# AUTOMATIC LINGUISTIC ANALYSIS

Pamela L. Coker & Mark A. Underwood

December, 1981

TN-3

national center for bilingual research, 4665 lampson avenue, los alamitos, CA 90720, telephone (213) 598-0481

The material in this publication was prepared under Cooperative Agreement 00-CA-80-0001 from the National Institute of Education, Department of Education. Its contents do not necessarily reflect the views of the National Institute of Education or of any other agency of the United States of Education or of any other agency of the United States government, and no official endorsement should be inferred.

# Abstract

The National Center for Bilingual Research (NCBR) intends to develop a large corpus of the language of bilingual children. This report surveys the available computer programs which could potentially aid in the linguistic analysis of the NCBR corpus by automating a number of labor-intensive and time-consuming linguistic analyses.

Two criteria guided the search for applicable computer programs. The automation of linguistic analyses which form the basis of the child language research for monolinguals were preferred over those analyses which are not typically used in child language research. The computer programs must be easily implemented on the UCLA IBM 370/3033 computer.

Eight computer programs which met at least one of the criteria were evaluated in terms of their potential usefulness to NCBR. It was determined that the Computer Assisted Language Analysis System (CALAS) was the most promising in terms of capabilities and cost. A series of programs which could be used immediately were located at UCLA, however; these programs are limited to word frequency counts and concordance programs based on terminal strings.

## Table of Contents

# Automatic Linguistic Analysis

## I. Introduction

The analysis of linguistic data has proven to be a time-consuming labor-intensive effort. The purpose of this report is to examine a series of computer assisted alternatives which reduce the amount of time and effort required for linguistic data analysis. In particular, a set of recommendations are made with respect to the needs of the National Center for Bilingual Research, which is presently collecting a large corpus of child language from bilingual children.

Computational Linguistics is a field that has been devoted to the automatization of linguistic information, whether it be for the machine translation of one language to another or for the analysis of textual and discourse information. Computational Linguistics became very active in the late 1950's with the advent of large computational machines. Since that time the field has developed in several directions, and has been supplemented by the newer field of Artificial Intelligence. This report will present a brief review of the goals and accomplishments of these two fields, followed by a discussion of desirable linguistic analyses for the NCBR corpus. Finally, a series of computer programs which could potentially aid in the automatization of the desirable linguistic analyses will be evaluated in terms of their ease of implementation by NCBR.

## II. Computational Linguistics and Artificial Intelligence

Research in computational linguistics generally falls into one of three areas: machine translation of one language to another, computer validation of linguistic theory, or computerized linguistic analysis of text or discourse. Machine translation of one language to another was an area of research heavily funded in the late 1950's and early 1960's. It was hoped that computer programs would be able to automatically translate written documents or even intercepted audio signals. It is generally agreed that these early efforts failed not because the computers did not have sufficient computational power but because we simply did not have an adequate understanding of the structure of the rules of natural languages (Chomsky, 1957, 1965). The most unsettling discovery of these early attempts was that given a dictionery of the words of a language and the syntactic rules of that language, the computer still could not generate the meaning of a sentence. What was missing was a set of rules which combines the meaning of individual words with the syntactic structure of a sentence to produce the meaning of that sentence. The delineation of these combination rules of semantic interpretation and the reassessment of the structure of syntactic rules have received considerable attention at the theoretical level during the last twenty years (Bresnan, 1976; Chomsky, 1965, 1976; Jackendoff, 1972; Katz & Fodor, 1964; Lakoff, 1971; Montague, 1974; Partee, 1975, 1976).

Today, although there are still efforts being made in machine translation of one language to another (see discussion below), a large part of the field of computational linguistics is devoted to the testing contemporary advances of linguistic theory. That is, a given formalism in linguistic theory is to be preferred if the correct meaning or the correct syntactic parse of a sentence can be assigned by computer. Simultaneous with this effort has been the emergence of the field of Artificial Intelligence which seeks to have the computer understand not only natural language but also solve complex problems. The goal of several projects has been the development of a computer program able to understand a sentence, to make an inference based on the meaning of that sentence, and then to use that inference as the partial solution to a given problem. Because so much of the effort in Artificial Intelligence involves the understanding of linguistic information, the Computational Linguist and the researcher in Artificial Intelligence have many shared goals.

A number of computer programs designed to parse the syntactic structure of a sentence have been written to test competing linguistic theories of syntactic structure. Mitchell Marcus, who is currently at Bell Laboratories, has written a deterministic syntactic parser which incorporates a number of constraints on linguistic rules proposed by Chomsky (Chomsky, 1973, 1976; Marcus, 1978). Ron Kaplan of Xerox Palo Alto Research Center is currently implementing a syntactic parser based on his previously developed Augmented Transition Network (ATN) parser and on Joan Bresnan's Realistic Grammar (Bresnan, 1978), which is a competing theory to Chomsky's. Martin Kay also of Xerox is currently implementing another parser based on Systemic Grammar. These parsers are similar because each was developed to test a theory, and, as such, none are comprehensive parsers of English. They consist only of a subset of the rules of English, and thus are not generally applicable to the task of analyzing a large corpus of naturalistic data.

In Artificial Intelligence, more ambitious researchers have produced computer programs which not only assign a syntactic structure to the sentence, but also interpret the meaning of a sentence. The interpreted meaning, along with other stored knowledge, is processed to yield inferences which aid in complex problem solving. For example, Winograd's SHRDLU conversed with a human in English about a small imaginary world of blocks. The conversation involved the computer responding to orders to move the blocks and keeping track of the relative positions of the blocks (Winograd, 1971,1972). SHRDLU both interpreted and produced English sentences. LUNAR, developed at Bolt, Beranek & Newman, Inc. is used by NASA to access and manipulate moon rock samples data. Again the conversation with LUNAR is in English. SOPHIE (Sophiticated Instructional Environment) is capable of conversing in English with a student about the student's ideas on electronic troubleshooting (Bobrow & Brown, 1975, Brown, Bell &

Burton, 1974, Brown, Burton & Bell, 1975). GUS (Genial Understanding Systme) communicated with travel agency clients who wished to travel to a single city on any of several air flights. These and other projects by Anderson and Bower (1973), Schank (1973, 1976, 1978, 1980), and Norman and Rumelhart (1975) are serious attempts to automate the understanding of linguistic information. They are, however, only attempts at what is possible. Typically, both the topics of conversation and the linguistic structures are restricted to those necessary for the tiny artificial domain of the system's "world". There has been no attempt to develop a comprehensive set of linguistic rules, and lexicons have been limited to include only a small, interrelated set of words; and, as such, the computer programs are not equipped to handle extensive semantic domains found in any spontaneous language corpus.

The third area of Computational Linguistics is the linguistic analysis of textual and discourse information. Computers aid in the analysis of literature and poetry. For example, choice of words by two or more authors can be compared by computer concordance programs which count the number of times a particular word or phrase appears and listing out the context of each instance of the word (Ross, 1972; Widman, 1975). In this way the choice and use of words of particular authors can be compared and analyzed. Concordance programs vary as to which linguistic features they can analyze, and have been used to count the number of occurrences of syntactic structures (Chrisholm, 1976) as well as to compute letter and word frequency, spelling patterns, and morphological complexity (Spolsky, Holm, Holliday & Embry, 1978). In addition to the linguistic analysis of literature, there are also programs which analyze scientific textual data. For example, the String Parser programs at New York University analyze medical texts and other scientific textual information (Fitzpatrick & Sager, 1974; Hobbs & Grishman, 1976; Sager, 1976). The input in each of these cases is well-formed grammatical sentences of English, and the syntactical rules in these programs assume grammatically correct input sentences.

