... and (2) the use of quantitative methods to replace anecdotal evidence and persuasive argument.

They go on to criticize - with justification - that at that time Greenberg was lacking in an over-all theory of language structure or language change. But in the next paragraph their foresight and insight grows dim:

We are encouraged by Greenberg's use of quantitative methods and his ability to isolate significant trends in structure. At the same time, one must admit that he is necessarily confined to surface structure at the lowest level of reliability which is common to the descriptions of the languages available to him. It is sometimes argued that one must have a comprehensive theory of language or language change as a whole, before one can begin to investigate language or language change seriously. If one holds to this doctrine, one would have to be extremely critical of Greenberg's workmanlike procedures. (138-139; emphasis added)

Returning to the present, it is evident that Weinreich et al. were correct in recognizing the potential in Greenberg's work, and in recognizing its need for a theoretical perspective. It is also clear that they underestimated the value of the study of word order.

Keeping this in mind, the volume under review must be seen as a breakthrough first, because it effectively synthesizes Greenberg's universals and generative theory, and secondly, because this synthesis leads to an unprecedented understanding of the causal relationships between the surface constituents of a sentence.

Of course, many of the contributors to this collection have published similar material that predates this volume. But their success (such as it might have been from case to case) was singular. The impact of this book derives from the strength of the common purpose of the twelve contributors. Certainly many of the claims made by these scholars will eventually have to be modified, (some can be shown already to be in conflict), or will have to be cast aside altogether. They readily admit that much of their work is tentative. The 'breakthrough' is not decisive. However, these studies of word order and word order change constitute a stimulus for new explanations in syntactic theory comparable to the stimulus provided to phonological theory by the notions of markedness or naturalness.

Hence, this volume represents the burgeoning of a third direction, an alternative to the major competing theories of language and language change. It is non-Labovian and non-generative, though it draws heavily from both. The study of language typology may contribute in an unexpected way to our understanding of linguistic change.

This book is recommended reading for any professional linguist. For the teacher it is a valuable asset because it can be used as one of the few references for historical syntactic change. Alas, not everybody likes phonology.

## REFERENCES

- Fries, Charles C. 1940. On the development of the structural use of word-order in Modern English. *Lg.* 16:199-205.
- Greenberg, Joseph. 1973. Numeral classifiers and substantival number: problems in the genesis of a linguistic type. *Working Papers on Language Universals*, 9. 1-39. Stanford.
- \_\_\_\_\_. (ed). 1962. *Universals of language* (second edition). Cambridge: M.I.T. Press.
- Hyman, Larry. 1971b. Consecutivization in Fe'fe. *JAL*, 10.2: 29-43.
- Lakoff, Robin. 1972. Another look at drift. *Linguistic change and generative theory* ed. by Robert Stockwell and Ronald Macaulay, 172-198. Bloomington: Indiana University Press.
- Lehmann, W.P. and Malkiel, Y., (eds.). 1968. *Directions for historical linguistics*. Austin: University of Texas Press.
- Sapir, Edward. 1921. *Language: an introduction to the study of speech*. New York: Harcourt, Brace & Co.
- Vennemann, Theo. 1973a. Explanation in syntax. *Syntax and semantics*, Vol II, ed. by J. Kimball, 1-50. New York: Seminar Press.
- Weinreich, U., Labov, W., and Herzog, M. 1968. Empirical foundations for a theory of language change. *Directions for historical linguistics*, ed. by W.P. Lehmann & Y. Malkiel, 95-188. Austin: University of Texas Press.

I N F O R M A L   S P E E C H  
ALPHABETIC AND PHONEMIC TEXTS WITH STATISTICAL ANALYSES AND TABLES

Edward C. Carterette and Margaret Hubbard Jones

*University of California Press  
Berkeley  
1974*

Reviewed by John B. Carroll  
The L. L. Thurstone Psychometric Laboratory  
University of North Carolina at Chapel Hill

Is "spoken language" different from "written language", and if so, how? This was the focal question addressed in this monograph. It was the authors' belief that "all previous statistical studies of language . . . have derived their material from written language." They hoped to show that "genuine spoken language is actually quite different from written language, even on such a gross level as proportional frequencies of letters and phonemes" (p.3). "Genuine spoken language," in their view, is the sort of language that would be used in free conversations among peers. They proceeded, therefore, to collect large samples of such language.

