American Journal of Computational Linguistics *Microfiche 51*

THE FINITE STRING

NEWSLETTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS

VOLUME 13 - NUMBER 6 SEPTEMBER 1976

Both the Current Bibliography and the promised index to AJCL are absent from the present packet, because of the difficulties inherent in operations based on voluntary effort.

American 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 14261*

Editorial Assistant: William Benzon

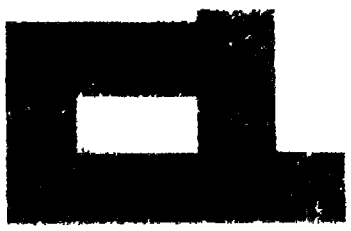
Editorial Address 5048 Lake Shore Road, Hamburg, New York 14075

Managing Editor: A. Hood Roberts

Assistant. James S. Megginson

Production and Subscription Address. 1611 North Kent Street, Arlington, Virginia 22209

CONTENTS



ASSOCIATION FOR COMPUTATIONAL LINGUISTICS

14TH ANNUAL MEETING

SAN FRANCISCO - OCTOBER 8 - 9, 1976

ACCOMMODATIONS: HILTON HOTEL

SESSIONS: ST. FRANCIS HOTEL

REGISTRATION: $15 FOR MEMBERS
$20 FOR NONMEMBERS
$10 FOR STUDENTS

PROGRAM: FRIDAY OCTOBER 8

8.30 Registration

TECHNIQUES IN LANGUAGE PROCESSING

9:00 AN EASILY COMPUTED METRIC FOR RANKING ALTERNATIVE PARSES

George E. Heidorn, IBM Research ... *Abstract, Frame 5*

9·30 MEDIAN SPLIT TREES: A FAST LOOKUP TECHNIQUE FOR FREQUENTLY OCCURRING KEYS

B. A. Sheil, Harvard University ... *Abstract, Frame 7*

10.00 TRANSLATING 'WELL-WRITTEN' ALGORITHM DESCRIPTIONS INTO CODE

Jerry R. Hobbs, City College of CUNY . . *Abstract, Frame 9*

10·30 GENERATING NATURAL LANGUAGE EXPLANATIONS OF COMPUTER PROGRAMS

George E. Heidorn, IBM Research ... *Abstract, Frame 13*

11:00 THE SEMANTIC INTERPRETATION OF MASS NOUN EXPRESSIONS IN THE PHLIQA 1 SYSTEM

Harry C. Bunt, Philips Research Laboratories ... *Frame 15*

11.30 · EMPIRICAL STUDIES OF NOUN MEANING

John Bennett, University of Wisconsin - Madison ... *Frame 18*

1:30 PANEL: EVALUATION OF NATURAL LANGUAGE SYSTEMS ... *Frame 21*

Chair: Joyce B. Friedman, University of Michigan

Participants: William A. Martin, M.I.T.

William H. Paxton, Stanford Research Inst

Stanley Petrick, IBM Research

Naomi Sager New York University

Eric van Utteren, Philips Research Labs

Terry Winograd, Xerox Research

FRIDAY, OCTOBER 9

MACRO MODELING OF NATURAL DISCOURSE

8 30 METHODS FOR MODELING DIALOGUE (SPECIAL GROUP PRESENTATION)
William C. Mann, James A Levin, and James A Moore
Information Sciences Institute, University of Southern California ... *Abstract, Frame 22*

10.00 A PROCESS MODEL FOR SPEECH ACT UNDERSTANDING
Allen Munro, University of California, San Diego ... *Frame 3*

10.30 NATURAL LANGUAGE UNDERSTANDING BY COGNITIVE NETWORKS
Richard Fritzson, SUNY at Buffalo ... *Abstract, Frame 32*

11·00 HOW TO REPRESENT AND USE KNOWLEDGE ABOUT CAUSALITY
Charles J Rieger, University of Maryland ... *Abstract, Frame 33*

11.30 A HEURISTIC SEARCH APPROACH TO DISCOURSE ANALYSIS
Jerry R. Hobbs, City College of CUNY .... *Abstract, Frame 35*

AN EASILY COMPUTED METRIC FOR RANKING ALTERNATIVE PARSES

GEORGE E. HEIDORN
IBM RESEARCH

The idea of attaching a modifier to its nearest potential modificand is often stated as a reasonable heuristic in natural language parsing. (This is essentially Kimball's Principle Number Two*, "Terminal symbols optimally associate to the lowest nonterminal node." Although Kimball calls this principle "right association" and illustrates it with right-branching examples, it can apply equally well to left-branching structures.)

In the sentence, "Are those invoices produced from the orders processed by the system in New York?", the prepositional phrase "in New York" could potentially modify "system", "processed", "orders", "produced" and "invoices", both syntactically and semantically. Similarly, "by the system" has four possibilities and "processed" has two. According to the heuristic stated above, the preferred analysis is: "in New York" modifies "system", "by the system..." modifies "processed", and "processed..." modifies "orders".

A "potential modificand" must be acceptable pragmatically as well as syntactically and semantically. This means that if the system referred to in the above example were spread

---

*Kimball, John. Seven principles of surface structure parsing in natural language. Cognition 2 (1), 15-47.

out over the entire country and just had one of its processing stations in New York, the preferred analysis would have had "in New York" modifying "processed" instead of "system". Similarly, dialogue context can have an influence on what qualifies as a potential modificand.

This paper describes an easily computed metric that involves attaching a number to each syntactic unit as it is formed, to rank the possible analyses for any portion of an utterance according to the above-stated heuristic. The technique is illustrated by its use in a system which supports a natural language dialogue for automatic business programming. This system utilizes syntax, semantics, pragmatics, and context to understand user utterances.

MEDIAN SPLIT TREES:
A FAST LOOKUP TECHNIQUE FOR FREQUENTLY OCCURRING KEYS

B. A. SHEIL
HARVARD UNIVERSITY

Median split (MS) trees are a new technique for searching sets of keys with highly skewed frequency distributions (such as the lexemes of a natural language). The basic idea is to modify frequency ordered binary search (FOBS) trees to contain two key values--a node value which identifies the key which resides at that node (as in a conventional binary search tree), and a split value which gives the largest key value to be found in the left subtree. (If the keys are expensive to store in the tree nodes, or costly to compare, they may be hashed into integers without affecting the algorithm.) Searching an MS tree proceeds as in a binary tree except that the decision to go left or right from a node whose node value does not match the current key is made by comparing the current key to the split, rather than to the node, value. The use of two different values allows one to prevent the search tree from being unbalanced as FOBS trees become when a high frequency key has an extreme key value. In fact, by selecting the median of the key values of a node's descendants as its split value, one can force the search tree to be perfectly balanced--which both allows a highly space efficient representation of the tree, and achieves high speed search.

The average cost per search in an MS tree, like a FOBS tree, depends on both the frequency distribution of the keys, and the ordering relation on them. Performance analyses for specific frequency distributions and key orderings of interest are presented, and it is shown that, unlike FOBS trees, MS tree search time is log n bounded and is very stable around a value that can be shown to be optimal for any binary tree search procedure. Furthermore, in addition to being substantially faster than "optimum" binary tree search, an MS tree can be build for a given ordering relation and set of frequencies in time n log n, as opposed to n^2

A discussion of the application of MS trees to dictionary lookup for English is presented, and the performance obtained is contrasted with hash and trie solutions.

TRANSLATING 'WELL-WRITTEN' ALGORITHM DESCRIPTIONS INTO CODE

JERRY R. HOBBS
CITY COLLEGE OF CUNY

In this paper we describe a system for the semantic analysis of "well-written" algorithm descriptions in English, such as one finds in Knuth (3), and the translation of the resulting structures into a program in a higher-level programming language, such as PL/I. There are two approaches one might take toward the sublanguage of algorithm descriptions. One may treat the English as a deviant programming language and try to smooth out the deviations. Or one may view the English as a well-behaved, highly regular form of natural language, and apply to the text the sophisticated techniques of analysis that have been developed for less tractable sorts of texts. We have taken the latter approach.