Better suited to the linguistic analysis of the NCBR corpus are the programs which analyze discourse. Computer programs have been designed to analyze interactive dialogue sessions between two or more people (Miron, 1973). Dialogues between teachers and students, therapist and patient (Wachal & Spreen, 1970; Colby, Parkinson & Fought, 1974), as well as schizophrenic and other pathological language (Pepinsky, 1978) have been analyzed by computer programs. The advantage of these programs is that they can analyze sentence fragments, one word utterances and discourse-specific features not found in written language.

## III. Analysis of Child Language Corpora

The National Center for Bilingual Research intends to tape record the language of young bilingual children in a three year longitudinal study. The tapes will be transcribed and entered into the computer by clerical personnel. The accuracy of the transciptions will be verified by personnel with linguistic training. Because the resulting corpus will be quite large, it is desirable to automate as much of the linguistic analysis as possible. But before considering the actual programs which might be used to automate certain types of linguistic analyses, a discussion of the particular analyses relevant to child language production data is in order.

Since the transcripts will not contain a phonetic transciption of the child's speech, phonological analysis of the corpus is not possible. However, the syntactic, semantic and conceptual information in the corpus offers a rich base of data from which to analyze the complexity of the child's linguistic and conceptual development at particular ages. In order to evaluate the complexity of the bilingual child's language, it is desirable to use at least some of the measures of linguistic complexity developed for the analysis of monolingual language development.

One of the most widely used measures of linguistic complexity has been the mean length utterance (MLU) in the child's spontaneous speech. It is the best single indicator of complexity up to about five morphemes per utterance (Brown, 1973). It indicates both syntactic and semantic complexity which is highly correlated with conceptual complexity. It would be highly desirable to compute MLU for the NCBR corpus, as it would provide the basis for comparison with the extensive child language literature on monolinguals.

Slobin (1973) has developed a number of indices as to what contributes to syntactic complexity. These are based on the following language acquisition universals (taken from Slobin, 1979).

1) For any given semantic notion, grammatical realizations as postposed forms will be acquired earlier than realizations as preposed forms.

2) The following stages of linguistic marking are typically observed: (1) no marking, (2) appropriate marking, (3) overgeneralization of marking, (4) full adult system.

3) The closer a grammatical system adheres to one-to-one mapping between semantic elements and surface elements, the earlier it will be acquired.

4) When selection of an appropriate inflection among a group of inflections performing the same semantic function is determined by arbitrary formal criteria, the child initially tends to use a single form in all environments.

5) Semantically consistent grammatical rules are acquired

early and without significant errors.

Using Slobin's universals of language acquisition, it is possible to predict which syntactic structures will be difficult to learn in any language. For example, syntactic rules which are inconsistently applied or which attach themselves to the beginnings of words rather than to the ends of words are considered as complex relative to rules which are consistent with Slobin's universals.

Consider the tense system of English. Semantically, English expresses three tenses: past, present, and future. Syntactically, however there are only two tense markers: past and present. Each semantic expression can be syntactically marked as either past or present, as the following examples indicate. The examples are taken from Culicover (1976). All three are syntactically marked in the present tense, though each one semantically represents a different time.

1) I come home and then John says to me "Where the devil have you been all day?" (semantic past)

2) I choose Mary. (semantic present)

3) I sail for England next Wednesday. (semantic future)

This system becomes very complicated for the child when he (or she) learns the past tense marker and it does not always refer to some time in the past as in (4).

4) I would like a glass of milk. (semantic present, would is marked syntactically past)

These examples illustrate Slobin's third universal, that when there is not a one to one mapping between semantic elements and surface syntactic markings, the language learning task becomes more difficult.

In order to make specific comparisons with regard to the syntactic complexity of the child's speech, the level of analysis must be quite detailed. For example, Brown and others (Brown, 1973, Brown & Bellugi, 1964; Brown, Cazden & Bellugi, 1969) have traced the development of 14 grammatical morphemes in English. Sone of these are: present progressive (-ing) the prepositions oon and in, plural, possessive ('s), uncontracted copula (is), articles (the and a), irregular and regular past tense. To automate this type of syntactic analysis, the computer program must be able to detect individual morphemes when they appear as parts of words.

Other syntactic analyses which are important in determining the syntactic complexity of the child's language include analysis at the phrasal level. For example, the syntactic structure of (5) is generally regarded as more complex than that of (6). This is because (5) includes an embedded sentence in the subject noun phrase of the sentence whereas (6) does not have this additional structure at the surface level of analysis.

(5) The dog which belonged to Mary died.

(6) Mary's dog died.

Thus, it would be very useful to be able to analyze the child's utterances according to their phrasal complexity. This involves first determining what part of speech each word in the sentence is and then determining which syntactic rule applies to the sequence of syntactic categories. In order to perform this type of analysis on the computer, it is necessary to have a lexicon of the common words coded as to their syntactic category. However, this is sometimes difficult to implement since part of speech determination is often dependent on the placement of the word in the phrase or sentence. So, if a lexicon with associated syntactic categories is to be maintained, we must allow for the occurrence of more than one syntactic category for a particular word. This introduces ambiguity into the analysis, which must be be resolved at some later stage of analysis.

The child's mastery of coordinate and subordinate structures must also be analyzed by a program with phrasal/sentential level capabilities. This is somewhat easier to determine automatically, since the program can search for coordinating and subordinating conjunctions which introduce these clausal structures. Although there is ambiguity as to the syntactic category of these conjunctions, it fairly easy to resolve the ambiguity via the surrounding syntactic structure of the sentence, which can be readily expressed in simple phrase structure rules. Concordance programs could search for all the instances of the coordinating conjunctions, and, or, and then, but first, and the subordinating conjunctions, because, although, when, while, before, after, until, since. The "hits" of the search then could be categorized as to whether the conjunctions conjoined sentences or phrases.

The use of subordinating conjunctions not only indicates a syntactic sophistication but also the mastery of difficult semantic concepts. These in addition to logical connectors such as if...then, either...or, and suppose indicate advanced semantic development. The line between syntactic development and semantic development is also blurred when we consider the development of complex verbs, such as believe, understand, volunteer, realize, imagine, etc. which take sentential or infinitival complements.

In sum, there are a variety of linguistic analyses which measure the syntactic/semantic and conceptual complexity of child language. Many of these measures require detailed linguistic analysis. To perform these analyses automatically required a sophisticated computer program.

## IV. Criteria for Evaluating Automatic Linguistic Analysis Programs

Two overriding criteria served as the basis for the evaluation of computer programs for the automatic linguistic analysis of NCBR corpus. The first was to seek computer programs which automated as much of the linguistic analysis as possible. That is, programs which could analyze the phrase structure of a sentence were considered more desirable than simple concordance programs which compute frequencies at the terminal string level only. The second and more important consideration was the amount of effort and time required to implement the computer program on the IBM 370/3033 at UCLA. From these two general considerations, the following list of questions was generated.

1) Is the program designed to analyze spontaneous discourse or textual information? The problem here is that if the program is designed with the assumption that each sentence will be a grammatical sentence of English, then a considerable amount of effort must be spent in writing a new set of syntactic surface structure rules which will allow for sentence fragments and one word utterances typical of spontaneous discourse. Additionally, since the grammatical rules of child language differ from adult grammatical rules, provisions must be made in the program for the addition of the rules of child grammar.

