[*Mechanical Translation*, Vol.7, no.2, August 1963]

Abstracts of Papers for the 1963 Annual Meeting of the Association for Machine Translation and Computational Linguistics

Denver, Colorado, August 25 and 26, 1963

Necessity of Introducing Some Information Provided by Transformational Analysis into MT Algorithms

Irena Bellert

Department of English Philology, Warsaw University

A few examples of ambiguous English constructions and their Polish equivalents are discussed in terms of the correlation between their respective phrase-marker representations and transformational analyses. It is shown by these examples that such an investigation can reveal interesting facts for MT, and therefore should be carried out for any pair of languages for which a given MT program is being constructed.

If the phrase-marker of the English construction is set into one-to-one correspondence with the phrase-marker of the Polish equivalent construction, whatever particular transformational analysis of this construction is to be taken into account, then the ambiguous phrase-marker representation can be used as a syntactical model for MT algorithms with good results.

If the phrase-marker of the English construction is set into one-to-many correspondence with the phrase-markers of the Polish equivalents, according to the transformational analyses of this construction, then the ambiguous phrase-structure representation has to be resolved in terms of transformational analysis, for only then is it possible to assign the corresponding phrase structure representation to the Polish equivalents.

A tentative scheme of syntactical recognition is provided for the multiply ambiguous adjectival construction in English[1] (which proved to belong to the latter case) by means of introducing some information obtained from the transformational analysis of this construction.

The Use of a Random Access Device for Dictionary Lookup

Robert S. Betz and Walter Hoffman

Wayne State University

The purpose of this paper will be to present a scheme to locate for single textual items and idioms in textual order their corresponding dictionary entries stored in an IBM 1301 random access mechanism.

Textual items are considered to be 24 characters in length (left justified with following blanks). A dictionary entry consists of a 24 character Russian form, grammar information for the form and a set of translations for that form. Dictionary entries are packed into sequential tracks of the 1301. This paper will cover the method used for dictionary storage.

The lookup for a textual item I first consists of a search for the first track that the dictionary entry E (if one exists) for I could be stored in. Once a track has been determined its contents are searched in core by a bisection convergence technique to find E. If E cannot be found, a "no entry" indication is made.

If E is found a further search is made of the dictionary to find the longest sequence of text, starting with the first item I, that has a dictionary entry. The last such entry found is picked up.

Included in the presentation will be examples of the dictionary lookup output for actual text.

1 cf. the paper by Robert B. Lees, "A Multiply Ambiguous Adjectival Construction in English", Language 36(1960).

Generative Processes for Russian Impersonal Sentences

C. G. Borkowski and L. R. Micklesen

IBM Thomas J. Watson Research Center

Impersonal sentences of Russian are those traditionally construed to consist of predicates only. Ever since the first Russian grammar was compiled, they have continued to pose a problem for grammarians. This paper is intended to be a review and evaluation of all types of the so-called impersonal sentences in the Russian language. The investigation of these sentences has been conducted in terms of their relationships to basic (kernel) sentences. Our paper attempts to define the origin for such impersonal sentences, i.e., how such sentences might be derived within the framework of a generative grammar from a set of rules possessing maximal simplicity and maximal generative power. The long-range aim of this investigation involves the most efficient manipulation of such sentences in a recognition device for Russian-English MT.

Concerning the Role of Sub-Grammars in Machine Translation

Joyce M. Brady and William B. Estes

Linguistics Research Center, The University of Texas

The comprehensive grammars being developed at the Linguistics Research Center of the University of Texas will be too large for easy access and manipulation in either experimental programs or practical translation. It is necessary, therefore, to devise some reliable method for selecting subsets of the grammar rules which will be reasonably adequate for a given purpose. Since

the majority of the rules are dictionary rules, this problem is closely related both to the problem of constructing microglossaries and to the subsequent problem of choosing a particular microglossary suitable to a given text.

Our current approach to this problem entails the construction of key word lists in the first stage of analysis which guide the computer in its choice of a previously constructed microglossary. Work to date indicates adaptations of this technique may not only contribute to the solution of storage and access problems but also facilitate analysis and simplify problems of semantic resolution.