Semantic analysis is done by a system that has been developed for and is being tested on a variety of natural language texts (1). It takes as input the text in a predicate calculus-like notation produced by a syntactic front end (4, 2). It applies semantic operations which access a data base of "world knowledge" inferences that are drawn selectively, depending on context and in response to the operations. The operations are described below with examples that show their importance in analyzing algorithm descriptions:

1. Predicate interpretation: Inferences are sought to satisfy the demands predicates make on the nature of their

arguments.. Among other things, this operation recovers omitted material, for example when the whole is used to refer to the part. In

"T points to a binary tree"

the predicate "point" requires its second argument to be a node. The knowledge stored in the data base about binary trees is searched for a dominant node, the root node is found, and the sentence becomes

"T points to the root node of a binary tree."

2. Intersentence relations are determined by matching successive sentences against a small number of patterns. These patterns are stated in terms of inferences to be drawn from the current sentence and a previous sentence. An ordered heuristic search of the data base is used to locate the desired inferences. Intersentence relations are especially important in algorithm descriptions, for they determine the flow of control in the program. The most common and easily handled patterns in these texts are Temporal Succession, Cause, and Enablement. They are translated into successive lines of code.

But also important is the Contrast pattern: "The predicates of sentences S_1 and S_2 are the same. One pair of corresponding arguments are contradictory. The other pair are different but similar." Consider

"If VAL(M) > VAL(N), set N equal to LINK(N).
If VAL(M) < VAL(N), set M equal to LINK(M)."

It is intended that these translate not into successive instructions, but into a case statement. This is keyed by the Contrast pattern. (The top-level predicate is "imply".)
Similarly, consider the sentences

(1) "Decrease N by 1, and if it is zero, reset it to MAX."

(2) "Decrease N by 1, but if it is zero, reset it to MAX."

In (1) the test for zero comes after decreasing N, in (2) before. This is because in (2), "but" invokes the Contrast pattern, forcing us to recover the implicit "If N is not zero" before "Decrease". (2) is translated into a case statement.

3. Knitting: Redundancies are recognized and merged This frequently resolves anaphora as a byproduct.

4. Resolving anaphora: Antecedents not found by knitting are searched for by trying to maximize the redundancy of the text. In

"Decrease N by J, and if it is zero, reset it to MAX."

"N" is chosen over "J" as the antecedent of "it" since it is N whose value is subject to change.

5. Search for Primitives: Certain predicates are "primitive" to this application, i.e the task component has substitution rules for them. Where possible, a decomposition of the text in terms of these primitives is looked for. For instance, a tree is, in part, a set of nodes, a node is a set of fields, and a field may be a number. "Set" and "number" are primitives. To decrease N by 1 is to cause N to be equal to one less than before, and "cause to be equal to" is a primitive.

Semantic Analysis augments and interrelates the text What is produced is then given to the task component which performs two functions:

1. The temporal succession relations discovered by operation 2 may form only a partial ordering on actions. A linear ordering is imposed.

2. Substitution rules are applied to primitives to translate the semantic representation into code. A set translates into an array declaration, a set of sets into a two-dimensional array. "Cause to be equal to" translates into an assignment statement.

Bibliography

1 Hobbs, Jerry R. A general system for semantic analysis of English and its use in drawing maps from directions. AJCL Microfiche 32, 1975.

2 Hobbs, J. R., and Grishman, Ralph. The automatic transformational analysis of English sentences: an implementation. International Journal of Computer Mathematics, to appear.

3 Knuth, Donald. The art of computer programming, 1. Addison-Wesley, Reading, Mass., 1968.

4 Sager, Naomi, and Grishman, Ralph. The restriction language for computer grammars of natural language. CACM 18 (7), July 1975, p. 30.

GENERATING NATURAL LANGUAGE EXPLANATIONS OF COMPUTER PROGRAMS

GEORGE E. HEIDORN
IBM RESEARCH

An automatic programming system that converses with a user in a natural language must be able to generate explanations of portions of the program being produced. In the system that we are developing for producing customized business application programs, these explanations must be given in application terms rather than program terms. For example, the system might say, "The quantity shipped is the lesser of the quantity ordered and the quantity available," rather than "QS=MIN(QO,QA)".

One could certainly imagine generating such sentences by having an inventory of sentence patterns and name phrases and using a fill-in-the-blanks approach. However, in addition to not being very interesting, such an approach would certainly not provide the same flexibility of expression that a more linguistically motivated approach would. Also, it seems reasonable to do this generation by utilizing information that the system already has for understanding the user's English utterances.

In our system, information is stored in the form of a semantic network with three parts. The program part contains information about the programming language and about a specific program; the application part contains information about business concepts; and the language part contains information about English words. Each of these parts has its own internal structure, and there are links between the parts where appropriate. For example,

within the application part the node for the concept QUANSHIPD has a SUPerset arc to QUANT and a KIND arc to SHP. Between parts of the network, QUANSHIPD is linked as the RAO (Related Application Object) arc of the program node QS, the language node QUANTIT (the stem of "quantity") is linked as the RLO (Related Language Object) of QUANT, and the language node SHIP is the RLO of SHP. In general, the links between adjacent parts are many-to-many.

The linguistic processing in our system is specified by augmented phrase structure rules, uniformly for the three levels of processing: semantic, syntactic, and morphological. Given a pointer to a node in the program part, along with some information about what aspect of that node is to be explained, these rules can generate an appropriate English statement by utilizing the structure of the network. Because these statements are constructed from very small pieces (i.e. word stems), considerable flexibility of expression is possible, including taking into account dialogue context.

This paper will describe in some detail the overall method of generation, with special emphasis on complex definite noun phrases.

THE SEMANTIC INTERPRETATION OF MASS NOUN EXPRESSIONS IN THE PHLIQA 1 SYSTEM

HARRY C. BUNT
PHILIPS RESEARCH LABORATORIES

A large number of the most commonly used English nouns belongs to the category of mass nouns. In recent years it has become increasingly clear that the semantic analysis of expressions containing mass nouns encounters great difficulties because of the uncertain logical status of mass noun referents. This uncertainty is reflected in the appearance of a variety of vague terms, such as "substance", "quantity", "matter", "bits", "pieces", "stuff", etc. in recent literature on the subject (see Pelletier 1975). It is indeed fair to say that "the logic of mass expressions ... at present is grossly underdeveloped" (McCawley 1975)

Work in computational linguistics so far has not contributed to a solution or even a better understanding of the problems that mass nouns pose. In language understanding systems that have been developed in the past few years these problems have been evaded, either by restricting the admitted use of mass nouns trivially, as in the SHRDLU dialogue system (Winograd 1972), or by explicitly excluding their use altogether, as in the LSNLIS question answering system (Woods 1972).

The recently implemented question answering system PHLIQA 1 (Landsbergen 1976, Medema et al. 1975, Scha 1976) may be considered as a substantial step forward in this

respect. This system performs the semantic analysis of complex mass noun expressions, including quantifying phrases, adjectives, amount expressions, etc. The analysis is based on a theory, in which a central role is played by a logical formalism, developed for the semantic representation of mass expressions. These representations were required to meet the following conditions:

1. They should be in agreement with the ontological commitments of mass noun uses. In our view. the typical use of a mass noun indicates that the noun referent is viewed as a continuum; i.e. no articulation of the referent into individual elements is taken into consideration--which is of course not to say that the referent actually is a continuum.

2. They should bring the logical properties of mass expression to the light, e.g. rendering analytic sentences tautologically true.

The paper develops this semantic theory of mass noun expressions, and describes its incorporation in the semantic analysis performed by the PHLIQA 1 system.

Bibliography

Landsbergen, S. P. J. Syntax and formal semantics of English. in PHLIQA 1. Preprints, COLING 76.

McCawley, J. D. Lexicography and the count-mass distinction. Proceedings, Berkeley Linguistic Society, 1975.

Medema, P., W. J. Bronnenberg, H. C. Bunt, S. P. J. Landsbergen, R. J. H. Scha, W. J. Schoenmakers, and E. P. C. van Utteren. PHLIQA 1: multilevel semantics in question answering. AJCL Microfiche 32, 1975.

Pelletier, F. J., ed. On mass terms. Synthese 31 (3/4),. 1975.

Scha, R. J. H. Semantic types in PHLIQA 1. Preprints, COLING 76.