2) What is the structure of the lexicon in the program, and how much effort is required to add new words to it? In particular, what attributes are associated with each word? (e.g. inflectional morphenes, syntactic categorization rules).

3) What is the output of the program? Does it count the number of occurrences of a particular structure? Does it keep track of where in the corpus the structure of

interest occurred? Is it possible to obtain a listing of the surrounding context of the structure in question? Is the type of output under user control?

4) How transportable is the program to the UCLA IBM 370/3033?

- * Is there a programmer who is currently assigned to maintain the code?

- * What is the current amount of usage of the program?

- * What machine does the program run on? Are there any machine-dependent utilities required for the implementation of the program?

- * What operating system does the program run under?

- * What programming language is the code written in?

5) Can the program be used via remote timesharing?

6) How much main memory does the program require?

7) How costly is it to use the program?

- * How long does the program take to analyze a 10 word sentence?

8) What is the relationship between the size of the lexicon and the amount of disk storage?

9) What documentation is available?

- * Are there user manuals?

- * Are there software maintenance manuals?

- Is there operations documentation?

## V. Surveyed Linguistic Analysis Computer Programs

As discussed in the introduction, the computer programs which purported to analyze textual and discourse information were deemed the most appropriate for the purposes of analyzing the NCBR corpus. This is because these programs attempted to be comprehensive in the development of their syntactic parsing rules and their lexical entries. Additionally, we discovered two machine translation programs which are very sophisticated despite a reduction in government funding for machine translation projects. We begin with the two machine translation programs, both of which are capable of translations between English and Spanish.

### V.A. Brigham Young University Project

The theoretical basis for this machine assisted translation project is Junction Grammar developed by Eldon Lytle (Lytle, Packard, Gibb, Melby & Billings, 1975). Junction Grammar representations consist of word-sense information interrelated by junctions which contribute syntactic and semantic information. In the first stage of the translation system, the program interacts with a human operator who aids the machine in resolving ambiguities, producing a representation of the meaning of the text. The second and third stages of the translation process are automatic transfer and synthesis into one or more target languages.

Currently, there are two versions of the Junction Grammar machine translation system. The first is still at Brigham Young University. It is a highly interactive system, which requires a linguist who is conversant in Junction Grammar to properly resolve the ambiguities which the machine presents to the human operator. It is capable of sophisticated linguistic parses, e.g. it can note the difference between restrictive and non-restrictive relative clauses; and can distinguish count versus mass nouns, generic versus specific senses, among others. Unfortunately, at the present time, the Brigham Young University project is under experimental revision, and the code is not transportable. When the code is intact, it runs on an IBM 370/130 and is written in PL1. Time-sharing is available.

The other version of the Junction Grammar project is a commercially available machine assisted translation program. This version was developed by Eldon Lytle and others and is available from APL Systems, 450 N. University, Provo, Utah, 84601. This version has eliminated the need for a trained linguist to resolve the ambiguities. The system is highly interactive and is capable of translating English text into Spanish, French and German. The lexicon is quite extensive with 5000 general purpose words, and specific lexicons in computer science, heavy equipment, and systems design. Dr. Lytle indicated that it is fairly easy to add more words to the lexicon and that it is suited to the analysis of dialogue as well as textual information. Also, it would not be difficult to add child language grammar to the other syntactic parsing rules. There are two drawbacks as far as using this system for the NCBR corpus. First, it runs on a Data General machine and is written in ALGOL. It would be an extensive project (as much as one man year) to convert the code to run on the UCLA IBM machine. ALP Systems expects to have their programs converted to run on other machines, though to date no specific plans have been for an IBM conversion. Secondly, because it is a commercial product, NCBR would have to purchase the program, which is fairly expensive due to the long development effort by the company.

## V.B University of Texas, Austin, Linguistics Research Center

The Linguistics Research Center has developed an English-

German translation program. It can take a sentence as input and generate the syntactic structure of the sentence. Currently it has a lexicon of 3,000 words, with specialized lexicons in telecommunications and electronic switching systems, and in computer systems. There are several drawbacks as to using this system for the NCBR corpus. First, a highly trained linguist would have to write the child language grammar to input into the system. Linguists trained in theoretical linguistics typically have not had the experience in writing the computationally unambiguous syntactic rules necessary for machine translation. Second, the funding of the Texas project is currently being taken over by private sources and thus all future versions of this project will either not be available or will be at commercial prices. Third, though the programs are highly portable because they are written in UCI LISP, a relatively machine-independent high level programming language, a conversion effort is still required to run under the IBM operating system. The present implementation at Texas is on a DEC 10 but the Texas system is currently being converted to INTERLISP which will run on the DEC 20. In sum, though the Texas project is well-developed, the change in their funding situation means that the currently available system will fall into disuse, with the task of software maintenance becoming the burden of NCBR.

The final report of the Texas translation project may be obtained after October 1, 1980 from Zbigniew L. Pankowicz, Foreign Technology Division, Rome Air Development Center, Griffiss AFB, NY 13441.

### V.C. Syracuse University

In the late 1960's and early 1970's Professor Murray Miron directed a number of projects which consisted of computer programs to perform frequency analysis of vocabulary and sentence patterns in Japanese, Swahili and English (Miron, 1973; Rubama, Miron & Pratt, 1973; Sukle, Miron & Pratt, 1973). While those programs are capable of relevant linguistic analyses, the programs have not be used in the last five years and thus it is extremely unlikely that are transportable to UCLA. Professor Miron currently has linguistic analyzer called General Inquirer II which was developed for use in analyzing dialogue. Professor Miron said that General Inquirer II would be ideal for the analysis of the NCBR corpus. That is, it is possible to add more syntactic rules to the parser and more words to the lexicon. Also it is capable of generating the types of output of interest to NCBR, e.g. frequency counts of parts of speech, phrasal and sentential structure, etc. General Inquirer II is currently being used to aid the FBI in analyzing threats. Professor Miron uses it to develop personality profiles. Professor Miron was very interested in developing a collaborative effort with NCBR with respect to the use and maintenance of General Inquirer II. As with many computer programs which are developed with Government funding on a project basis, not enough resources are allowed for documentation and software maintenance. Professor Miron estimated that if NCBR wanted to use the program at Syracuse University, it would take one man year of programming effort to make the modifications for child language analysis. Furthermore, to transfer the program to UCLA would be next to impossible as the code is a potpouri of different programming languages, with no overall design. There is no documentation. Finally, to run the program it takes a large amount of random access memory (RAM) which is expensive.

## V.D. New York University, Linguistic String Parser Project

The Linguistic String Parser developed at NYU is designed for the analysis of scientific texts (Fitzpatrick & Sager, 1974). The parser takes well-formed complete sentences of English and outputs a parse tree for the sentence. Although it would accept a noun phrase without a verb or an object, in general it is unacceptable for discourse data. Another drawback is that it is a non-interactive system, and at the present time there are no provisions for outputs other than parse trees. The Linguistic

String Parser has a large set of syntactical rules as well as an extensive lexicon. The lexicon stores a variety of attributes of the word, including morphological variants, grammatical categories, selectional restrictions, and subcategorization rules. Currently the program is running on a CDC 6600 and uses a large amount of RAM memory (600KB). Though it is written in FORTRAN, it would still need to be converted to the IBM operating system. It is also extremely costly; a ten word sentence takes 1 second of CPU time to parse.