Word-Meaning and Sentence-Meaning*

Elinor K. Charney

Research Laboratory of Electronics, Massachusetts Institute of Technology

A theory of semantics is presented which (1) defines the meanings of the most frequently occurring semantic morphemes ('all', 'unless', 'only', 'if', 'not', etc.), (2) explains their role, as semantically interdependent structural-constants, in giving rise to sentence-meanings, (3) suggests a possible approach to a sentence-by-sentence recognition program, and (4) offers a feasible method of coordinating among different language systems synonymous sentences whose grammatical features and structural-constants do not bear a one-to-one correspondence to one another. The theory applies only to morphemes that function as structural-constants and their interlocking relationships, denotative terms being treated as variables whose ranges alone have structural significance in sentence-meaning. The basic views underlying the theory are: In any given sentence, it is the particular configuration of structural-constants in combination with specific grammatical features which produces the sentence-meaning; the defined meaning of each individual structural-constant remains constant. The word-meanings of this type of morpheme, thus, must be carefully distinguished from the sentence-meanings that configuration of these morphemes produce. Sentence-synonymy is not based upon word-synonymy alone. Contrary to the popular view that the meanings of all of the individual words must be known before the sentence-meaning can be known, it is shown that one must comprehend the total configuration of structural-constants and syntactical features in a sentence in order to comprehend the correct sentence-meaning and that this understanding of the sentence as a whole must precede the determination of the correct semantic interpretation of these critical morphemes. In fact, the structural features that produce the sentence-meanings may restrict the possible meanings of even the denotative terms since a structural feature may demand, for example, a verbal rather than a noun phrase as an indispensable feature of the configuration. Two or more synonymous sentences whose denotative terms are everywhere the same but whose structural configurations are not isomorphic express the same fundamental sentence-meaning. The fundamental sentence-meanings can be explicitly formulated, and serve as the mapping functions to co-ordinate morphemically-unlike synonymous sentences within a language system or from one system to another. The research goal of the author is to establish empirically these translation rules that state formally the structural characteristics of the sentence configurations whose sentence-meanings, as wholes, are related as synonymous.

Translating Ordinary Language into Symbolic Logic*

Jared L. Darlington

Research Laboratory of Electronics, Massachusetts Institute of Technology

The paper describes a computer program, written in COMIT, for translating ordinary English into the notation of propositional logic and first-order functional logic. The program is designed to provide an ordinary language input to a COMIT program for the Davis-Putnam proof-procedure algorithm. The entire set of operations which are performed on an input sentence or argument are divided into three stages. In Stage I, an input sentence 'S', such as "The composer who wrote 'Alcina' wrote some operas in English," is rewritten in a quasi-logical notation, "The X/A such that X/A is a composer and X/A wrote Alcina wrote some X/B such that X/B is an opera and X/B is in English." The quasi-logical notation serves as an intermediate language between logic and ordinary English. In Stage II, S is translated into the logical notation of propositional functions and quantifiers, or of propositional logic, whichever is appropriate. In Stage III, S is run through the proof-procedure program and evaluated. (The sample sentence quoted is of course 'invalid', i.e. non-tautological.) The COMIT program for Stage III is complete, that for Stage II is almost complete, and that for Stage I is incomplete. The paper describes the work done to date on the programs for Stages I and II.

The Graphic Structure of Word-Breaking

J. L. Dolby and H. L. Resnikoff

*Lockheed Missiles and Space Company***

In a recent paper[1] the authors have shown that it is possible to determine the possible parts of speech of

* This work was supported in part by the National Science Foundation, and in part by the U.S. Army Signal Corps, the Air Force Office of Scientific Research, and the Office of Naval Research.

** This work was supported by the Lockheed Independent Research Program.

[1] "Prolegomena To a Study of Written English," J. L. Dolby and H. L. Resnikoff.

English words from an analysis of the written form. This determination depends upon the ability to determine the number of graphic syllables in the word. It is natural, then, to speculate as to the nature of graphic syllabification and the relation of this phenomenon to the practice of word-breaking in dictionaries and style manuals.

It is not at all clear at the start that dictionary word-breaking is subject to any fixed structure. In fact, certain forms cannot be broken uniquely in isolation since the dictionary provides different forms depending upon whether the word is used as a noun or a verb. However, it is shown in this paper that letter strings can be decomposed into 3 sets of roughly the same size in the following manner: in the first, strings are never broken in English words; in the second, the strings are always broken in English words; and in the third, both situations occur. Rules for breaking vowel strings are obtained by a study of the CVC forms. Breaks involving consonants can be determined by noting whether or not the consonant string occurs in penultimate position with the final *c.* The final *e* in compounds also serves to identify the forms that are generally split off from the rest of the word.

A thorough analysis is made of the accuracy of the rules given when applied to the 12,000 words of the Government Printing Office Style Manual Supplement on word-breaking. Comparisons are also drawn between this source and several American dictionaries on the basis of a random sample of 500 words.

Writing of Chinese Recognition Grammar for Machine Translation

Ching-yi Dougherty

University of California, Berkeley

Our approach to this problem is based on the stratificational grammar outlined and the procedures proposed by Dr. Sydney Lamb. How the theory and the procedures can be applied to written Chinese is briefly discussed. For the time being our research is limited to the particular kind of written Chinese found in chemical and biochemical journals. First the Chinese lexes are classified by detailed syntactical analysis, then binary grammar rules are constructed for joining two primary or constitute classes. How a more and more refined classification can eliminate one by one the ambiguity resulting from all possible constructions arising from juxtaposition of two distributional classes is discussed in detail.

The Behavior of English Articles

H. P. Edmundson

Thompson Ramo Wooldridge Inc.