Winograd, T. Understanding natural language. Edinburgh University Press, 1972.

Woods, W. A., R. M. Kaplan, and B. Nash-Webber. The lunar sciences natural language information system. Final Report, Bolt Beranek and Newman, Cambridge, Mass., 1972.

EMPIRICAL STUDIES OF NOUN MEANING

JOHN BENNETT
UNIVERSITY OF WISCONSIN--MADISON

A central problem for computational models of understanding is to formalize the meanings of words which are manipulated by the model. This paper reports on research directed towards the formalization of the meanings of nouns. The goal of the research is to describe nouns sufficiently to enable the performance of a particular task of understanding: the appropriate selection of verb senses for multiply-sensed verbs.

The hypothesis being tested is that noun meaning can be described by primitive "units of meaning," here called features. The concepts of semantic features and componential analysis are familiar; the contributions of this work are the development of empirical methods of analysis and the generation of a general set of features.

The effect of noun meaning on the sense of a verb can be clearly seen by examining simple sentences with identical syntactic structure and no extra-sentential context. To study noun meaning in a regular, non-subjective fashion, a large number of simple sentences of the form noun-verb-noun have been generated. In groups of the sentences, the subject noun and verb have been fixed, while the object noun is varied. The variance in the object noun causes the verb to have different readings, or different mappings of the lexical verb onto one of its senses.

From that group of nouns, in the object position; which cause the selection of the same sense for the verb in question, the common thread of meaning is extracted and posited as a semantic feature descriptive of that group of nouns. Repetition of this process for various verbs and nouns yields a set of semantic features. These features are considered as formalizing noun meaning at least for purposes of verb sense selection.

The major problems in the above approach are 1) the inevitable intrusion of subjectivity into the feature extraction task and 2) the lack of generality of the feature set beyond the particular group of nouns and verbs studied. Several methods have been used to control subjectivity: repetition of the process using the same verbs with different nouns and vice-versa, and the generation of features from the regular study of other tasks of understanding which involve nouns--noun-adjunct:noun pairs (e.g. school house) and noun analogies. To ensure as much generality as possible, the words considered in the study have been chosen from the set of (statistically) most common English words and do not comprise any unified semantic field.

Results have been interesting and encouraging. In addition to the commonly used features ("human," "animate," "object"), features like "emotion," "authority," "concept," and "avocation" have been generated. Further work will look into the use of features in other tasks of understanding,

the use of empirical methods to study verbs, and the implications of features for semantic fields, meaning extension, and case structures.

PANEL: EVALUATION OF NATURAL LANGUAGE SYSTEMS

JOYCE B. FRIEDMAN, CHAIR
UNIVERSITY OF MICHIGAN

WILLIAM A. MARTIN, M.I.T.; WILLIAM H. PAXTON, STANFORD RESEARCH INSTITUTE; STANLEY PETRICK, IBM RESEARCH; NAOMI SAGER, NEW YORK UNIVERSITY; ERIC VAN UTTEREN, PHILIPS RESEARCH LABORATORIES; TERRY WINOGRAD, XEROX RESEARCH

Diverse natural language systems share a common problem of evaluation. This panel will address the topic of evaluation, emphasizing the evaluation of performance. The panelists will consider such questions as: What measures are used to evaluate a system? How sensitive is performance to vocabulary size and scope of syntax? What alternatives can be studied analytically or experimentally?

Performance can be considered absolutely (how effective is a system, with respect to resources and goals?), relative to problem variation (how would it respond to changes in, e.g., vocabulary size). relative to system variation, and relative to other systems. In each case both the results of evaluation of particular systems and procedures of evaluation are of interest.

The panel will consist of experts, all of whom have written large systems. These systems include speech systems and text systems, business-oriented question-answering systems and artificial intelligence systems, and systems with contrasting linguistic theories.

Methods for Modeling Dialogue

(Special Group Presentation)

William C. Mann, James A. Levin, and James A. Moore
Information Sciences Institute
University of Southern California

Research methods for modeling dialogue

When people communicate with machines, they do so by specializing and extending their existing ability to communicate. To design systems which interact effectively with humans, we need to better understand how people communicate with each other.

In particular, we need to view human communication as a problem-solving activity, in which the people so engaged are using language as a method to achieve certain of their goals. The existing research into natural language has not focused on this aspect, so from our point of view the work so far has been fragmentary and hard to use, particularly in terms of helping to enhance man-machine comm-nication.

Constructing a useful model of human communication is an extremely complex task. We view the controlling of this complexity (without sacrificing utility) as the central problem in designing our research approach. In response to this challenge, our research methodology contains some innovations, specifically intended to maintain control on the complexity, while retaining the usefulness of our results for designers of man/machine communication systems.

Specifically, these innovations include

--Case analysis, rather than functional design.

--Independent judgment of the adequacy of the model.

--Results in the form of individual algorithms, not full systems.

--Algorithms which are transferable into nonresearch systems.

We describe a research methodology, and the resulting Dialogue Modeling System which incorporates these innovations. The methodology analyzes naturally-occurring dialogues rather than prepared examples. For each dialogue we choose to analyze, a separate model is constructed. Each model is judged against annotations of the dialogue prepared by a trained observer who uses a predefined set of categories of phenomena.

Although the processes employed in each model are to be as general as we find practical, they may also be as ad-hoc as necessary to make the model work. After several models have been produced in this fashion, we expect to find that, although many processes have been too idiosyncratic to be used in more than one model, many others have seen repeated application and thus have demonstrated their more general utility. These processes are regarded as the principal results, since they have been verified against a diversity of observers' annotations.

Subsequent speakers will provide details on the memory organization of the model, the processors which operate on these memories, and the issues raised by the inclusion of the observer in the model development cycle.

KNOWLEDGE STRUCTURES FOR MODELING DIALOGUE

This paper describes the knowledge structures developed for our Dialogue Modeling System. Permanent knowledge is stored as Concepts in a semantic network Long Term Memory (LTM), using a predicate/parameter representational format Concepts that are currently salient to the Processors of the System are represented by tokens of those concepts, called Activations, in the Workspace (WS). Activations are transient entities, with new ones frequently appearing and existing ones changing or disappearing. The WS at any moment forms the current context, which drives current processing. All Processors intercommunicate through the WS, modifying the set of Activations and being influenced by them. The WS simplifies the integration of multiple sources of information about any given entity

Two kinds of higher level knowledge structures are used to model dialogues. There are Execution Scenes, representing organized clusters of actions and/or objects being discussed in the dialogue. These are similar to the other currently proposed higher level units, like "scripts" or "frames". There are also higher level units called Dialogue-games, specifications concerning the topic of discussion, the participants, the goals they are trying to achieve with a given type of dialogue, and a set of conventional subgoals for mutually achieving their goals.

In our analysis of naturally occurring dialogues, we have found many different kinds of Dialogue-games, including Helping, Smalltalk, Gripe, Polite-conversation, Order, Information-seeking, and Information-probing. An analysis of fourteen helping dialogues uncovered a surprising amount of consistency in the structure of these interactions, which is captured in the representation of the Helping Dialogue-game.

PROCESSORS FOR MODELING DIALOGUE

This paper provides details of the major Processors incorporated into our Dialogue Modeling System. These Processors are independent and asynchronous, monitoring the activity in the Workspace, and modifying it as appropriate.

PROTEUS PROCESSOR -- Every activation in the Workspace has a "level of activation" (a number) which corresponds to its degree of salience to the model's current activity. Each of these activations imparts some of its salience to its immediate neighbors (including those in Long Term Memory). Similarly, the activation level of each node is augmented by that imparted to it by all of its neighbors. Whenever a concept in LTM accumulates activation above a certain threshold, an activation of that concept is created in the Workspace. Thus, if enough concepts, closely related to X are active, X will also become active. Proteus is the Processor which attends to all of this, thus providing a mechanism for bringing a concept into attention, based on the attention directed toward related concepts.

MATCH PROCESSOR -- This identifies concepts in LTM that are congruent to existing activations in the WS The Dialogue Modelling System contains a number of equivalence-like relations, which Match uses to identify a concept in LTM as representing the same thing as an activation of some seemingly-different concept Once such an equivalent concept is found, it is activated. Match thus provides the same sort of attention direction that Proteus does, except that Proteus is driven by the explicit connectivity of the memory, while Match responds to more indirectly-represented similarities.