### V.E. IBM Projects

Currently there are two projects of interest at the IBM Thomas J. Watson Research Center. The first one is called TQA for transformational question and answering program. It is designed to be the natural language interface to a data base management system (DBMS). Thus it understands and produces English discourse. Presently, it is being used as an interface to a municipal data base on land use assessments. Though it is capable of extensive syntactic and semantic analysis, this program is proprietary to IBM is thus not available for dissemination.

The second project at IBM is syntactic parser based on Controlled Partition Grammar (Muckstein, 1979). This parser takes the output from a speech recognition system, operating bottom-up to generate a written version of the text. The syntactic parser is constructed to recognize and define surface syntactic dependencies based on the parts of speech which have been generated by a part-of-speech label algorithm. This parser has been used to analyze the text of depositions of patent attorneys. The sentences average 35 words in length and tend to be well-formed grammatically. Dr. Muckstein indicated that it would take a considerable amount of effort to adapt the program to a child language corpus. Furthermore, since the research was supported by IBM and not by Government funds, the computer programs are most likely proprietary to IBM and hence not available.

## V.F. SRI International, DIAGRAM

SRI has developed a natural language understanding system called DIAGRAM, which produces parse trees as its output. These parse trees are then semantically interpreted and produce the logical meaning of the sentence. The logical meaning can then be queried by other computer systems. DIAGRAM currently has a lexicon of 3,000 words in English and Spanish. The structure of lexical entries is detailed and complex. The verbs alone are categorized by some 20 attributes, such as whether they are transitive, intransitive, or detransitive; whether they take particles, etc. In terms of modifying the syntactic rules and lexicon to accommodate child language grammar, a highly trained linguist would need to spend some time with the project linguist, Dr. Jane Robinson, in order to learn the system of grammatical rules implemented by DIAGRAM. The development of the lexical entries is the most difficult task. Mr. Gary Hendricks of the project estimated that if SRI were to add 500 new lexical items for NCBR and also gave NCBR a two week training session, the cost would be approximately $50,000. If NCBR were to do all of the linguistic work, then it would cost approximately $10,000 for training. Because DIAGRAM was developed under Government funding, the code is available at no charge.

To install DIAGRAM on the UCLA computer, it would require the conversion of the code, written in INTERLISP, to the IBM operating system. At SRI, DIAGRAM runs on a DEC 10 and a Foonley which emulates the DEC 10. The operating systems it runs under are 10X and TOPS 20. Mr. Hendricks indicated that SRI would make timesharing available on their DEC 10 at the end of the year, and that timesharing costs for Government programs are inexpensive. In terms of the documentation available for DIAGRAM, there are two 20 page manuals for programmers and no user manuals. There are five users at SRI.

DIAGRAM has received praise from the Stanford research community and so it deserves careful consideration. Mr. Hendricks of SRI suggests NCBR send some sample data to SRI and have them run it through DIAGRAM to see if the resultant parse would be useful for NCBR's purposes. In terms of CPU time, a full parse with semantic interpretation takes approximately one second for a ten word sentence and a syntactic parse without a semantic interpretation takes about 250 msec. Technical reports on DIAGRAM are available from Dr. Jane Robinson of SRI. She can be reached at (415) 326-6200, extension 4573.

## V.G. Computer Assisted Language Analysis System (CALAS)

CALAS was developed to analyze discourse and dialogue information. It has been used to analyze interactions between students and teachers in a classroom setting and between therapist and patient in a clinical setting. CALAS consists of three stages. Stage 1, called EYEBALL, assigns the part of speech to each word in the sentence. Ambiguities of parts of speech are resolved by a human editor. Stage 2, PHRASER, assigns aggregates of words to phrase structures. Again a human editor eliminates possible ambiguities. Finally in Stage 3, CLAUSE/CASE assigns semantic roles according to Case Grammar. All human editing can be done either interactively or off-line.

Because CALAS relies on human editing, the computer programs are not as complex and costly to run as some of the other programs we have discussed ( DIAGRAM, Linguistic String Parser, and General Inquirer II ). The human editor need not be a linguist; a good working knowledge of freshman English is adequate. The editing process is the most important at Stage 1, as EYEBALL has an 85% accuracy rate in assigning syntactic categories to the words. If the errors are caught in this stage, the remaining editing proceeds smoothly. Errors that escape the editor in the first Stage 1 can play havoc with the next two stages.

CALAS is a flexible program and can be easily modified to analyze child language data. The program was designed for the analysis of discourse, and it is a simple matter to add new lexical items to the dictionary as well as change the syntactic/semantic rules. For example, the user is asked each time he or she logs onto the system whether lexical items are to be added or deleted and whether the syntactic/semantic rules are to be changed. This feature means that different child grammars can be tested for different aged children (or different languages) in the corpus. This feature seems ideally suited to NCBR's needs.

Another attractive feature of the CALAS program is that print routines are designed to feed into SPSS programs. For example, frequencies could be computed for: number of words per noun phrase, number of complex noun phrases, number of plural markers, number of adjectives, nouns, etc., number of words per utterance. While this last item is not mean length utterance as used in the child language literature, most, if not all, of the information used to calculate typical MLU counts can be taken from the CALAS program.

In terms of transportability, CALAS will run on any IBM 370 series including the IBM 370/3033 at UCLA. Dr. Naomi Meara at University of Tennessee recently installed CALAS on an IBM 370/3031 with little difficulty. Most of the programs are written in PL1 and one program is written in SPITBOL, which is version of SNOBOL. In order to run CALAS, it is necessary to interface through another time sharing machine. The DEC PDP11/34 should be sufficient for this purpose. Dr. Meara and Dr. Pepinsky at Ohio State University are currently writing a user's manual. There are programmers at each institution who have served as consultants on CALAS and would be willing to assist by phone or letter in the installation of CALAS at UCLA. Both Dr. Pepinsky and Dr. Meara thought the installation would proceed smoothly.

CALAS can by obtained from Dr. Pepinsky at Ohio State University simply by mailing him a tape or sending him $35 for a tape with the program on it. Dr. Pepinsky can be reached at (614) 422-5470.

## V.H. UCLA Word Frequency and Concordance Programs

If NCBR would like to begin some simple linguistic analyses immediately, we located a set of programs which are available now and could be used with little programming resources on the part of NCBR. The advantage of these programs is that they already run on the UCLA IBM 370/3033 computer, and they are used frequently enough to expect that they are well-maintained. The disadvantage is that they only perform word frequency counts and concordances of terminal strings specified by the user. But because they are relatively simple programs as compared to most of those reviewed, they also are inexpensive to run. The amount of RAM memory needed is dependent on the size of the corpus to be analyzed. The size of the NCBR corpus could be reduced by categorizing the corpus into meaningful subcategories, such as analyses by individual child, by a calendar period, by age of the children, by language, etc.

In addition to word frequency counts, the concordance programs can list the sentence in which each word of interest appears, as well as list the word in the middle of a page, along with the preceding and succeeding 60 characters on either side of the word. In this way, the context in which the word or phrase appears will be listed out for further analyses. These programs have a number of other useful features, and we suggest that NCBR contact Dr. Rand in the ESL Department for further information. Dr. Rand has worked with SWRL programmers in the past on the LAP project and understands NCBR's needs in terms of this project.