Machine translation has often been conceived as consisting of three steps: analysis of source-language sentence, transformation of analyzed pieces, and synthesis of target-language sentence. This paper is concerned with one aspect of the last step, namely, the rules of behavior of English articles. Since the classical definitions of definite and indefinite articles are operationally imprecise, proper mechanistic rules must be formulated in order to permit the automatic insertion or non-insertion of English articles. The rules discussed are of syntactic origin; however, note is also taken of their semantic aspects. This paper describes the methods used to derive these rules and offers ideas for further research.

On Representing Syntactic Structure

E. R. Gammon

Lockheed Missiles and Space Company

The idea of sentence depth of Yngve (A Model and an Hypothesis for Language Structure, *Proc. Am. Phil. Soc*., Vol. 104, No. 5, Oct. 1960) is extended to the notion of "distance" between constituents of a construction. The distance between constituents is defined as a weighted sum of the number of IC cuts separating them. Yngve's depth is then a maximum distance from a sentence to any of its words.

Various systems of weighting cuts are investigated. For example, in endocentric structures we may require that the distance from an attribute to the structure exceeds the distance from the head to the structure, and in exocentric structures that the distances from each constituent to the structure are equal.

Representations of constructions are considered which preserve the distance between constituents. It is shown that it is impossible to represent some sentences in Euclidean space with exact distances, but a representation may be found if only relative order is preserved. If more general spaces are used then exact distances may be represented. It follows that for a wide class of sentence types, there is a weighting, and a space, in which the distance preserving representations are identical with the diagrams of traditional grammar.

La Traduction Automatique et l'Enseignement du Russe

Yves Gentilhomme

Centre National de la Recherche Scientifique, Paris

Les recherches effectuées depuis quelques années en vue de la Traduction Automatique ont conduit à des méthodes de travail et à des résultats intéressant la pédagogie des langues.

Une expérience d'enseignement du russe a l'usage des scientifiques fondée sur ces données a été poursuivie pendant deux ans à Paris (Centre National de la Recherche Scientifique et Faculté des Sciences), et a abouti à la publication d'un manuel.

Le present compte-rendu a pour objet de préciser les principes généraux utilisés, la réaction des étudiants et le rendement pédagogique obtenu.

1. *Graphes morphologiques:* Les mots d'une même famille. Notion de base. La double ramification. Les graphes abstraits. Les néologismes scientifiques.
2. *Graphes syntaxiques:* La double structure d'une phrase. Multiplicité des modèles. Point de vue psychologique. Notion de fonction. Continuité et discontinuité.
3. *Les séparateurs:* La segmentation d'une phrase. Le vocabulaire prioritaire.
4. Théorie de la valence: macro et microcontexte. Qu'est-ce-que "connaître un mot"?
5. Point de vue de l'étudiant; point de vue du traducteur humain; et point de vue de l'Enseignant.

Word and Context Association by Means of Linear Networks

Vincent E. Giuliano

Arthur D. Little, Inc.

This paper is concerned with the use of electrical networks for the automatic recognition of statistical associations among words and contexts present in written text. A general mathematical theory is proposed for association by means of linear transformations, and it is shown that this theory can be realized through use of passive linear electrical networks. Several small-scale experimental associative networks have been built, and are briefly described in the paper; one such device will be demonstrated in the course of the oral presentation of the paper. Some of the devices generate measures of association among index terms used to characterize a document collection, and between the index terms and the documents themselves. Another uses syntactic proximity within sentences as a criterion for the generation of word association measures. Examples are given of associations produced by these network devices. It is conjectured that the network-produced association measures reflect two distinct types of linguistic association—"synonymy" association which reflects similarity of meaning, and "contiguity" association which reflects real-world relationships among designata.

A Study of the Combinatorial Properties of Russian Nouns

Kenneth E. Harper

Rand Corporation

A statistical study was made of the extent to which Russian nouns enter into certain kinds of syntactic combination. The basis of the study was a corpus of 180,000 running words of Russian physics text prepared for analysis by the Automatic Language Data Processing group at The Rand Corporation; for each sentence of text the syntactic dependency of each word had been previously coded. A data retrieval program was applied, showing for each noun in text the number of occurrences (a) with at least one genitive noun dependent, (b) with at least one adjective dependent, and (c) with either type of dependent. A listing of all nouns in text (64,026 occurrences of 2,993 nouns) was prepared, ordered by frequency, and showing counts for a, b, and c above. Separate listings were prepared, showing for each noun that occurred 50 times or more the probability P that it would be modified in each of these three ways; these listings were ordered on P.

The data suggests, among others, the following conclusions: there is statistical significance in the variability with which nouns enter into the given combinations; the partial interchangeability of adjective and genitive noun modification is supported; a general correspondence exists between combinatorial groupings of nouns and morphological or semantic groupings (concrete nouns have low P for genitive complementation, abstract nouns have high P, etc); the use of words in a given field of discourse can be determined empirically (*e.g.,* the use of deverbative nouns either to indicate a process or the result of a process). It is suggested that the distributional approach is a useful supplement to traditional syntactic and semantic classification schemes, and that it is of direct utility in automatic parsing programs.