DEDUCE PROCESSOR -- This attempts to apply a rule, whenever that rule has become active. Rules are concepts of the form. "(Condition)->(Action)". They become active by virtue of the activity of the concept which is their condition half. Deduce senses the activity of a rule and attempts to apply it by activating the concept for the action. Whatever correspondences were evolved in the course of creating the activation of the condition (left) half of the rule are carried over into the activation of the action (right) half. The combination of Match and Deduce gives us all the capability of a production system, where the system is represented by all the Rules in LTM.

DIALOGUE-GAMES PROCESSOR -- Dialogue-games are a theoretical construct we have developed to represent certain communicative forms used by people to interact in achieving goals each person has. Each Dialogue-game has a set of parameters, including two roles and a topic. Each parameter has a set of

specifications on what can serve as a value for itself. We have found that these dialogue forms are identified in conversation by utterances that attempt to establish these parameters. Once a Dialogue-game has been activated as possibly the communication form being proposed for a dialogue, the Dialogue-game Processor operates on it to verify that the parameters are properly specified, and then to establish the subgoals that are specified in LTM as the components of the particular Dialogue-game.

PRONOUN PROCESSORS -- The Dialogue Model System contains a set of Pronoun Processors, including an I-Processor, a You-Processor, and an It-Processor. Each of these is invoked whenever the associated surface word appears in an input utterance, and operates to identify some preexisting activation that can be seen as referring to the same object.

OBSERVATION METHODS FOR MODELING DIALOGUE

Most models of natural language are constructed using only the designer's own linguistic intuitions. and,evaluated by others using their own intuitions. This paper describes observation methods which have been used as a guide for the generation of process models of dialogue phenomena and as a basis for empirical evaluation of these models.

The methods have concentrated on a number of specific dialogue phenomena. Observers were given the annotation instructions, and after some training, produced highly reliable

annotations of:

Requests - Annotations of requests and the various kinds of responses to these requests by the other participant in a dialogue were useful for evaluating various control issues.

Repeated reference - Annotations of repeated reference were used to evaluate models of anaphoric reference (cases in which two sets of words in a dialogue describe the same thing).

Topic - Observers' annotations of the topic units in naturally-occurring dialogues provide a basis for specifying high level multi-utterance knowledge structures.

Expressions of comprehension - Observers annotated the expressions of comprehension (or lack thereof) by each participant of the other's previous utterances. These annotations are being used to evaluate our model of the stopping rule for comprehension, which determines when the model has successfully comprehended an utterance.

Similar expressions - Observers judged possible paraphrases of the utterances of dialogues for similarity in meaning in isolation and also in the context of the dialogue. This is serving as useful data for determining the effect of context on comprehension.

Correction actions - Observers' annotations of utterances that corrected some information conveyed previously in the dialogue allow us to evaluate our representation of the comprehension of the previous utterances.

While these phenomena don't span all possible issues of interest in natural language comprehension, this same general methodology can be extended to other language phenomena. It can provide a valuable guide for building natural language models, and an empirical basis for evaluating models.

A PROCESS MODEL FOR SPEECH ACT UNDERSTANDING

ALLEN MUNRO
UNIVERSITY OF CALIFORNIA, SAN DIEGO

In a conversation, hearers must make judgments about the intention of a speaker's utterance. A restaurant waiter realizes that when a patron says, "Waiter, this soup is cold," he is not merely thereby making an assertion, but is also making a complaint, and perhaps requesting rectification.

Utterances can be understood to have more than one intention. Two types of intentions are described in this paper, both of which depend largely on contextual factors for their recognition or understanding in a dialogue situation. Those of the first type (type A) are broadly applicable to many discourse situations and are understood primarily on the basis of the syntactic structure of utterances and certain facts about the status relationships that hold between the speaker and the hearer. Type A intentions are very similar to the Speech Acts of Searle or of Gordon and Lakoff. Speech acts of the second type (type B) are much more dependent on the nature of the particular contextual situation. They are recognized or understood on the basis of shared cultural norms concerning the particular kind of conversation the participants are engaged in. Type B speech acts are much less general in nature than type A. If a salesman says, "This little two-tone model has been priced to sell," then the customer should recognize that the type A intention is to make an assertion or a claim, and that the type B intention is to offer an item for sale.

A processing model for speech act recognition or understanding is presented, in which procedures called speech act schemata are activated. In the example with the salesman, the customer should have experienced the activation of both an assertion schema and a sales-offer schema The schemata are formulated in SOL, the high level language of the LNR memory model (Norman, Rumelhart, and LNR), which permits the modeling of conceptual entities which have both structural and procedural properties. In addition, the type B speech act schemata are subprocedures of larger information organizers which are sensitive to the particular context. These larger devices are called scripts and are related to the computational entities of that name developed by Schank and Abelson. In the example with the salesman, the sales-offer speech act is part of the "Commercial Transaction" script.

A detailed model for one such script, an "Airline Reservation Phone Call" script is presented, together with an account of the interactions between scripts, the two types of speech act schemata, and the syntactic structure of utterances.

NATURAL LANGUAGE UNDERSTANDING BY COGNITIVE NETWORKS

RICHARD FRITZSON
SUNY AT BUFFALO

A model of human cognition which stores and utilizes linguistic knowledge in the same structures as 'general' or 'world' knowledge supports an efficient and intuitively satisfying technique for the understanding of natural language. Cognitive comprehension and syntactic parsing can proceed in parallel, cooperatively, using the same procedures, without the loss of the syntax/semantics distinction.

In this paper, a system which uses such a representation is described. The representation is a cognitive network, a directed graph model of human cognition. The grammatical knowledge it contains is shown to be similar in structure to an augmented recursive transition network, but the understanding process is not ubilt around an ATN parser. Instead a "syntactically aided comprehension" procedure is used; cognitive processes, taking advantage of the uniform representation of knowledge, utilize linguistic knowledge as is required for tasks such as the disambiguation of conceptual relationships, and the location, in the network, of referents designated by phrases instead of names. Although the system has the syntactic knowledge needed to produce a correct parse for an utterance, it does not always have to apply it.

A few brief comparisons with other natural language understanding techniques will be included.

HOW TO REPRESENT AND USE KNOWLEDGE ABOUT CAUSALITY

CHARLES J. RIEGER
UNIVERSITY OF MARYLAND

A system for representing and using knowledge of cause and effect, called Commonsense Algorithms (CSA), will be described. The CSA project, ongoing for about a year now, has as its goal the unification of current ideas about problem solving with current ideas about contextual language comprehension. To this end, the CSA project currently consists of three parts, all built on top of the declarative CSA representation formalism: (1) a plan synthesizer, (2) a sentence-in-context interpreter (designed as the 2-sentence case of an eventual n-sentence story comprehender), and (3) a mechanism description "laboratory", whose purpose is to provide a framework for describing the "causal topology" of man-made devices and mechanisms. The LISP system which implements the CSA theory incorporates processes which are demon-like and processes which are more planful, and the CSA theory identifies how and why the planner and the population of demons interact. The presentation will cover the organization and access techniques for large numbers of CSA causal patterns in these two modes (planful and demonic), drawing on scenarios which illustrate the level of sophistication of the current implementation. One scenario will be taken from the children's story, The Magic Grinder (a Walt Disney book-of-the-month book), which the CSA group is employing to focus the language

comprehension part of the project. It is believed that any language comprehender must possess CSA-like knowledge, and that any theory of language comprehension must deal with issues quite similar to those encountered by plan synthesizers.

A HEURISTIC SEARCH APPROACH TO DISCOURSE ANALYSIS

JERRY R. HOBBS
CITY COLLEGE OF CUNY

This paper describes a computational approach to the problems of discourse structure of real-world English paragraphs, using the techniques of heuristic search. Procedures for detecting intersentence relations are being developed within the framework of a system for semantic analysis of English texts (1). This system takes as input the text in a predicate calculus-like notation produced by a syntactic front end (3, 2). Various semantic operations are applied:

1) Inferences are sought to satisfy the demants made by predicates on the nature of their arguments (predicate interpretation).