Being interested also in age trends, they sampled conversational speech from first-, third-, and fifth-grade children, and from "adults" (actually, junior college students from elementary psychology classes, mean age not specified). All were selected as being ostensibly middle-class native speakers of a "Southern California dialect." The conversations recorded might well be described

as..."bull sessions" among three participants that, in the case of the children, were put into motion by a friendly adult who faded into the background after the conversation warmed up, and in the case of the "adults," took place in what was purported to be an experiment in small group process. Conversations were tape-recorded and then transcribed both in regular orthography and in phonemes. By the authors' reckoning, the resulting corpus, from 58 peer-group sessions involving 174 different individuals, contained 84,164 lexical words, 313,694 alphabetic letters (in the conventionally spelled form) and 251,360 phonemes (in the phonemically transcribed form). If word and sentence marks (spaces and periods) are included, the figures are 405,906 alphabetic characters and 282,240 phonemes. The total corpus thus gathered occupies pages 57 through 439 of the book; the conventionally printed form and the phonemically transcribed form are on facing pages. No information is given as to whether the corpus is now available in machine-readable form, although it must have existed in that form at some stage of its analysis.

Various types of printed material were also analyzed, to represent "written language": school readers at several grade levels, trade books rated as liked by children, and adult material from a previous study by Newman and Waugh (Information and Control, 1960, 3, 141-153).

The principal mode of analysis--in fact essentially the sole mode of analysis--was inspired by information theory, and concentrates on the phonemes and graphemes of the corpus. Extensive tables (pp. 441-646) give data on first-order frequencies and probabilities

of letters and phonemes, and some of the higher-order sequential frequencies and probabilities (up to "triphones" for phonemes, and word-initial and word-final "tetragrams" for graphemes). The discussion of these statistical tables occupies pages 23-45.

The book itself is a handsome production, printed in a tasteful style on high-quality stock and nicely bound in cloth. On closer examination of its contents, however, one is tempted to conclude that the authors, having completed their manuscript and handed it over to the printer in about 1968, proceeded to divest themselves of any further responsibility for its editing and publication. Only in this way can one explain the many egregious typographical errors, inexcusable editorial changes, and glaring omissions of important materials that are to be noted in the book. The defects are in many instances serious enough to destroy much of the book's potential usefulness.

I infer that the manuscript was completed in 1968 or thereabouts because a reference to a 1968 article is cited as "in press," and there are no references to any publications subsequent to that year. Along with illiterate spellings and typographical errors such as pulication, concensus, idiosyncracies, and diagrams (for digrams, p. vii), we find inconsistent mathematical notation (pp. 18-20) and incorrect plotting of data points (in Figure 3.3.1 the points do not everywhere increase monotonically, although they must, in theory, and do, by the values recorded in Table 7.6, p. 451). But these matters are trivial beside the blatant alterations in some of the tables. On page 43 the authors state that in trigram tables the symbol (/) stands for a word space, in the tables them-

selves (pp. 479-535) no such symbols are to be found. Apparently the original manuscript contained the symbols, but the printer converted all of them to blank spaces and left-justified the entries. Thus, there is no indication as to whether the most frequent trigram, printed as "th", is to be taken as /TH or TH/. In this particular case, it is almost certainly to be taken as /TH/ (from word-initial 'TH- in frequent determiners and some content words), but what about a less frequent trigram such as what is printed as "en": is this /EN or EN/ ? Fortunately a similar error does not occur in the tables of "triphones" (pp. 537-613) where the printer indicated the space character as a carat (^), although the authors intended use of the slash (/).

The gross omissions are of certain summary statistical tables that were, according to the authors' text discussion (pp. 43-45), supposed to accompany the tables of triphones and trigrams. Tables 8.7.1-8.7.4 are not, as promised, preceded by tables giving frequency distributions of trigrams; similarly, the tables in the 8.9 series are not accompanied by the promised frequency distributions. I gather also that these omitted tables contained reports of certain information-theoretic statistics such as  $H$  or  $H'$ . With much effort, a user of this book could construct the frequency distributions from the detailed tabulations, but it is still inexcusable for the printer to have omitted them.

But the printer (or the publisher) is not to be blamed for everything. There are also problems with the manuscript, and the research that lay behind it. Questions must be raised about the purposes of the work, the procedures, and the methods of analysis.