Connectability Calculations, Syntactic Functions, and Russian Syntax

David G. Hays

*Common Research Center, EURATOM, Ispra**

A program for sentence-structure determination can be divided into routines for analysis of word order and for testing the grammatical connectability of pairs of sentence members. The present paper describes a connectability-test routine that uses the technique called *code matching.* This technique requires elaborate descriptions of individual items, say the words or morphemes listed in a dictionary, but it avoids the use of large tables or complicated programs for testing connectability. Development of the technique also leads to a certain clarification of the linguistic concepts of *function, exocentrism,* and *homography.*

In the present paper, a format for the description of Russian items is offered and a program for testing the connectability of pairs of Russian items is sketched. This system recognizes nine dominative functions: subjective; first, second, and third complementary; first, second, and third auxiliary; modifying; and predicative.

* On leave from The RAND Corporation, 1962-63. The work reported in this paper was accomplished in part at RAND and completed at EURATOM. A fuller account of the connectability-test routine for Russian dominative functions is to appear as a EURATOM report.

The nature of a program for testing connectability with respect to coordinative functions (coordination, apposition, etc.) is suggested.

Punctuation and Automatic Syntactic Analysis*

Lydia Hirschberg

University of Brussels

In this paper we discuss how algorithms for automatic analysis can take advantage of information carried by the punctuation marks.

We neglect stylistic aspects of punctuation because they lack universality of usage and we restrict ourselves to those rules which any punctuation must observe in order to be intelligible. This involves a concept we call "coherence" of punctuation. In order to define "coherence", we introduce two characteristics, which we prove to be mutually independent, namely "separating power" and "syntactic function".

The *separating power* is defined by three experimental laws expressing the fact that two punctuation marks of different separating power prevent to a different extent syntactic links from crossing them. These laws are defined independently of any particular grammatical character of the punctuation marks or of the attached grammatical syntagms.

On the other hand, whichever grammatical system we choose, we may assimilate the punctuation marks to the ordinary words, to the extent that we can assign to them a known *grammatical character and function,* well defined in any particular context. They differ however from the other words by their large number of homographs and synonyms i.e. by the fact that almost every punctuation mark can occur with almost every grammatical value in each particular case, and in quite similar contexts.

The syntactic functions, in general, and in particular those of the punctuation marks, *can be ordered* according to an arbitrary scale of decreasing "value" of syntactic links, where the "value" of a link is directly related to the number of syntactic conditions the links must satisfy.

The law of coherence, then, shows that in a given context, a particular punctuation mark cannot indistinctly represent all its homographs, so that a certain number of assumptions about its syntactic nature and function can be discarded. This law can be stated as follows: "When moving from a punctuation mark to its immediate (left or right) neighbor in any text, the separating power cannot increase if the value of the syntactic function increases and vice-versa".

In addition we review two related topics, namely the stylistic character of punctuation and the necessity and existence of intrinsic criteria of grammatically, i.e. independent of punctuation. We propose such a criterion, and suggest a formalism related to the parenthesis free notation of logic.

* This investigation was performed under EURATOM contract No. 018-61-5-CET.B.

Application of Decision Tables to Syntactic Analysis

Walter Hoffman, Amelia Janiotis, and Sidney Simon

Wayne State University

Decision tables have recently become an object of investigation as a possible means of improving problem formulation of data processing procedures. The initial emphasis for this new tool came from systems analysts who were primarily concerned with business data processing problems. The purpose of this paper is to investigate the suitability of decision tables as a means of expressing syntactic relations as an alternative to customary flow charting techniques. The history of decision tables will be briefly reviewed and several kinds of decision tables will be defined.

As an example, parts of the predicative blocking routine developed at Wayne State University will be presented as formulated with the aid of decision tables. The aim of the predicative blocking routine is to group a predicative form together with its modal and temporal auxiliaries, infinitive complements, and negative particle, if any of these exist. The object of the search is to define such a syntactic block, but it may turn out instead that an infinitive phrase is defined or that a possible predicative form turns out to be an adverb.

Simultaneous Computation of Lexical and Extralinguistic Information Measures in Dialogue

Joseph Jaffe, M.D.

College of Physicians and Surgeons, Columbia University

An approach to the study of information processing in verbal interaction is described. It compares patterns of two indices of dispersion in recorded dialogue. The lexical measure is the mean segmental type—token ratio, based on 25-word segments of the running conversation. It is computed from a key punched transcript of the dialogue without regard to the speaker of the words. The extralinguistic measure is the H statistic, computed from the temporal pattern of the interaction. The latter is prepared from a two-channel tape recording by a special analogue to digital converter (AVTA system) which key punches the state of the vocal transaction 200 times per minute. Probabilities of the four possible states (either A or B speaking, neither speaking, both speaking) are the basis for the computation. All analyses are done on the IBM 7090. The methodology is part of an investigation of information processing in dyadic systems, aimed toward the reclassification of pathological communication.