2) Redundancies are recognized and merged (knitting).

3) Anaphora are resolved.

The operations work by accessing a data base of world knowledge inferences, which are drawn selectively in response to the operations, and which carry a measure of their salience that varies with the context.

Intersentence relations are determined by matching the current sentence and the previous text against a small number of patterns. These patterns are stated in terms of inferences to be drawn from the current sentence and an "eligible" previous sentence, and the modification to be performed on the text if the pattern is matched. The patterns are of two

kinds--coordinating and subordinating. One coordinating relation is Temporal Succession: "The current sentence asserts a change whose initial state is implied by an eligible previous sentence." Other coordinating relations are Cause, Contrast, Parallel, and Paraphrase. Two subordinating relations are

1) Example: "The predicate and arguments of the assertion of the current sentence stand in a subset or element-of relation to those of the assertion of an eligible previous sentence." (See below.)

2) "Relative clause in disguise": "The current sentence provides further information about an entity fairly deeply embedded in an eligible previous sentence." (E.g. the second and third sentences of this abstract.) When this pattern is matched, the modification made to the text is equivalent to turning the current sentence into a relative clause.

As analysis proceeds, a tree-like structure is constructed for the paragraph, with subordinating relations building the tree downward and coordinating links building it to the right. A previous sentence is "eligible" if it is on the right frontier of this tree. For example, in a text S_1 S_2 S_3, if S_2 is a disguised relative clause on S_1, then S_3 may be coordinated with either S_1 or S_2, but if S_2 succeeds S_1 temporally, then S_3 may be coordinated only with S_2.

The heart of the operation is an ordered heuristic search of the data base for desired inferences. These

desired inferences are kept on a goal list, and have strengths associated with them. The strength of a goal depends on the type of text and the patterns found previously in the text. Also, the presence of conjunctions and some other elements advance certain patterns--"and" promotes Temporal Succession and a repetition of the previously recognized pattern, a dash and "i.e." promote Paraphrase, and the article "this" frequently signals a disguised relative clause. The evaluation function which orders the heuristic search is based on the strength of the goal, the length of the chain of inference, and the current salience of inferences in the chain.

When a <u>partial</u> match is found, the difference, or the remainder of the goal, is placed high on a goal list for subsequent processing. Consider the text

> "Republicans are discouraged about their prospects
>
> The party chairman is convinced that many GOP congressmen will lose their bid for reelection."

Suppose for simplicity "be discouraged" decomposes into "believe something bad will happen" We have a partial match with the Example pattern since the second sentence asserts that a particular Republican believes something. Hence a principal goal for our processing of the "that" clause is to show that what is believed is than an event bad for a Republican will occur. This is shown by accessing the fact about a political party that one of its purposes is to win elections.

Goals generated by a partial match are also passed on to subsequent sentences. In a Newsweek paragraph analyzed, the first sentence S_1 asserts a change. The next three sentences S_2 S_3 S_4 assert stages along the course of this change. S_2 is matched with the initial state of the change in S_1, thus generating as a goal the matching of S_3 and S_4 with succeeding states. Finally, the structure "S_2 then S_3 then S_4" is taken as an Example of S_1.

This work has significance for future attempts to summarize paragraphs automatically. For instance, in a text tied together by the Example pattern, the sentence "exampled" must be a major part of the summary.

Bibliography

1. Hobbs, Jerry R. A general system for semantic analysis of English and its use in drawing maps from directions. AJCL Microfiche 32, 1975.

2. Hobbs, J. R. and Grishman, Ralph. The automatic transformational analysis of English sentences: An implementation. International Journal of Computer Mathematics, to appear

3. Sager, Naomi, and Grishman, Ralph. The restriction language for computer grammars of natural language. CACM 18 (7), July 1975, p. 390.

ACM COMPUTER SCIENCE CONFERENCE

EMPLOYMENT REGISTER

MARRIOTT HOTEL - ATLANTA

JANUARY 31 - FEBRUARY 2, 1977

The Fifth Annual Register will provide books of listings of applicants and positions during the Conference.

Applicant listings include education, publications, experience, interests, references, position, and salary desired.

Employer listings include position available, starting date, salary, and benefits; education, experience, and specialization requirements.

Deadline: January 7, 1977

Fee: $20 for employers

$ 5 for applicants, plus $5 for anonymous listing

Free to students

Address: A form on which to submit information will be supplied by

Orrin E. Taulbee

Computer Science Employment Register

Department of Computer Science

University of Pittsburgh

Pittsburgh, Pennsylvania 15260

NATIONAL SCIENCE FOUNDATION
ADVISORY PANEL FOR LINGUISTICS

WILLIAM O. DINGWALL	University of Maryland
VICTORIA FROMKIN	University of California, Los Angeles
IVES GODDARD	Harvard University
ROGER SHUY	Georgetown University and CAL
CARLOTA SMITH	University of Texas
ARNOLD ZWICKY	Ohio State University

The Panel is expected to meet three times annually, in fall, winter, and spring. All proposals are to be reviewed at Panel meetings; proposals received too late for the spring meeting will be held until fall, possibly extending the customary six-month interval between submission and decision.

Program Director for Linguistics at NSF is Paul G. Chapin.

THIRD EUROPEAN CONGRESS ON

INFORMATION SYSTEMS AND NETWORKS

OVERCOMING THE LANGUAGE BARRIER

3 - 6 MAY 1977

LUXEMBOURG

EURONET, supplying scientific, technical, and economic information from diverse origins, will begin to operate in 1977. The Congress deals with the language problems raised by the network.

1. Teaching and utilization of languages in the European community.
2. Semantics, terminology, and lexicography, including equipment for rapid access to terminological resources.
3. Linguistics, grammar, and syntax, including computer-aided control and representation of syntactical structures
4. Human translation and interpretation, translation aids, and computer-aided or semi-automatic translation.
5. Multilingual thesauri and information retrieval systems.
6. Automatic translation.

A closing panel will discuss the evolution of multilingual systems from the viewpoint of policy makers and users.

Information: Loll Rolling, Information Management (XIII-B), Commission of the European Communities, European Center, Luxembourg.

NEW JOURNAL

TRANSACTIONS ON DATA BASE SYSTEMS

Publisher: ASSOCIATION FOR COMPUTING MACHINERY
P. O. Box 12105, Church Street Station
New York, New York 10249

Subscriptions: $40 for nonmembers of ACM
$15 for members

In the first issue: Selected papers from the International Conference on Very Large Data Bases, Framingham, Massachusetts, September 1975.

5TH INTERNATIONAL CONGRESS OF APPLIED LINGUISTICS

MONTREAL, AUGUST 21-26 1978

ORGANIZED BY THE CANADIAN ASSOCIATION OF APPLIED LINGUISTICS (CAAL)
UNDER THE AUSPICES OF THE
INTERNATIONAL ASSOCIATION OF APPLIED LINGUISTICS (IAAL)

Experts from all over the world will meet and discuss the latest developments in the many fields of applied linguistics, including :

First and Second Language Teaching and Learning

Bilingualism and Multiculturalism
Constrastive Linguistics
Translation
Lexicology
Computational Linguistics
Stylistics
Semiology
Communication Theory
Psycholinguistics
Neurolinguistics
Speech Therapy
Applied Phonetics
Sociolinguistics
Language Planning and Policy

Previous Congresses

NANCY	1964
CAMBRIDGE	1969
COPENHAGEN	1972
STUTTGART	1975

Information :

Jacques D. Girard
Secretary of the AILA Congress 1978
University of Montreal
Box 6128
Montreal 101, Canada

AMERICAN FEDERATION OF INFORMATION PROCESSING SOCIETIES

OFFICES AND HONORS

President 1976-77: THEODORE J. WILLIAMS
Professor of Engineering and Director
Laboratory for Applied Industrial Control
Purdue University

Chairman NCC Board: ALBERT S. HOAGLAND
Vice President and Manager
Advanced Recording Media
IBM San Jose Research Laboratory

(This Board conducts the National Computer Conferences)

Harry Goode Award. LAWRENCE G. ROBERTS
President, Telenet Communications Corp.

For the technical and organizational leadership under which the ARPA net was built.