Dr. Rand can be reached at (213) 825-4647 and has office hours daily from 1:00 pm to 2:00 pm. Dr. Rand suggested that NCBR take him a sample of data punched on cards, to run it through the word frequency and concordance programs. In this way, NCBR will be able to quickly determine if the programs are suitable. Additionally, Dr. Rand may know of other programs available at UCLA once he has a clear picture of the linguistic analysis requirements of the NCBR corpus.

## VI. Conclusions and Recommendations

Eight computer programs which met at least one of the general criteria listed in Section IV were discussed in detail tc determine whether or not they could be used to analyze the NCBR child language corpus. The first criterion was to locate programs which could automate as much of the linguistic analysis as possible and the second criterion was the amount of effort and cost of implementing the computer program on the UCLA IBM 370/3033.

Six projects met the first criterion, but were unsatisfactory in terms of the second criterion. These were: the two machine translation projects based on Junction Grammar in Provo, Utah; the machine translation project at the Linguistics Research Center at the University of Texas, Austin; General Inquirer II at Syracuse University; the Linguistic String Parser at New York University; the two projects at IBM Thomas J. Watson Research Center; and DIAGRAM at SRI International. Of these, DIAGRAM may be acceptable in terms of ease of implementation if a timesharing agreement between SRI International and NCBR could be negotiated. Problems still remain as to how adaptable DIAGRAM is to child language data.

In terms of satisfying both criteria, CALAS appears to be the optimal choice. It is relatively sophisticated in terms of the linguistic analyses that it can perform and it should be fairly straightforward to install CALAS on the UCLA IBM 370/3033. Additionally, a number of researchers have already used CALAS, so NCBR has the basis to adequately evaluate the program before deciding to use it. It is recommended that NCBR contact Dr. Pepinsky and Ohio State University and Dr. Meara at the University of Tennessee for a first hand assessment of the capabilities of CALAS.

And finally, the word frequency and concordance programs at UCLA best satisfy the second criterion but are deficient in terms of the complexity of the linguistic analysis they are able to perform. Since the use of these programs require very little programming or technical support by NCBR, it is recommended that NCBR explore the possible analyses offered by these programs with Dr. Rand at UCLA.

## References

Anderson, J. R. & Bower, G. H. *Human associative memory*. New York: Halstead Press, 1973.

Bobrow, D. G., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H., & Winograd, T. GUS, A frame driven dialogue system. *Artificial Intelligence*, 1977, *8*, 155-173.

Bobrow, R. J. & Brown, J. S. Systematic understanding: Synthesis, analysis, and contingent knowledge in specialized understanding systems. In D. G. Bobrow A. Collins (Eds.), *Representation and understanding studies in cognitive science*. New York: Academic Press, 1975.

Bresnan, J. A realistic transformational grammar. In M. Halle, J. Bresnan and G. Miller (Eds.), *Linguistic theory and psychological reality*. Cambridge, Mass.: The MIT Press, 1978.

Brown, J. S., Bell, A. G. & Burton, R. R. *Sophisticated instructional environment for teaching electronic troubleshooting*. (AFHRL-TR-74-77). Brooks Air Force Base, TX, Air Force Human Resources Laboratory. (NTIS No. AD A-002 148 5ST).

Brown, J. S. & Burton, R. R. Multiple representations of knowledge for tutorial reasoning. In D. G. Bobrow & A. Collins (Eds.), *Representation and understanding studies in cognitive science*. New York: Academic Press, 1975.

Brown, J. S., Burton, R. R. & Bell, A. G. SOPHIE: A step toward creating a reactive learning environment. *International Journal of Man-Machine Studies*, 1975, *7*(5), 675-696.

Brown, R. *A first language: The early stages*. Cambridge, Mass.: Harvard University Press, 1973.

Brown, R. & Bellugi, U. Three processes in the child's acquisition of syntax. *Harvard Educational Review*, 1964, *34*, 133-151.

Brown, R., Cazden, C. & Bellugi, U. The child's grammar from I to III. In J. P. Hill (Ed.) *Minnesota symposium on child psychology* (Vol. 2). Minneapolis: University of Minnesota Press, 1969.

Chomsky, N. *Syntactic structures*. The Hague, Netherlands: Mouton, 1957.

Chomsky, N. *The aspects of the theory of syntax*. Cambridge, Mass.: The MIT Press, 1965.

Chomsky, N. Conditions on transformations. In S. Anderson & P. Kiparsky (Eds.) Festschrift for Morris Halle. New York: Holt, Rinehart & Winston, 1973.

Chomsky, N. Conditions on rules of grammar. Linguistic Analysis, 1976, 2, 303-351.

Colby, K. M., Parkinson, R. C. & Fought, B. Pattern-matching rules for the recognition of natural language dialogue expressions. The Finite String, 1974, 11(1), (Published as part of The American Journal of Computational Linguistics, 1974, 1, Microfiche 5, 1-82).

Culicover, P.W. Syntax. New York: Academic Press, 1976.

Fitzpatrick, E. & Sager, N. The lexical subclasses of the linguistic string parser. The Finite String, 1974, 11(1). (Published as part of The American Journal of Computational Linguistics, 1974, 1, Microfiche, 2, 1-70).

Hobbs, J. R. & Grishman, R. The automatic transformational analysis of English sentences: An implementation. International Journal of Computer Mathematics, 1976, 5(4), 267-285.

Jackendoff, R. Semantic interpretation in generative grammar. Cambridge, Mass.: The MIT Press, 1972.

Kaplan, R. Augmented transition networks as psychological models of sentence comprehension. Artificial Intelligence, 1972, 3, 77-100.

Katz, J. J. & Fodor, J. A. The structure of semantic theory. In J. A. Fodor & J. J. Katz (Eds.), The structure of language: Readings in the philosophy of language. Englewood Cliff, New Jersey: Prentice-Hall, 1964.

Lakoff, G. On generative semantics. In D. Steinberg & L. A. Jakobovits (Eds.), Semantics: An interdisplinary reader in philosophy, linguistics, and psychology. New York: Academic Press, 1971.

Lytle, E. G., Packard, D., Gibb, D., Melby, A, K., Billings, F. H. Jr. Junction grammar as a base for natural language processing. American Journal of Computational Linguistics, 1975, 3,(77).

Marcus, M. P. A theory of syntactic recognition for natural language. Doctoral Dissertation, Massachusetts Institute of Technology, 1977.

Miron, M. Spoken language vocabulary and structural frequency count: English data analysis. (SURC-TR-73-117). Syracuse University Research Corporation, Syracuse, NY, March, 1973. (NTIS No. AD-775 924/4)

Montague, R. The proper treatment of quantification in ordinary English. In J. Hintika, J. Moravcsik & P. Suppes (Eds.) Approaches to natural language. Dodrect, Holland: D. Reidel, 1973.

Muckstein, E. M. A natural language parser with statistical applications. IBM Research Report, RC 7516 (#32506), 1979.

Norman, D. & Rumelhart, D. E. (Eds.) Explorations in cognition. San Francisco: W. H. Freeman & Company, 1975.

Partee, B. Montague grammar and transformational grammar. Linguistic Inquiry, 1975, 6, 203-300.

Partee, B. Montague grammar. New York: Academic Press, 1976.

Pepinsky, H. B. A computer-assisted language analysis system (CALAS) and its applications. Columbus, Ohio: Ohio State University, 1978. ERIC Document Reproduction Services No. ED 162 663.

Pester, A. R. The use of the computer in linguistic and literary research. Association for Literary and Linguistic Computing Bullentin, 1976, 4,(3), 245-250.