Design of a Generalized Information System

Ronald W. Jonas

Linguistics Research Center, The University of Texas

While mechanical translation research involves the design of a computer system which simulates language processes, there is the associated problem of collecting the language data which are to be used in translation. Because large quantities of information will be needed, the computer may be useful for data accumulation and verification.

A generalized information system should be able to accept the many types of data which a linguist encodes. A suitable means of communication between the linguist and the system has to be established. This may be achieved with a central input, called Linguistic Requests, and a central output, called Information Displays. The requests should be coordinated so that all possible inputs to the system are compatible, and the displays should be composed by the system such that they are clearly understandable.

An information system should be interpretive of the linguist's needs by allowing him to program the data manipulation. The key to such a scheme is that the linguist be permitted to classify his data freely and to retrieve it as he chooses. He should have at his disposal selecting, sorting, and displaying functions with which he can verify data, select data for introduction to a mechanical translation system, and perform other activities necessary in his research.

Such an information system has been designed at the Linguistics Research Center of The University of Texas.

Some Experiments Performed with an Automatic Paraphraser

Sheldon Klein

System Development Corporation

The automatic paraphrasing system used in the experiments described herein consisted of a phrase structure, grammatically correct nonsense generator coupled with a monitoring system that required the dependency relations of the sentence in production to be in harmony with those of a source text. The output sentences also appeared to be logically consistent with the content of that source. Dependency was treated as a binary relation, transitive except across most verbs and prepositions.

Five experiments in paraphrasing were performed with this basic system. The first attempted to paraphrase without the operation of the dependency monitoring system, yielding grammatically correct nonsense. The second experiment included the operation of the monitoring system and yielded logically consistent paraphrases of the source text. The third and fourth experiments demanded that the monitoring system permit the production of only those sentences whose dependency relations were non-existent in the source text. While these latter outputs were seemingly nonsensical, they bore a special logical relationship to the source. The fifth experiment demanded that the monitoring system permit the production of sentences whose dependency relations were the converse of those in the source. This restriction was equivalent to turning the dependency tree of the source text upside down. The output of this experiment consisted only of kernel type sentences which, if read *backwards,* were logically consistent with the source.

The results of these experiments determine some formal properties of dependency and engender some comments about the role of dependency in phrase structure and transformational models of language.

Interlingual Correspondence at the Syntactic Level*

Edward S. Klima

Department of Modern Languages and Research Laboratory of Electronics, M.l.T.

The paper will investigate a few major construction types in several related European languages: relative clauses, attributive phrases, and certain instances of coordinate conjunction involving these constructions. In each of the languages independently, the constructions will be described as resulting from syntactic mechanisms further analyzable into chains of partially ordered operations on more basic structures. Pairs of sentences equivalent in two languages will be examined. Sentences will be considered equivalent if they are acceptable translations of one another. The examples used will, in fact, be drawn primarily from standard translations of scholarly and literary prose. Equivalence between whole sentences can be further analyzed, as will be shown, into general equivalence 1) between the chains of operations describing the constructions and 2) between certain elements (e.g., lexical items) in the more basic underlying structures. It will be seen that superficial differences in the ultimate shape of certain translation pairs can be accounted for as the result of minor differences in the particular operations involved or in the basic underlying structure. We shall examine two languages (e.g., French and German) in which attributive phrase formation and relative clause formation on the whole correspond and in which, in a more or less abstract way, the rules of relative clause formation are included as intermediate links in the chain of operations describing attributive phrases. The fact that in particular cases a relative clause in the one language corresponds to an attributive phrase in the other will be found to result from, e.g., differences in the choice of perfect auxiliary in the two languages.

* This work was supported in part by the National Science Foundation, and in part by the U.S. Army Signal Corps, the Air Force Office of Scientific Research, and the Office of Naval Research.

Sentence Structure Diagrams

Susumu Kuno

Computation Laboratory, Harvard University

A system for automatically producing a sentence structure diagram for each analysis of a given sentence has been added to the program of the multiple-path syntactic analyzer. A structure code, consisting of a series of structure symbols or phrase markers that identify the successive higher-order structures to which the word in question belongs, is assigned to each word of the sentence. The set of structure codes for the words of a given sentence is equivalent to an explicit tree diagram of the sentence structure, but more compact and easier to lay out on conventional printers.

The diagramming system makes some experimental assumptions about the dependencies of certain structures upon higher-level structures. All the major syntactic components of a sentence (i.e., subject, verb, object, complement, period, or question mark) are represented in the current system as occurring on the same level, all being dependent on the topmost level, "sentence". A floating structure such as a prepositional phrase or adverbial phrase or clause, whose dependency is not determined in the analyzer, is represented as depending upon the nearest preceding structure modifiable by such a floating structure. Different assumptions as to structural dependencies would yield different diagrams without requiring modification on the main flow of the diagramming program.