Social Implications Committee Chairman:

HERBERT B. SAFFORD, CDP
Systems Supervisor
GTE Data Services, Inc.
Marina del Rey, California

ASSOCIATION FOR COMPUTING MACHINERY

OFFICERS - JULY 1, 1976 - JUNE 30, 1978

President	HERBERT R. J. GROSCH Consulting editor, Computerworld
Vice President	DANIEL D. MCCRACKEN Consultant
Secretary	GEORGE G. DODD General Motors Research Laboratories
Council	ROBERT L. ASHENHURST Institute for Computer Research, U. of Chicago
	ANITA COCHRAN Bell Laboratories
	RAYMOND E. MILLER Mathematical Science Department, IBM Research
	SUSAN H. NYCUM Chickering and Gregory, Attorneys, San Francisco

L A T S E C :

CONGRESSMAN QUESTIONS PAYMENT

A supplementary payment of $400,000 on an indexing contract was criticized on the floor of the House by Congressman Aspin of Wisconsin. According to the Congressional Record (May 4, H3909), Aspin said that the payment was made "under the provisions of a special bailout law (Public Law 85-804)" when Peter Toma "said that he would be 'impelled to leave the United States' without the aid."

The bailout law, according to Aspin, "should only be used when the Pentagon really needs equipment that is vital to the national security." The Latsec contract was one of several that Aspin characterized as not qualifying for the supplementary payments. He called for a new study of the whole concept.

Toma was quoted in newspaper reports as saying that "If Congressman Aspin were to look at our system and what it is doing for this country, he would withdraw his criticism." The newspaper story describes the basic contract as one with the Foreign Technology Division at Wright-Patterson Air Force Base, originally worth $460,000 for a three-year period.

THE FUTURE OF M T

RUDOLPH C. TROIKE, DIRECTOR
CENTER FOR APPLIED LINGUISTICS

There is a widespread myth among linguists that machine translation--or, properly, machine-aided translation--which was the object of intense effort and research a decade and a half ago, was found to be a failure and has since been abandoned. Nothing, in fact, could be further from the truth.

Although a number of institutions and agencies in the U. S. and elsewhere undertook extensive efforts in the late 1950s and early 1960s to develop computer programs for translation, only one, the Georgetown University program, succeeded in becoming fully operational (without requiring extensive pre- or post-editing). The Georgetown program was the ultimate basis for two of the major functioning MT programs in the U.S. today, that at Oak Ridge and at Wright-Patterson Air Force Base. These and other programs every year produce thousands of usable scientific and technical translations. However, they are all built on a research base which is now nearly twenty years old.

The 1966 report of the Automatic Language Processing Advisory Committee (ALPAC), which concluded that MT results had not been fully satisfactory, led to the virtual elimination of government

---

Reprinted with permission from The Linguistic Reporter, 18(9), May-June 1976, page 2.

support for MT research. While the conclusion was not strictly justified (for example, scientists at Oak Ridge and Euratom, given choice between human and machine translation, both opted for the latter), the reduction in funding was timely, since the extant programs had largely exhausted the then-available possibilities in computer technology and linguistics.

Unfortunately, much of the money spent on machine translation projects was applied to theoretical research rather than being used for translation--which was often disparaged as being merely practical and lacking in theoretical interest. It is therefore ironic that had more research been done directly on translation, the development of linguistic theory itself might have been accelerated by five to ten years. (Interestingly, transformational linguists, so often linked with computers in the public mind, had little involvement with MT.)

In the ten years since the ALPAC report, there has been considerable development in computer technology and in linguistics. The state of the art has advanced in both fields to the point where a new synthesis is now possible, which could produce greatly improved translations on a more cost-effective basis. (Unfortunately, one of the few projects in recent years to try this, at Berkeley, was curtailed last year for lack of funds.)

The time has now come for a new effort in MT to be undertaken. Properly conducted, such an effort would not only improve the quality and efficiency of translation, but would add to our knowledge of substantive universals and semantics, as well as

deepen our understanding of particular languages. MT can make an important contribution to the building of the information base on which the growth of linguistic theory must depend, at the same time that it produces a result of great practical value.

American Journal of Computational Linguistics *Microfiche 51 : 50*

MACHINE TRANSLATION:

MOSCOW INTERNATIONAL SEMINAR

I. I. OUBINE

Head, Machine Translation Department
All-Union Centre for Translation of Scientific
and Technical Literature and Documentation
Ul. Krzhizhanovskogo 14, korp. 1
117218 Moscow B-218

An International Seminar on Machine Translation took place in Moscow from 25 to 27 November, 1975. About 200 scientists from People's Republic of Bulgaria, the German Democratic Republic, Czechoslovak Socialist Republic, People's Republic of Poland, and the Soviet Union participated in the Seminar which was organized by the All-Union Center for Translation of Scientific and Technical Literature and Documentation.

In his opening speech the head of the All-Union Center for Translation (ACT), Dr. V. N. Gerasimov, said that the expansion of interational contacts and internationalization of science had promoted the growth of the translation activities in the USSR. In the country at present translation is being done by various specialized and departmental organizations. In 1975, for example, the ACT translated more than 30,000 author's sheets of scientific and technical literature and documentation. In the

nearest future ACT alone will reach the volume of 50-80,000 printer's sheets a year. Dr. Gerasimov sees the way out of this situation in the speediest possible development and installation of industrial systems of machine translation. This was also the conclusion of the Temporary Scientific and Technical Commission on Machine Translation of the State Commission on Machine Translation of the State Committee on Science and Technology of the Council of Ministers of the USSR which worked in 1972-1973

At the plenary session, four reports dealing with general principles of construction of machine translation systems and ways of improvement of MT quality were delivered. Yu. N. Marchuk (ACT, Moscow) draws our attention to the increased role of computer dictionaries in automatic data processing systems and namely in MT systems. Since the quality of MT largely depends on the word-list and volume of the dictionary and the information contained in the dictionary entries, these dictionaries must, according to Marchuk, be compiled with due regard to distributional and statistical methods. It is important to make wide use of contextual information. Inter- and post-editing are absolutely indispensable in running the first industrial MT systems.

This idea is developed in a joint report by B. D. Tikhomirov, Yu. N. Marchuk, and I. I. Oubine (ACT, Moscow), who put forward a new approach to inter- and post-editing for correction of input and translation errors. Interediting follows a search in the

dictionary, automatic analysis of new words and formation of the information required for future processing. The intereditor either confirms or corrects the grammatical information of new words and can also correct input errors, supply translations for new words, etc. At the intereditor's command the system accumulates new words for the posteditor.

The posteditor exercises phrase-by-phrase control of the translation. If the translation of a phrase does not satisfy him he edits it with the help of a display unit and, if necessary, sends the original phrase and its machine translation to a special storage device for further analysis aimed at the improvement of the MT algorithm.

In his summarizing report, Dr. A. Ljudskanov (PRB) singled out the main stages of MT development, and showed what an unfavorable effect the ALPAC report and the pessimistic statements of Dr. Y. Bar-Hillel had produced on MT. MT systems aimed to operate on an industrial scale in the nearest future are to be developed on the basis of the "selective" strategy, i.e. they are to use only that information "which is necessary and sufficient for the given aim for the given pair of languages." Dwelling upon this thought, Dr. Ljudskanov said in his second report, "Lexeme dictionary for MT systems", that in the MT system which is under development in Bulgaria "deep" difficulties are shifted to "surface" levels (lexical and morphological) and solved with the help of context analysis.

The fourth report at the plenary session was read by Dr. R. G. Piotrovsky, who thinks that for the time being those who develop MT systems must not strive to achieve 100% efficiency of text processing and these MT systems must be based not on deductive generative grammar but on deductive-inductive linguistics of text. According to the linguistics of text, MT systems should be developed consecutively from simple algorithms operating with units of the surface levels of the text system to more complicated algorithms oriented to the deep levels of the text systems. Dr. Piotrovsky considers that vocabulary contains the lion's share of syntactic and semantic information of the text and consequently the basis of any MT system must be a bilingual automatic dictionary with the output word list of the thesaurus type: Such MT systems can be developed within reasonably short periods of time and will, according to Dr. Piotrovsky, enable the consumer to derive all semantic information from the input text.