Plath, W. J. String transformations in the REQUEST system. The Finite String, 1974, 11(2). (Published as part of the The American Journal of Computational Linguistics, 1974,11, Microfiche, 8, 1-81.)

Ross, D. Beyond concordance: Algorithms for description of English clauses and phrases. In A. J. Aitken, R. W. Bailey, & N. Hamilton-Smith (Eds.), The computer and literary studies. Edinburgh: Edinburgh University Press, 1973.

Rubama, I, Miron, M., & Pratt, C. C. Spoken language vocabulary and structural frequency count: Swahili data analysis (SURC-TR-73-229). Syracuse University Research Corporation, Syracuse, N.Y., June, 1973. (NTIS No. AD-775 926/9)

Sager, N. Evaluation of automated natural language processing in the further development of science information retrieval. New York: New York University, 1976. (ERIC Document Reproduction Service No. ED 149 590)

Schank, R. C. & Colby, K. C. Computer models of thought and language. San Francisco: W. H. Freeman and Co., 1973.

Schank, R. C. *Research at Yale in natural language processing.* New Haven, Conn.: Yale University, 1976. (ERIC Document Reproduction Service No. ED 144 560)

Schank, R. C. Computer understanding of natural language. *Behavior Research Methods & Instrumentation,* 1978, *10*(2), 300-306.

Schank, R. C. Language and memory. *Cognitive Science,* 1980, *4*(3), 243-284.

Slobin, D. I. Cognitive prerequisites for the development of grammar. In C. A. Ferguson and D. I. Slobin (Eds.) *Studies of child language development.* New York: Holt, Rinehart, & Winston, 1973.

Slobin, D. I. *Psycholinguistics,* 2nd edition. Glenview, IL: Scott, Foresman & Company, 1979.

Spolsky, B., Holm, W., Holiday, B. & Embry, J. A computer-assisted study of the vocabulary of young Navaho children. *Computers and the Humanities,* 1973, *7*(4), 209-218.

Sukle, R. J., Miron, M., & Pratt, C. *Spoken language vocabulary and structural frequency count: Japanese data analysis.* (SURC-TR-73-228) Syracuse University Research Corporation, Syracuse, New York, July, 1973. (NTIS No. AD 775-925/1)

Wachal, R. S. & Spreen, O. *A computer-aided investigation of linguistic performance: Normal and pathological language* (THEMIS-UI-TR-29). Iowa University Department of Mathematics, July 1970. (NTIS No. AD-714 144)

Widman, R. L. Trends in computer applications to literature. *Computers and the Humanties,* 1975, *9*(5), 231-235.

Winograd, T. Understanding natural language. *Cognitive Psychology,* 1972, *3*,(1), 1-191.

Winograd, T. What does it mean to understand language? *Cognitive Science,* 1980, *4*(3), 209-241.

# Appendix

## Automatic Linguistic Analysis

This appendix consists of abstracts taken from Lockheed Dialog and NTIS searches. The abstracts are grouped in the following categories:

Prepared for the National Center for Bilingual Research

by

**Integrated Research & Information Systems, Inc.**

10150 Sorrento Valley Road
Suite 320
San Diego, California 92121

# COMPUTER MODELS
# OF THOUGHT AND LANGUAGE
# SECTION 1

Understanding Natural Language Using a Variable Grammar

Dartmouth Coll Hanover N H Dept of Mathematics*Office of Naval Research, Arlington, Va. (404325)

Technical rept.
AUTHOR: Harris, Larry R.
C4491A3 FLD: 6D, 9B, 5G, 95P, 62 USGRDR7511
Mar 75 91p
REPT NO: TR75-1
CONTRACT: N00014-73-A-0261
MONITOR: 18

ABSTRACT: A natural language understanding system is described that is designed to work with variable grammars. This is distinct from most natural language systems which can make automatic lexical changes in the dictionary, but can alter the grammar only by actual programming changes. This parsing scheme was developed as part of a larger system that could detect limitations to its grammar and automatically update the grammar, thereby improving its performance. Thus, the need to parse with a variable grammar. The key issue is not the ability to syntactically parse with different grammars, but the ability to mesh semantics with the parses defined by grammars of varying complexity.

DESCRIPTORS: *Natural language, *Artificial intelligence, *Computer applications, Computer programming, Computational linguistics, Context free grammars, Context sensitive grammars, Robots

IDENTIFIERS: Parsing, NTISDCDN

AD-A007 573/9ST NTIS Prices: PC$4.75/MF$2.25

Research At Yale in Natural Language Processing

Yale Univ New Haven Conn Dept of Computer Science (407051)

Technical rept.
AUTHOR: Schank, Roger C.
D1715C4 Fld: 5G, 9B, 92D GRAI7709
1976 32p
Rept No: RR-84
Contract: N00014-75-C-1111
Monitor: 18

Abstract: This report describes the state of the computer programs at Yale that do automatic natural language processing as of the end of 1976. The theory behind the programs shown here as well as descriptions of how these programs function, has been described elsewhere. This report is summarizes the capabilities of 5 computer programs at the present time.

Descriptors: *Natural language, *Information processing, *Reading machines, Artificial intelligence, Computational linguistics, Semantics, Parsers, Concept formation, Reading machines, Machine translation, Computer programs, Intelligibility, Man computer interface, Planning, Newspapers

Identifiers: NTISDODXA

AD-A035 874/7ST NTIS Prices: PC A03/MF A01

Comprehension by Computer: Expectation-Based Analysis of Sentences in Context

Yale Univ New Haven Conn Dept of Computer Science (407051)

Research rept.
AUTHOR: Riesbeck, Christopher K.; Schank, Roger C.
DC131L3 Fld: 5G, 9B, 92D, 62B GRAI7701
Oct 76 82p
Rept No: RR-78
Contract: N00014-75-C-1111
Monitor: 18

Abstract: ELI (English Language Interpreter) is a natural language parsing program currently used by several story understanding systems. ELI differs from most other parsers in that it: produces meaning representations (using Schank's Conceptual Dependency system) rather than syntactic structures; uses syntactic information only when the meaning can not be obtained directly; talks to other programs that make high level inferences that tie individual events into coherent episodes; uses context-based exceptions (conceptual and syntactic) to control its parsing routines. Examples of texts that ELI has understood, and details of how it works are given.

Descriptors: *Comprehension, *Natural language, *Computer applications, Artificial intelligence, Computational linguistics, Semantics, Ambiguity, Interpreters, Concept formation, Computer programming

Identifiers: ELI(English Language Interpreter), English language interpreter, Natural language processors, NTISDODN

AD-A031 587/9ST NTIS Prices: PC A05/MF A01

59-09067 DOC YEAR: 1978 VOL_NO: 59 ABSTRACT NO: 09067
A system for primitive natural language acquisition.
Harris, Larry R.
Dartmouth Coll
International Journal of Man-Machine Studies 1977 Mar Vol 9(2) 153-206
LANGUAGE: Engl CLASSIFICATION: 21. 60

Notes that natural language acquisition deals with 2 very difficult problems in artificial intelligence: computer learning and natural language processing. The present paper focuses on the problems involved in the acquisition of primitive linguistic capability (i.e., when words are first correlated to concepts and when the ordering of the words of utterance first become important). Techniques of acquiring the capability to deal with nested dependent clauses are described. This work is of interest in the field of computer learning inasmuch as it provides an example of an adaptive system that, rather than tuning numeric weights, actually varies its primary structural element, namely the grammar that defines its current language. This work is of interest in the field of natural language processing in that it requires the development of a parsing algorithm robust enough to deal with grammars and dictionaries that vary with time. The ability to automatically extend the grammar to include new sentence forms is also requisite for language acquisition. (20 ref)