The diagrams thus obtained contribute greatly to the rapid and accurate evaluation of the analysis results, and they are also useful for obtaining basic syntactic patterns of analyzed structures, and for detecting the head of each identified structure.

Linguistic Structure and Machine Translation

Sydney M. Lamb

University of California, Berkeley

If one understands the nature of linguistic structure, one will know what design features an adequate machine translation system must have. To put it the other way around, it is futile to attempt the construction of a machine translation system without a knowledge of what the structure of language is like. This principle means that if someone wants to construct a machine translation system, the most important thing he must do is to understand the structure of language.

Any MT system, whether by conscious intention on the part of its creators or not, is based upon some view of the nature of linguistic structure. By making explicit the underlying theory for various MT systems which have been proposed we can determine whether or not they are adequate. Similarly, by observing linguistic phenomena we can determine what properties an adequate theory of language must have, and such determination will show what features an MT system must have in order to be adequate.

It can be shown that some of the approaches to MT now being pursued must necessarily fail because their underlying linguistic theories are inadequate to account for various well-known linguistic phenomena.

On Redundancy in Artificial Languages

W. P. Lehmann

Linguistics Research Center, The University of Texas

Artificial languages are one concern of work in computational linguistics, if only as a mnemonic device for interlinguas which will be developed. Even if it does not gain wider use, the structure of an artificial language is of general interest.

In contrast to the artificial languages which have been widely proposed, linguistic principles underlying a well-designed artificial language and its usefulness are well-established, particularly through Trubetzkoy's article, TCLP 8.5-21. which indicates phonological limitations for such a language. Since Trubetzkoy's specifications yield a total of approximately 11,000 morphemes, if an artificial language incorporated the degree of redundancy found in natural languages it would be severely handicapped by the size of its lexicon. The paper discusses the problem particularly with regard to suprasegmentals, which Trubetzkoy almost entirely ignored.

A Procedure for Automatic Sentence Structure Analysis

D. Lieberman

IBM Thomas }. Watson Research Center

The two main considerations in the design of this procedure were the economical recognition and representation of multiple readings of syntactically ambiguous sentences, and general applicability to "all" languages (English, Russian, Chinese). The following features will be discussed: types of structural descriptions, form of linguistic rules, use of linguistic heuristics to achieve economical multiple analyses, application to linguistic research and application to production MT systems. Also, the relation between this procedure and other existing sentence analysis procedures will be discussed.

An Algorithm for the Translation of Russian Inorganic-Chemistry Terms

L. R. Micklesen and P. H. Smith, Jr.

IBM Thomas J. Watson Research Center

An algorithm has been devised, and a computer program written, to translate certain recurring types of inorganic-chemistry terms from Russian to English. The terms arc all noun-phrases, and several different types of such phrases have been included in the program. Examples are:

AZOTNONATRIEVA4 SOL6 sodium nitrate
SOL6 ZAKISI/OKISI JELEZA ferrous/ferric salt
ZAKISNA4 OKISNA4 SOL6 JELEZA
GIDRAT ZAKISI/OKISI JELEZA ferrous/ferric salt

etc., where the stems underlined may be replaced by any of a number of other stems (up to 65 in some positions) in the particular type.

Translation of each type encounters problems common to almost all the types: (1) The Russian noun is translated as an English adjective, while the noun of the resulting English phrase is found among the modifiers of the Russian noun. (2) The Russian noun (English adjective) may be a metal with more than one valence state, the state indicated (if at all) by the modifiers. (3) The number of the resulting English noun-phrase is determined by some member of the Russian phrase other than the noun. (4) The phrase elements may occur compounded in the chemical phrase but free in other contexts, and dictionary storage must provide for this. The program permits translation of conjoined phrase elements as well.

The paper also includes an investigation into the deeper grammatical implications of this type of chemical nomenclature, and some excursions into the semantic correlations involved.

The Application of Table Processing Concepts to the Sakai Translation Technique

A. Opler, R. Silverstone, Y. Saleh, M. Hildebran, and I. Slutzky

*Computer Usage Company**

In 1961, I. Sakai described a new technique for the mechanical translation of languages. The method utilizes large tables which contain the syntactic rules of the source and target languages.

As part of a study of the AN/GSQ-16 Lexical Processing Machine, a modification of the Sakai method was developed. Five of six planned table scanning phases were implemented and tested. Our translation system (1) converts input text to syntactic and semantic codes with a dictionary scan, (2) clears syntactic ambiguities where resolution by adjacent words is effective, (3) resolves residual syntactic ambiguities by determining the longest meaningful semantic unit, (4) reorders word sequence according to the rules of the target language and (5) produces the final target language translation.

French to English was the source-target pair selected for the study. An Input Dictionary of 3,000 French stems was prepared and 17,000 entries comprised the Input Product Table (allowable syntactic combinations).