Purpose of the work. At the time the work was undertaken, probably in the early 1960's, it may have been correct to say, as the authors do, that all previous statistical analyses of language had been based on written or printed material. Nevertheless, one can think of exceptions: even the authors cite the study of telephone speech by French, Carter and Koenig (Bell System Technical Journal, 1930, 9, 290-324), although this study has its limitations. Concurrently with the work reported in this book, a number of statistical studies of spoken language appeared (e.g., D. Howes, A word count of spoken English, Journal of Verbal Learning and Verbal Behavior, 1966, 5, 572-606; F. Goldman-Eisler, Psycholinguistics: Experiments in Spontaneous Speech, New York, Academic Press, 1968), and several investigators collected samples of spoken language for various purposes (e.g., W. F. Soskin & V. P. John, The study of spontaneous speech, in R.G. Barker (Ed.), The stream of behavior, New York, Appleton-Century-Crofts, 1963). These and other studies could have been cited by Carterette and Jones in their list of references; they were not.

What is relatively unique about Carterette and Jones' samples is that they were collected and analyzed by a uniform, specified procedure, with considerable attention to insuring that the speakers were representative of some defined population at several age levels. The extensive comparative analysis of these texts in both their spoken (phonemic) and written (graphemic) form is also unique. Although one may disagree with the modes of analysis that Carterette and Jones chose to use, we have them to thank for making their corpus available for any other types of analysis that might be desir-

able or feasible.

Are the authors correct, however, in saying they are studying differences between "spoken language" and "written language"? Evidently the authors mean to draw the distinction not between "written" language and "spoken" language as such (for any sample of language can be either written down or spoken aloud), but between "informal language" and "formal language," i.e. edited language. That is, they are concerned with how language is generated. They apparently feel that the most "natural, genuine" language is generated in informal conversational situations where the speakers have little if any pressure to make their speech conform to artificial norms of correctness, grammaticality, or even communicative efficiency. It is probably for this reason that they titled the book Informal Speech. They were also interested in ~~the~~ development of speech generation, at least over a certain age range <sup>from</sup> from the first grade to the junior college level. Note, however, that they did not sample the informal (relatively unedited) speech of highly educated, mature speakers such as might be found in a congressional cloakroom or the salons of an ivy-league faculty club. It is quite possible that speech sampled in such circumstances would conform fairly closely to the norms of edited, written language, at least in some respects (and probably in the respects studied by the authors).

If one bears in mind, therefore, the limitation that it is not speech as such (as opposed to writing) that these authors have studied, but rather unedited speech of relatively immature speakers,

the work has considerable unique value by virtue of its presentation of extensive samples of such speech. The authors do not really analyze, however, the ways in which unedited language differs from edited language, nor the ways in which speakers develop strategies of editing their speech.

Another objective of the authors was to use "the rather powerful tools of information theory in the description of informal speech over the age range, in an effort to trace the role of redundancy in shaping language as a person uses it and presumably understands it in discourse." I will have more to say about the authors' use of information theory below.

Data collection procedures. For their purposes, the procedures were excellent--certainly superior to procedures (interviews, contrived play situations, classroom discourse) used by other investigators, for the procedures certainly elicited informal, unedited speech full of interrupted sentences, hesitations, false starts, etc. The content covered a wide range of topics. Nevertheless, it was all conversation; the participants were merely exchanging ideas, and declarative and interrogative utterances abound. They were not directing each others' physical activity; thus, there appears to be a low incidence of imperatives, requests, etc. The corpus is certainly large enough for the kinds of analysis employed by the authors at a phonemic or graphemic level, but it might not be sufficiently large or representative for certain types of lexical or syntactical analysis.

Transcription procedures. Transcription of the corpus was a formidable and time-consuming project, not only in terms of a con-

ventional printed form but also and particularly so, in terms of a phonemic version. One can only say that the authors made approximately the best of a very difficult job. They found it impossible always to identify speakers, and decided to omit any speaker identifications, showing only changes of speakers. For the phonemic transcription, they used a modification of the Trager-Smith transcription suggested by Peter Ladefoged, but found it hard to get hired "phoneticians" to adhere to the system consistently. The system used was admittedly only partially phonemic; for example, the glottal stops that were recorded may or may not be phonemic. One wonders how consistently such distinctions as those between /aydownow/, /aydənnow/, and /aydownnow/ were observed. The treatment of pause phenomena was particularly bothersome. Pause phenomena were represented in the phonemic transcriptions only by spaces and periods; thus, the phonemic transcriptions contain a high proportion of very long "phonemic words" like /naktowvərəməlkbadəlwənnəywərgowɪŋənɛr./, transcribed in graphemic form as "knocked over a milk bottle when they were going in there" (pp. 312-313). But in the printed version the location of the junctures between these "phonemic words" is unfortunately not shown, although it would have been fairly simple to have done so, perhaps by the use of slashes (/). For certain purposes, it is unfortunate, also, that certain kinds of material were omitted from the transcriptions, e.g. repetitions of words when in answer to wh- questions, and certain kinds of interruptions in continued sentences. It is curious that proper names were generally deleted in the printed version but left in the phonemic