SUBJECT TERMS: LANGUAGE DEVELOPMENT. GRAMMAR. COMPUTER SIMULATION. COMPUTER SOFTWARE; 27760. 21530. 10950. 10960
INDEX PHRASE: computer programs. primitive natural language acquisition

7728730 77-3-000646
A Parser for English and Its Application in an Automatic Programming System
Ginsparg, Jerrold Martin
Dissertation Abstracts International, Pt. A US ISSN 0419-4209, Pt. B US ISSN 0419-4217, Ann Arbor, MI, 1977, 38:2756B
Doc Type: journal article
Descriptors: linguistics - linguistics, general - linguistics, computational - mechanolinguistics - programming languages
Descriptor Codes: 0302020003

CAI systems that process natural language.
Koffman, Elliot B.
U Connecticut
Educational Technology 1974 Apr Vol 14(4) 37-42
CLASSIFICATION: 16

Surveys a number of generative systems of computer assisted instruction which have the ability to construct tutorial sequences and respond to student queries by manipulating a data base of relevant information. The systems are oriented toward the humanities and textual manipulation. The use of artificial intelligence research as a theoretical foundation for the natural language processing aspects of these systems is discussed.

SUBJECT TERMS: COMPUTER ASSISTED INSTRUCTION. COMPUTER SOFTWARE; 10920, 10960

INDEX PHRASE: computer-assisted instruction systems, processing natural language

ED145707 FL008979
An Overview of OWL, a Language for Knowledge Representation.
Szolovits, Peter; And Others
Massachusetts Inst. of Tech., Cambridge. Lab. for Computer Science.

Jun 77 28p.; Paper presented at the Workshop on Natural Language for Interaction with Data Bases (Schloss Laxenburg, Austria, January 1977) ; Print is fuzzy on some pages

Sponsoring Agency: Advanced Research Projects Agency (DOD), Washington, D.C.

Contract No.: N00014-75-C-0661

EDRS Price MF-$0.83 HC-$2.06 Plus Postage.

This is a description of the motivation and overall organization of the OWL language for knowledge representation. OWL consists of a linguistic memory system (LMS), a memory of concepts in terms of which all English phrases and all knowledge of an application domain are represented; a theory of English grammar which tells how to map English phrases into concepts; a parser to perform that mapping for individual sentences; and an interpreter tc carry out procedures which are written in the same representational formalism. The system has been applied to the study of interactive dialogs, explanations of its own reasoning, and question answering. (Author/AM)

Descriptors: *Artificial Intelligence/ *Computational Linguistics/ *Computer Programs/ Computer Science/ *English/ Grammar/ Information Processing/ Phrase Structure/ *Programing Languages/ Semantics/ Sentence Structure

Identifiers: *Language Processing/ *OWL/ Parsing

55-00019 DOC YEAR: 1976 VOL NO: 55 ABSTRACT NO: 00019
SOPHIE: A step toward creating a reactive learning environment.
Brown, John S.; Burton, Richard R.; Bell, Alan G.
Bolt Beranek & Newman, Inc, Computer Science Div, Cambridge, MA
International Journal of Man-Machine Studies 1975 Sep Vol 7(5) 675-696
CLASSIFICATION: 21

Describes a fully operational assisted-instruction-computer--assisted-instruction system which incorporates artificial intelligence techniques to perform question answering, hypothesis verification, and theory formation activities in the domain of electronic troubleshooting. Much of SOPHIE's (SOPHIsticated Instructional Environment) logical or inferencing capabilities is derived from uses of simulation models in conjunction with numerous procedural specialists. The system also includes a highly tuned structural parser for allowing the student to communicate in natural language. Although the system is extremely large, it is sufficiently fast to be thoroughly exercised in a training or classroom environment.

SUBJECT TERMS: COMPUTER ASSISTED INSTRUCTION, MAN MACHINE SYSTEMS DESIGN; 10920, 29360

INDEX PHRASE: design philosophy & mechanisms of SOPHIE, operational computer assisted instruction system producing ''reactive'' learning environment

SEMANTIC CATEGORIES

Tracor Inc Austin Tex (352100)
AUTHOR: Schank, Roger C.
461314 FLD: 5G USGRDR6813
Apr 68 24p
REPT NO: TRACOR-68-551-U

ABSTRACT: In order to generate coherent sentences, a conceptual semantics must be utilized that limits possible conceptual dependencies to statements about the real world. This is done by the creation of semantic files that serve to spell out the defining characteristics of a given concept and enumerate the possibilities for relation with other concepts within the range of conceptual experience. The semantic files are created, in part, from a hierarchical organization of semantic categories. The semantic category is part of the definition of a concept and the information at the nodes dominating the semantic category in the hierarchical tree may be used to fill in the semantic file. This report is concerned with the system of semantic categories and their use in the construction of the semantic files. (Author)

DESCRIPTORS: (*Computational linguistics, *Semantics), Classification, Language, Synthesis, Artificial intelligence, Word association, Perception(Psychology)

AD-668 916 CFSTI Prices: PC$6.00 MF$0.95

75077156 v3n9
Transition network grammars for syntactic pattern recognition
Chou, S.M.
Conference on Computer Graphics, Pattern Recognition, and Data Structure A752348 Beverly Hills, California 14-16 May 75
UCLA Extension--in cooperation with the IEEE Computer Society and the ACM Special Interest Group on Computer Graphics
Proceedings available at time of conference, price n a: IEEE Computer Society, Publications Office, 5855 Naples Plaza, Suite 301, Long Beach, Calif. 90803.
Descriptors: TRANSITION; NETWORK; PATTERN; RECOGNITION
SECTION HEADING: MATHEMATICS
Section Class Codes: 6500

ED144560 IR005130 Research
Research at Yale in Natural Language Processing. Research
Report #84.
Schank, Roger C.
76 32p.
Sponsoring Agency: Advanced Research Projects Agency (DOD), Washington, D.C.
Contract No.: N00014-75-C-1111
EDRS Price MF-$0.83 HC-$2.06 Plus Postage.
This report summarizes the capabilities of five computer
programs at Yale that do automatic natural language processing
as of the end of 1976. For each program an introduction to its
overall intent is given, followed by the input/output, a short
discussion of the research underlying the program, and a
prognosis for future development. The programs discussed are:
SAM, a script-based story understanding program; FRUMP, a fast
program designed to skim a newspaper looking for events in
which it is interested; PAM, a plan based program designed to
understand stories that call upon general knowledge of human
goals and relationships; TALESPIN, a program intended to make
up stories to tell in an interactive mode; and WEIS/POLITICS,
a program designed to read newspaper headlines and both code
the sentences into a political coding scheme, and simulate a
person with an ideological belief system being informed of the
event in the headlines. (WBC)
Descriptors: *Artificial Intelligence/ *Computational Linguistics/ *Computer Programs/ *Programing Languages/ *Research
Identifiers: Natural Language Processing/ *Yale University

DO56543* FL002405

Procedures as a Representation for Data in a Computer rogram for Understanding Natural Language.

Winograd, Terry

Massachusetts Inst. of Tech., Cambridge.

Feb 71 464p.; Revised version of a doctoral dissertation. assachusetts Institute of Technology

Sponsoring Agency: Department of Defense, Washington, D.C. dvanced Research Projects Agency.