Since Sakai was working with highly dissimilar languages, he found it necessary to use an intermediate language. Because of the structural similarity between French and English, we found an intermediate language was unnecessary.

The method proved straightforward to implement using the table lookup logic of the Lexical Processor. The translation was actually performed on an IBM 1401 which we programmed to simulate the concept of the AN/GSQ-16 Lexical Processor. In our implementation magnetic tapes replaced the photoscopic storage disk.

* This work was performed while under contract to IBM Thomas J. Watson Research Center, Yorktown Heights, New York.

Slavic Languages—Comparative Morphosyntactic Research

Milos Pacak

Machine Translation Research Project, Georgetown University

An appropriate goal for present-day linguistics is the development of a general theory of relations between languages. One necessary requirement in the development of such a theory is the identification and classification of inflected forms in terms of their morphosyntactic properties in a set of presumably related languages.

According to Sapir, "all languages differ from one another, but certain ones differ far more than others". As for the Slavic languages he might well have said that they are all alike, but some are more alike than others. The similarities stemming from their common origin and from subsequent parallel development enable us to group them into a number of more or less homogeneous types.

The experimental comparative research at The Georgetown University was focused on a group of four Slavic languages, namely, Russian, Czech, Polish and Serbocroatian.

The first step in the comparative procedure here described is the morphosyntactic analysis of each of the four languages individually. The analysis should be based on the complementary distribution of inflectional morphemes. The properties whose distribution must be determined are:

1) the graphemic shape of the inflectional morphemes,
2) the establishment of distributional classes and subclasses of stem morphemes and (on the basis of 1 and 2),
3) the morphosyntactic function of inflectional morphemes which is determined by the distributional subclass of the stem morpheme.

f(x,y)-l, where x is the distributional subclass of the stem morpheme (which is a constant) and y is the given inflectional morpheme (which is a free variable). On the basis of this preliminary analysis the patterns of absolute equivalence, partial equivalence, and absolute difference can be established for each class of inflected forms in each language under study.

Once this has been accomplished, the results can be used in order to determine the extent of distributional equivalences among the individual languages. The applicability of this procedure was tested on the class of adjectivals. Within the frame of adjectivals the follow-

ing morphosyntactic properties were analyzed within each language first and compared among the four languages:

1) the category of gender,
2) the category of animateness,
3) the category of case and number.

The product of this comparative analysis is a set of formation rules which embody a system for the identification of the inflected forms. The detailed result will be presented in an additional report.

Types of Language Hierarchy

E. D. Pendergraft

Linguistics Research Center, The University of Texas

Various relations lead to hierarchical systems of linguistic description. This paper considers briefly a typology of descriptive metalanguages based on such relations and sketches possible consequences for computational linguistics.

Its scope is accordingly limited to metalanguages having operational interpretations which specify individual linguistic processes and structural interpretations which specify language data of individual languages. Immediate-constituent, context-free metalanguages are used to illustrate hierarchical types.

Path Economization in Exhaustive Left-to-right Syntactic Analysis

Warren J. Plath

Computation Laboratory, Harvard University

In exhaustive left-to-right syntactic analysis using the predictive approach, each path of syntactic connection which originates at the beginning of a sentence must be followed until it is clear whether or not it will lead to the production of a well-formed analysis. The original scheme of following each path until it terminates either in an analysis or in a grammatical inconsistency has been considerably improved through the. incorporation of two path-testing techniques. Using the first technique, the program abandons a path as unproductive whenever a situation is detected where the prediction pool contains more predictions of a given type than can possibly be fulfilled by the remaining words in the sentence. Employment of the second technique, which is based on periodic comparison of the current prediction pool with pools formed on earlier productive paths, eliminates repeated analysis of identical right-hand segments which belong to distinct paths.

Taken together, the two path-testing procedures frequently enable the program to terminate the processing of a path well before its end has been reached. For most sentences, this means a considerable reduction in the total path length traversed, accompanied by a corresponding increase in the speed of analysis. Comparison of runs performed using both versions of the program indicates that employment of the new techniques reduces the average running time per sentence to less than one-fifth of its former value.

A Computer Representation for Semantic Information

Bertram Raphael

Computation Center, Massachusetts Institute of Technology

This paper deals with the problem of representing in a useful form, within a digital computer, the information content of statements in natural language. The model proposed consists of words and list-structure associations between words. Statements in simple English are thought of as describing relations between objects in the real world. Sentences are analyzed by matching them against members of a list of formats, each of which determines a unique relation. These relations are stored on description-lists associated with those words which denote objects (or sets of objects). A LISP computer program uses this model in the context of a simple question-answering system. Functions are provided which may grow, search, and modify this model. Formats and functions dealing with set-relations, part-whole and numeric relations, and left-to-right spatial relations have been included in the system, which is being expanded to handle other types of relations. All functions which operate on the model report information concerning their actions to the programmer, so that the applicability and limitations of this kind of model may more easily be evaluated.