transcription (e.g. compare pages 432-433); the authors apparently felt that privacy would be preserved in the mystique of a phonemic transcription but not in conventional spelling.

One can only guess as to what stress and intonational phenomena occurred in the conversations. The printed version contains no question or exclamation marks, and the phonemic transcription contains no indication of intonations. Whether the tape-recorded material is archived somewhere, available for further analysis, is not stated.

Analysis. As mentioned previously, all analyses of the material were at a phonemic or graphemic level. The intention was to use the "powerful tools of information theory" to trace the development of "redundancy." This mission was certainly enough to occupy the authors throughout the course of their study; it was apparently their intention to leave other types of analysis to future workers. One may raise the question, however, whether information theory analysis was really so "powerful", and indeed how informative it was when applied solely to zero- and higher-order phenomena such as word distributions, distributions of syntactical patterns, etc.? Perhaps I can illustrate my attitude by relating my own experiences with such analyses on the phonemic level. In 1951, in connection with an interdisciplinary seminar of psychologists and linguists, a group of the participants decided to investigate the information-theory characteristics of sequences of phonemes in American English speech. Not having readily available any authentic samples of speech, we decided to

make a phonemic transcription of a series of one-act plays written for high school student performances. One of the linguists, (Fred Agard) transcribed some 20,000 phonemes from this corpus and subsequently I did a number of information-theory analyses of the data. The results were incorporated in a mimeographed report that, incidently, was cited by Carterette and Jones in their reference list as, however, "not seen." (A copy of the report could easily have been made available to them if they had asked me for it.) It seemed to me, having done these computations, that they meant very little. I tabulated zero-order, first-order, and second-order sequential probabilities, estimates of information ( $H$ ), and the like, but it seemed to me that all that was being shown was that certain phonemes and combinations of phonemes were more frequent than others because of their appearance in words or sequences of words having the higher frequencies. People generate language, I reasoned, not by selecting phonemes but by selecting words and sequences of words; therefore, the frequencies of phonemes and their combinations were mere epiphenomena. Tabulations of these sequences might conceivably have some uses in designing stimulus material for psycholinguistic experiments, to control for the frequencies of habit patterns, but beyond that they would be of little interest either linguistically or psychologically. When Lee Hultzen requested use of my material for his analysis (Hultzen, Allen, and Miron, Tables of transitional frequencies of English phonemes, University of Illinois Press, 1964), I was only too happy to turn it over to him.

Now, what do we find in the work of Carterette and Jones?

These authors must have been disappointed with their findings on the zero-order distributions of letters and phonemes. They find that for the distributions of letters in the conventionally printed version of their conversational samples there is very little change over age. As they say, "the largest change is in m, which decreases steadily from first grade to adult speech." This is "partly accounted for by a decrease in the use of 'um' as a noise word, with the concomitant [sic] rise in the use of 'you know.'" (p. 23). Furthermore, the distribution of letters is about what many other investigators have found for samples of printed English. Etaoin Shrdlu can still be the linotyper's friend! Changes in zero-order distributions over different age levels seem mainly to reflect changes in the frequencies of certain "noise words" like "um", "well," etc. Yet the authors take pains to compare their results to those of other investigators of phoneme distributions, claiming that "the highest correlations usually occur between phonemic systems derived from material closest to natural speech, whereas the lowest correlations occur with phonemic systems based on material furthest from natural speech" (p. 29). Their argument is not convincing, however, for they seem to underestimate the effect of different systems of phonemic transcriptions used in various studies, and also the effect of the "editing" that occurs as one goes from highly informal speech (with its "noise" words) to more highly edited speech (e.g., contrived speech in high-school plays).