Thus the main speakers put forward as the cardinal task of the present day the creation of MT systems not yet oriented to high-quality translation but working industrially

More than 60 reports were made at the seminar which were distributed into four sections: (a) computer dictionaries; (b) automatic analysis and synthesis of texts; (c) semantic analysis of texts; (d) mathematical and program maintenance.

Various trends in the theory and practice of machine translation were represented at the section on computer dictionaries. A number of reports were read by members of the Speech Statistics

group (the work is under the scientific guidance of Dr. R. G. Piotrovsky). Scientists of this group compile dictionaries for MT within the framework of one scheme in accordance with the postulates of the linguistics of text; their MT dictionaries consist of two parts: a dictionary of commonly used words and changeable terminological dictionaries for various fields of science and technology. The input entries of these dictionaries are either single word forms or combinations of several word forms. Morphological and syntactic information is written straight in the dictionary entries. The volume of syntactic information is relatively small. This information includes indices of grammatical word-class, transitivity-intransitivity for verbs, type of government and some other syntactic characteristics. Dictionaries of input and output languages have charts for coding information and correspondence charts which in this MT system are the principal linguistic algorithm. All programs of automatic processing of linguistic information are based on these charts. The authors of these reports pay great attention to methods of coding information and compressing codes in the computer. The computer realization of such dictionaries makes it possible to get an interlinear version which is not actually a translation but gives the consumer the main contents of the original text. During the discussion, many participants of the seminar, criticizing this approach to MT, pointed out that automatic dictionaries should not be compiled separately but as component parts of MT systems. The coding of lexical information

in an MT system should be preceded by the elaboration of blocks of grammatical analysis and synthesis which use lexical information and in their turn may affect its composition and coding.

A different approach to principles for development of MT systems was presented by Z. M. Shaljapina in "The ARAT system: dictionary, grammar, and their use in automatic analysis" (Moscow, State Institute of Foreign Languages). The central component of this Anglo-Russian Automatic Translation system (ARAT) is a dictionary of a special type (Anglo-Russian Multiaspect Automatic Dictionary--ARMAD) in which every entry gets complete and diverse characteristic properties of its linguistic behavior on all levels of description: morphological, lexical, syntactic, and semantic. For formal recording of all required information, ARMAD employs syntactic patterns, lexical functions, demands imposed by the word on the linear and structural context, rules of standard and individual modifications of lexical and syntactic structures with this word, formal semantic representation, semantic selection of syntactic valencies of the key word, etc.

The abundance and variety of information contained in the dictionary entry of practically every word is the principal advantage of ARMAD as compared with other types of automatic dictionaries. Wide use of syntactic and expecially semartic information makes it possible with the help of this dictionary in the framework of the ARAT system to solve such a cardinal problem of MT as grammatical and lexical ambiguities. Data on lexical co-occurrence of the original English word and its Russian equivalent

as well as on required syntactic transformations, it is hoped, will ensure an exact and idiomatic translation of the original text But at the same time the abundance and variety of the information in ARMAD naturally leads to a more complicated entry structure and multiplies the difficulties which linguists have to overcome when writing entries. At present the difficulties and possibilities of algorithmic realization of full-scale dictionaries of this type have not been investigated.

A number of reports were devoted to compiling various types of auxiliary dictionaries and creation of whole systems for automatic lexicographic work. In their reports, Yu. N. Marchuk, N. G. Tikhonova, and I. I. Oubine (all of ACT, Moscow) informed the participants of the seminar of the work now being done to develop a bilingual automatic lexicographic system to compile frequency and semantic frequency dictionaries and bilingual concordances as auxiliary material for MT dictionaries.

E. V. Vertel and V. A. Vertel (State Institute of Foreign Languages, Minsk) submitted a joint report on the elaboration of an algorithm and a set of programs to compile a frequency-alphabetic dictionary and concordance on a medium-size computer. The set of programs is designed to process an original text of up to 300,000 words.

An interesting report on the compilation of a reverse French dictionary was submitted by E. L. Kozmina (Applied Mathematics Institute, Moscow), who worked out a method of getting a wordform by its ultima or by one wordform representing a whole class of

words inflecting uniformly. This method makes it possible to reduce a reverse dictionary of word-forms to the size of an ordinary dictionary.

In this session the participants in the Seminar discussed the principles of compilation of MT dictionaries and came to the conclusion that it is necessary:

- to consider an MT system as a whole in which all parts are interlocked, while the dictionary is its principal component;

- to design MT systems for sublanguages and consequently compile automatic dictionaries for well-defined topics;

to pay special attention to strict selection of entries for automatic dictionaries, and for this purpose compile various auxiliary dictionaries and concordances;

- to increase the volume of diverse syntactic, semantic, and lexical information in automatic dictionaries;

- to attach great importance to the structure of automatic dictionaries. According to a number of the participants of the seminar, the optimum structure is a combination of a dictionary of stems and a dictionary of wordforms. The dictionary of word-forms contains more frequent words and the dictionary of stems contains the less frequent words.

The section on automatic analysis and synthesis of texts attracted the largest number of reports, reflecting the intense interest of linguists in these important parts of MT systems. A series of interconnected reports was read by scientific workers of the Computational Center of Leningrad State University Dr.

G. S. Tseitin proposes to use for automatic syntactic analysis models which do not simply contain a set of rules for construction (or "filtering") of admissible syntactic structures but also mark out among these rules more or less "preferable" ("normal", "nuclear", "productive"). B. M. Leikina (Leningrad State University) is working on a model of an English grammar which permits under certain limitations nonprojectivity while fulfilling the projectivity demands for the majority of cases. At present this group is working on a more complicated model which takes into account among other things the order of establishment of syntactic links in a particular structure. A number of experiments have been carried out with the first variant of this grammar. These experiments were aimed at checking the presence of the correct analysis and absence of a fixed incorrect analysis and were carried out under the conditions of man-machine interaction during which a part of the intermediate results were rejected by the man. Complete analysis has been tested only on several sentences.

The principles of analysis in the ARAT system were expounded in this section. Besides the report of Z. M. Shaljapina already mentioned, three reports on less general problems were read: L. A. Afonasjeva, "Disambiguation of homonymy in the ARAT system", T. N. Nikanorova, "Prepositions in the ARAT system", and O. A. Sternova, On a model of Russian inflexion in the ARAT system" (all State Institute of Foreign Languages, Moscow). In this MT system the disambiguation of homonymy is not singled out into a

separate block but is carried out during the analysis simultaneously with the fulfillment of other tasks. Prepositions in the ARAT system are treated like any other lexical unit of the text. Meaningless prepositions, i.e. prepositions serving as surface-syntactic indications of strong government and having no influence on the meaning of the text are eliminated during the transition from combined syntactic structure (CSC) to semantic representation. Meaningful prepositions, i.e. prepositions having their own meaning, are elements of the semantic structure and are preserved during the transition from CSC to semantic representation.

E. E. Lovtsky (ACT, Moscow) proposed formal means for the description of the syntax of natural languages. The description is an oriented graph, in the nodes of which are symbols of syntactic classes and of subgrap The information on the dependency structure of strings is recorded on the edges of the graph. The description of a natural language syntax, made with the help of the suggested formal means, can be used in a system of automatic syntactic analysis. For the analysis of a phrase, it is necessary to find in the graph all occurrences of a string of syntactic classes corresponding to this phrase and to read from the edges of the graph the information about the structure of the string. This procedure is done automatically by a special algorithm. As a result of the analysis we get immediate constituent trees and dependency trees of the analyzed phrase. In case several trees correspond to the analyzed phrase then the

choice of the tree corresponding to the lexical units of the phrase is made on the basis of the information contained in the automatic dictionary. The said description is being done for the analysis of English scientific and technical texts.

The joint report of E. Benesova, A. Bemova, S. Machova, and J. Panevova (Czechoslovakia) contains characteristics of cardinal components and principles of functional generative description designed to get semantic representations which in their turn are the basis for synthesis of sentences of a natural language.

Other reports at the section on automatic analysis and synthesis of texts were of a less general character. Dr. Klimonov (GDR) presented a set of criteria for automatic identification of antecedents of personal and possessive pronouns in Russian and German. Dr. Starke (GDR) read a report on a method of transformation of Russian structures with German equivalents having different dependency representations. Each transformation is performed by means of elementary operations: insertion, deletion, alteration of a node, etc.