Specifications for Generative Grammars Used in Language Data Processing

Robert Tabory

IBM Thomas]. Watson Research Center

It becomes more and more evident that successful pragmatics (i.e. automatic recognition and production procedures for sentences) cannot be performed without previously written generative grammars for the languages involved, using an underlying meta-theoretical framework proposed by the present school of mathematical linguistics. Two aspects of grammar writing are examined:

1. A taxonomy over the non-terminal vocabulary, using a subscripting system for signs and fitting into the more general string taxonomy of phrase structure components. The resulting more complex lexical organization is studied.

2. A command syntax for phrase structure components limiting the full, not necessarily needed generative power of these grammars. The proposed restrictions correspond to a priori linguistic intuition. Applicational order and location of the rules is studied.

Finally, the recognitional power and generative capacity of a computer are examined, the machine being structured according to a Newell-Shaw-Simon list system. It is well known that pushdown stores are particular cases of list structures, that context-free grammars

are particular cases of phrase structure grammars and that pushdown stores are the generative devices for context-free grammars.

Collecting Linguistic Data for the Grammar of a Language

Wayne Tosh

Linguistics Research Center, The University of Texas

Establishing the grammatical description of a language is one of the major tasks facing the technician in machine translation. Another is that of creating the system of programs with which to carry out the translation process. The Linguistics Research Center of The University of Texas recognizes the advantages in maintaining the specialties of linguistic research and computer programming as two separate areas of endeavor.

We regard the linguistic task as a problem in convergence. We do not expect ever to have a final description of a language (except theoretically for a given point in the history of that language). We do expect, however, to begin with almost immediate application of the very first grammatical description. We shall make repeated revisions of the grammar as we learn how to make it approximate better the language text fed into the computer.

The grammatical description of any one language is based primarily on specific text evidence. We are not attempting to describe "the language". We are, however, attempting to make descriptive decisions sufficiently general that new text evidence does not require extensive revision of earlier descriptions.

Corpora selected for description are chosen so as to have similar texts within the same scientific discipline for the several languages. Tree diagrams are drawn for each sentence in detail. The diagrams are inspected for consistency before corresponding phrase-structure rules are compiled in the computer. The grammar is then verified in the computer system and revised as necessary.

Derivational Suffixes in Russian General Vocabulary and in Chemical Nomenclature

John H. Wahlgren

University of California, Berkeley

A grammar based upon a conventional morphemic analysis of Russian will have a rather large inventory of derivational suffixes. A relatively small number of these recur with sufficient generality to acquire lexemic status (i.e., to be what is usually termed "productive"). Names of chemical substances in Russian may likewise be analyzed as combinations of roots or stems with derivational affixes, in particular, suffixes. The number of productive suffixes in the chemical nomenclature is considerably larger than in the general vocabulary. These suffixes derive from adoption into Russian of an international system of chemical nomenclature. A grammar of this system is basically independent of any grammar of Russian. It must, however, be consistently incorporated into the grammar and dictionary which are to serve in a machine translation system for texts in the source language containing chemical names.

Grammatical analysis of chemical suffixes and connected study of general Russian derivational suffixes has raised certain practical problems and theoretical questions concerning the nature of derivation. On the practical side, where a complex and highly productive system is involved, effective means of detecting and dealing with homography have required development. Theoretical consideration has been given to the question of grammaticality in chemical names and to problems of sememic analysis and classification of root and stem lexemes into tactic classes on the basis of co-occurrence with derivational suffixes.

On the Order of Clauses*

Victor H. Yngve

Department of Electrical Engineering and Research Laboratory of Electronics, Massachusetts Institute of Technology

We used to think that the output of a translation machine would be stylistically inelegant, but this would be tolerable if only the message got across. We now find that getting the message across accurately is difficult, but we may be able to have stylistic elegance in the output since much of style reflects depth phenomena and thus is systematic.

As an example, the order of the clauses in many two-clause sentences can be reversed without a change of meaning, but the same is not normally true of sentences with more than two clauses. The meaning usually changes when the clause order is changed. Equivalently, there appear to be severe restrictions on clause order for any given meaning. These restrictions appear to follow from depth considerations.

The idea is being investigated that there is a normal depth-related clause order and any deviations from this order must be signalled by special syntactic or semantic devices. The nature of these devices is being explored.

When translating multi-clause sentences, there may be trouble due to the fact that the clause types of the two languages are not exactly parallel. Therefore the list of allowed and preferred clause orders in the two languages will not be equivalent and the special syntactic and semantic devices available to signal deviations from the normal order will be different. Thus one would predict that multi-clause sentences in language A often have to be split into two or more sentences when translated into language B, while at the same time multi-clause sentences in language B will often have to be broken into two or more sentences when translating into language A.

* This work was supported in part by the National Science Foundation, in part by the U.S. Army Signal Corps, the Air Force Office of Scientific Research, and the Office of Naval Research, and in part by the National Bureau of Standards.