What the authors make most point of are the differences among various types of material, spoken or written, in "redundancy"

or "relative sequential constraint" as defined by information-theoretic statistics. Actually, there are no differences between first-grade speech and the "adult" speech samples in phoneme redundancy--the curves of relative sequential constraint across second-symbol positions (Figure 3:3.1) are virtually identical, leveling off at about .30. I would interpret this to mean merely that both first-graders and adults are using the same (Southern California English) language, and that the same system of phonemic transcription has been used in the two cases. I would be much surprised to learn that first-grade and adult speech samples could not be differentiated in many ways--in lexical selection, in grammatical patterns, etc. Sequential constraint of phonemes is probably not a sensitive way of indexing anything useful or interesting about language samples. It is at least misleading for the authors to state that "[i]n terms of simple sound pattern redundancies, therefore, 6-year-old speech is already adult" (p. 30).

The case is slightly different when the redundancy statistics are applied to letters (graphemes) in the transcriptions of speech, or in printed materials. First grade speech is shown to be slightly more redundant than adult speech; I would think that a large part of the difference could be traced to differences in lexical distributions--differences that show up in letters but not in phonemes because lexical boundaries are observed in the letter statistics but are rarely preserved in the phonemic transcriptions. First grade readers are also much more redundant in letter distributions than even first-grade speech; but it is well known that

first-grade readers are typically highly redundant in their lexical distributions. Redundancy statistics based on grapheme distributions apparently reflect these lexical distributions, but how reliably, it is difficult to tell. There is the suggestion, arising from these results, that redundancy statistics based on grapheme distributions and their sequences could be a useful surrogate for other types of indexes based on incidence and sequences of words, grammatical patterns, etc. But the authors' suggestion that differences thus revealed between natural speech and written material are somehow important to take into account in the teaching of reading seems rather forced and gratuitous.

The authors also pay some attention to words and sentence lengths, both in the printed material and the phonemic transcriptions. Their "phonemic word" is defined "only in terms of a prosodic feature, specifically a pause in the flow of sound:" They find phonemic words to be three times as long, on the average, as lexical words. They suggest, "Insofar as the units of spoken language and written language are different, the learning of written language (reading) will be difficult," but do not explore the implications of this suggestion further.

Final evaluation. The authors at several places state that the results presented should be used "with great caution." I would say that this caveat must be taken to apply to the whole work. Some linguists, and psychologists, and educators may find uses for the transcribed speech samples, but the limitations of

these samples--particularly in the way in which they are presented--must be borne in mind. One can conceive uses for the statistical tables, perhaps by psychologists seeking ways to control experimental stimulus material for phoneme frequencies. In general, however, one wonders whether it was worthwhile to pursue the statistical analyses and tabulations of phoneme and grapheme frequencies to the extremes reached by Carterette and Jones. It is little wonder that these authors seemed to abandon their interest in their research after completing the manuscript represented in this strangely unfinished book. But more importantly, the authors have not persuaded me that "spoken language" is different from "written language" in any interesting ways. It is conceivable that interesting differences exist between "informal" and formal speech, but the authors' analysis has not revealed them.

Pp. xiv + 646

\$25.00.

**PERSONAL NOTES**

AGRESTI, HENRI. \*Department of Linguistics, University of California San Diego, La Jolla 92037. Methodology in theoretical linguistics and discourse analysis.

ARKWRIGHT, THOMAS D. To Automated Systems Division, Defense Language Institute, Presidio of Monterey, California 93940. From the University of Quebec.

CARROLL, JOHN B. \*Kenan Professor of Psychology and Director of the L. L. Thurstone Psychometric Laboratory, University of North Carolina, Chapel Hill 27514. Measurement and analysis of individual differences in language abilities; in relation to psycholinguistic and cognitive studies of learning, memory, and information processing.

HAWKINSON, LOWELL B. \*Project MAC, Massachusetts Institute of Technology. Mailing address: Morningside Lane, Lincoln, Massachusetts 01773. Semantics and discourse; computation; grammar.

PAGE, ELLIS B. Elected, by the Division of Educational Psychology of the APA: Program Chairman, then President, then Past-president & Executive Committee Member. Visiting Universidad Javeriana, Bogota, Colombia, until October.

STYGAR, PAUL. \*TRW Systems Group, Box 282, McLean, Virginia 22101. Conversion of unstructured textual data to structured data. Parsing. Language design.

*\*Specialties and improved addresses received since publication of the 1974 Membership Directory.*