The reports submitted to the section on semantic analysis touched on some narrow problems. A number of reports show their author's efforts to solve their problems within the framework of automatic text processing. These reports are the definition of the concept of answer in question-answer structures of texts (Dr. Konrad, GDR); semantics of prepositions (Leontjeva, ACT, Moscow) and Nikitina (Institute of Linguistics, Academy of

Sciences, Moscow); semantic and syntactic analysis of Russian words with the meaning "quantity" (Iljin and Smirnova, Leningrad State University); communicative structure of the English sentence (Korolev, ACT, Moscow); semantic analysis of headings of Japanese patents (Zeinalova, Samarina, and Shevenko, Institute of Oriental Studies, Academy of Sciences, Moscow).

An interesting report was submitted by N. A. Kuzemskaya and Dr. E. F. Skorokhodko (Institute of Cybernetics, Kiev), who reported work on quantitative criteria for estimation of translation quality. They propose to solve this problem with the help of a semantic language which would meet the following conditions: (a) ability to express the required information; (b) possibility of getting the necessary quantitative estimates; (c) possibility of translating into this semantic language from the original language and from the language of translation.

Semantic networks can be used as such a semantic language. Comparison of two speech semantic networks--the primary corresponding to the original text and the secondary to the text of the translation--makes it possible to define a number of parameters characterizing various aspects of the quality of the translation. The main aspects of translation quality are completeness and accuracy. Completeness is calculated from the proportion of the primary network contained in the second, and accuracy from the proportion of the secondary that coincides with the primary.

The problem of automatic recognition of key words in text was the main concern of the reports of R. A. Kovalevitch, V. A. Sorkina (State Institute of Foreign Languages, Minsk), and G. S Osipov, A. I. Chaplja (Mahachkala State University). Kovalevitch and Sorkina singled out more than 50 elementary semantic units and, relying on formal characteristics of the text and using distributional statistical methods, they singled out and formalized rules of combination of elementary semantic units in free word-groups. The result of the analysis is a matrix of relations of components in English word groups and their Russian translations with all necessary grammatical information. The set of formal characteristics of all components of a free English word group determines the corresponding Russian equivalent. In their report, Osipov and Chaplja spoke about an attempt to construct (by the thesaurus method) formal means useful for elaboration of an algorithm of recognition of key words in a text. These authors build this formal apparatus on the basis of estimating the degree of synonymy of words in the structures of the input text.

To our regret it should be admitted that in spite of the universally recognized importance of the semantic component for MT, the seminar lacked reports in which authors put forward methods and instruments of formal representation of word meanings for the solution of problems of analysis and synthesis of texts during MT. The participants of the seminar did not pay sufficient attention to the search for optimum ways of elaborating formal semantic languages for MT.

In the software section, ten reports were made, three of which are characterized by a broad approach to this important part of MT systems. These are the reports of N. A. Krupko and G. S. Tseitin (Leningrad State University), V. S. Krisevitch and I. V. Sovpel (Institute of Foreign Languages, Minsk), and D. M. Skitnevsky (Institute of Foreign Languages, Irkutsk)
The joint report by Krupko and Tseitin sums up the research work carried out during the last 13 years, involving development of software to be used in computer experiments on MT. At present the software of the MT model at the MT laboratory of Leningrad State University comprises the following components: (a) the system of symbolic representation of the linguistic content of the model; (b) the machine representation of the linguistic content; (c) the compiler for translating the symbolic representation into the machine representation; (d) the interpreter which employs the machine representation to process a given text in accordance with the aim of the experiment. Such a structure of the software is motivated mostly by the necessity of changing the linguistic content in the course of experiments and by the impossiblity of keeping in core memory more complicated linguistic information in the form of a compiled program. The same approach with an appropriate shift in the structural role of each of the component parts seems, in the opinion of the authors of the report, to be quite reasonable in developing practical MT systems.

The report by Skitnevsky considers some basic principles of software development for linguistic investigations. The main idea of Skitnevsky's model is to consider as interdependent the following three levels of software: (a) the conceptual level--a model of linguistic process; (b) the compositional level--an input language; (c) the programming level--the implementation system. The development of software includes the following five stages: the first stage is an analysis of a set of algorithms which are given in their verbal description (i.e. as flow charts). The analysis results in an informal definition of the class of algorithms. The second stage consists in specifying this formal definition. As a result we have a logical and mathematical framework, which is a sufficiently precise and dynamic description of basic concepts occurring in the informal definition (MODEL). At the third stage a data bank and a service program package using the standard software are formed on the basis of the set of objects and operations of the model. The fourth stage consists of working out an input language which permits the user to state his tasks in terms of the MODEL. The fifth stage designs an operational system which uses the standard software and remains intact in respect to relatively extended MODEL, its data bank and service program package. The MODEL built along these lines and an input language form a convenient metalanguage for the description of the computer semantics of natural languages.

The report by Krisevitch and Sovpel emphasizes that an MT system operates with large information files including input and

output, automatic dictionaries containing grammatical and semantic information; therefore the efficiency of an MT system depends to a great extent on the degree of optimization of information processing. The authors believe that the efficiency of use of storage and the minimization of average access time are determined by such factors as the nature of information, the request frequency, and the properties of storage devices. In keeping with these principles, the authors constructed their model of the information base for an MT system. The other reports discuss less general problems of software relating to automatic text processing. The report by N. G. Arsentjeva (ACT, Moscow) describes the results of the machine experiment which was aimed at forming surface syntactical characteristics of a word on the basis of its semantic description. The input information is a semantic description of some word, represented as a formula containing semantic elements and operation signs. The author has devised an algorithm which supplies the linguist with a list of possible realizations in the text of the syntactic pattern of the processed word on the basis of the semantic description of the word, its syntactic pattern and some other linguistic information. The program is written in the algorithmic language of recursive functions. The experiment shows that the said language is quite applicable for linguistic purposes.

S. A. Ananjevsky and P I Serdukov (Institute of Cybernetics, Kiev) informed the seminar of their work on the development of a system of automatic syntactic analysis of text for purposes of

automatic abstracting, information retrieval, and MT. The system permits one to obtain the following characteristics of the English verb: search area (zone of syntactic dependents of a verb in a sentence), position of a verb (predicate), which is the initial point of the search; the immediate distribution of the verb; syntactical compatibility of the verb; syntactic compatibility of the dependents (string of verb dependents). This system is implemented on the Minsk-32 and permits one to obtain a complete formal description of syntactical dependencies of the English verb.

The report by L. N. Beljaeva and E. M. Lukjanova (Gertsen Pedagogical Institute, Leningrad) formulates principles of compiling an automatic polytechnical dictionary for MT. The dictionary is a system of sets of linguistic data as well as algorithmic, linguistic, and programming units of a Russian text, different for different linguistic algorithms. The structure of such a polytechnical dictionary gives the possibility of organizing this system as a linguistic data bank. E. V. Krukova and T. I. Gushchina reported the results of the application of the PL/I language for an automatic analysis of texts. The authors have written a number of programs in PL/I, designed to compile frequency and reverse dictionaries, for morphological analysis of Russian words, and for resolution of ambiguity of French nouns. In their opinion, PL/I is well suited to the purposes of automatic text processing.

Summing it up, it should be noted that the reports submitted to this section varied considerably from the point of view of the problems discussed and the methods employed due to lack of uni formity in using programming languages and types of computers

The main characteristic feature of the seminar, in our opinion, is an emphasis on developing practical systems of MT for well-defined sublanguages of natural languages rather than devising global language models of which MT is only a component. The most essential thing now is not to carry out experiments but to develop practical systems of MT operating on a large scale.

At the concluding session the participants of the seminar adopted a memorandum which admitted the importance of the seminar for the development of practical MT systems and appealed to ACT and the International Centre of Scientific and Technical Information to hold another seminar in Moscow in 1977. The participants of the seminar think it necessary to promote further coordination and integration of research in MT in the USSR as well as in the framework of the International System of Scientific and Technical Information. They find it reasonable to determine the main body responsible for problems and to entrust AGT with this important task.