Report No.: MAC-TR-84

Available from: National Technical Information Service. pringfield, Va. 22151 (AD-721 399, MF $.95, HC $3.00)

Document Not Available from EDRS.

This paper describes a system for the computer understanding f English. The system answers questions, executes commands, nd accepts information in normal English dialogue. It uses emantic information and context to understand discourse and o disambiguate sentences. It combines a complete syntactic nalysis of each sentence with a heuristic understander which ses different kinds of information about a sentence, other arts of the discourse, and general information about the orld in deciding what the sentence means. The objectives of he project are a practical language-understanding system, a etter understanding of what language is and how it is put ogether, and an understanding of what intelligence is and how t can be put into a computer. (Author/VM)

Descriptors: *Computational Linguistics/ *Computer Programs/ omputers/ Deep Structure/ Discourse Analysis/ English/ rammar/ Language/ *Language Skills/ Linguistic Theory/ Logic/ rograming Languages/ *Semantics/ Sentences/ Sentence tructure/ Structural Analysis/ Structural Linguistics/ Syntax/ Transformation Theory (Language)

OUTLINE OF A CONCEPTUAL SEMANTICS FOR GENERATION OF COHERENT DISCOURSE

Tracor Inc Austin Tex (35210C)
AUTHOR: Schank, Roger C.
4605A3 FLD: 5G USGRDR6813
Mar 68 45p
REPT NO: TRACOR-68-462-U

ABSTRACT: The paper develops a method for generating coherent sentences. A conceptual semantics is presented, that when coupled with a conceptual dependency abstraction of meaning, allows concepts to be linked in a manner consonant with the system's knowledge of the world. The paper is part of a series of papers concerned with the problem of language synthesis for artificially intelligent systems. (Author)

DESCRIPTORS: (*Computational linguistics, *Semantics), Artificial intelligence, Language, Synthesis

AD-668 724 CFSTI Prices: PC$6.00 MF$0.95

ED039510 AL002310

Spinoza II: Conceptual Case-Based Natural Language Analysis.

Schank, Roger C.; And Others

Stanford Univ., Calif. Artificial Intelligence Project.

Jan 70 107p.

Sponsoring Agency: Department of Defense, Washington, D.C. dvanced Research Projects Agency.; National Inst. of Mental ealth (DHEW), Bethesda, Md.

Report No.: M-AIM-109

EDRS Price MF-$0.76 HC Not Available from EDRS. PLUS POSTAGE

This paper presents the theoretical changes that have eveloped in Conceptual Dependency Theory and their amifications in computer analysis of natural language. The ajor items of concern are: the elimination of reliance on grammar rules" for parsing with the emphasis given to onceptual rule based parsing; the development of a conceptual ase system to account for the power of conceptualizations; he categorization of ACT's based on permissible conceptual ases and other criteria. These items are developed and iscussed in the context of a more powerful conceptual parser nd a theory of language understanding. (Author/AMM)

Descriptors: *Case (Grammar)/ *Computational Linguistics/ oncept Formation/ Conceptual Schemes/ Deep Structure/ Linguistic Theory/ Semantics/ *Structural Analysis/ *Thought rocesses/ Translation/ Verbs

Identifiers: *Conceptual Dependency Theory/ Spinoza II

A CONCEPTUAL DEPENDENCY REPRESENTATION FOR A COMPUTER-ORIENTED SEMANTICS

Stanford Univ., Calif. Dept. of Computer Science. (094 120)

Technical rept.
AUTHOR: Schank, Roger C.
619424 FLD: 9B, 5G, 906 USGRDR6914
Mar 69 209p
REPT NO: CS-130, AI Memo-83
CONTRACT: PHS-MH-06645-07

ABSTRACT: Machines that may be said to function intelligently must be able to understand questions posed in natural language. Since natural language may be assumed to have an underlying conceptual structure, it is desirable to have the machine structure its own experience, both linguistic and nonlinguistic, in a manner concomitant with the human method for doing so. Some previous attempts at organizing the machine's data base conceptually are discussed. A conceptually-oriented dependency grammar is posited as an interlingua that may be used as an abstract representation of the underlying conceptual structure. The conceptual dependencies are utilized as the highest level in a stratified system that incorporates language-specific realization rules to map from concepts and their relations, into sentences. In order to generate coherent sentences, a conceptual semantics is posited that limits possible conceptual dependencies to statements about the system's knowledge of the real world. The system has been programmed; coherent sentences have been generated and the parser is operable. The entire system is posited as a viable linguistic theory. (Author)

DESCRIPTORS: (*Learning machines, Artificial intelligence), ( *Programming languages, *Computational linguistics), English language, Semantics, Programming(Computers), Grammars, Theses

PB-183 907 CFSTI Prices: HC$6.00 MF$0.95

48-09128 DOC YEAR: 1972 VOL NO: 48 ABSTRACT NO: 09128
Understanding natural language.
Winograd, Terry
Massachusetts Inst. of Technology
Cognitive Psychology 1972, Jan. Vol. 3(1), 191 p
CLASSIFICATION: 11

Describes a computer system that answers questions, executes commands, and accepts information in an interactive English dialogue. It is based on the assumption that in modeling language understanding, we must deal in an integrated way with all of the aspects of language syntax, semantics, and inference. The system contains a parser, a recognition grammar of English, programs for semantic analysis, and a general problem solving system. It can (a) remember and discuss its plans and actions as well as carrying them out; (b) enter into a dialogue with a person, responding to English sentences with actions and English replies; and (c) ask for clarification when its heuristic programs cannot understand a sentence through the use of syntactic, semantic, contextual, and physical knowledge. Knowledge in the system is represented in the form of procedures, rather than tables of rules or lists of patterns. By developing special procedural representations for syntax, semantics, and inference, flexibility and power are gained. Since each piece of knowledge can be a procedure, it can call directly on any other piece of knowledge in the system. (3 p. ref.)

SUBJECT TERMS: Language, Computers, Syntax, Grammar, Semantics; 27740, 10970, 51220, 21530, 46390

INDEX PHRASE: language understanding computer system, special procedural representations for syntax & semantics & inference

7400083 7400083
Understanding natural language
BOOK AUTHOR: Winograd, T.
Johnson-Laird, P. N.
the Quarterly Journal of Experimental Psychology~ 1973, 25
(3), 444-446. CODEN: qjxp-a
Series: REVIEW
New York: Academic Press, 1972.for U. S., Canada, Central
America, and South America. Academic Press, 111 Fifth Ave.,
New York NY 10003; and for all other countries, Academic
Press, 24-28 Oval Rd., London NW1 England;
Section Heading Codes: 012 LANGUAGE: Engl.
A favorable review of Winograd's computer program for
understanding natural language. What the program evidently
does is to converse by Teletype about a small imaginary world
of blocks, boxes, and cubes, and, in response to orders, it
moves around the objects to make up any required
configuration. A number of language and problem solving skills
interact in a complex fashion enabling the program to carry
out the tasks given and to conduct lucid conversations with
interlocutors. It is accomplished by a highly skilled
deployment of a whole set of programs. Syntactic analysis is
based on a systemic grammar developed originally by Halliday,
and an integrated approach is taken to the interpretation of
sentences. Meaning is liberated from its specific verbiage by
treating it as a matter of underlying concepts, and
selectional restrictions are treated in a standard way to
determine which particular meaning of a word is relevant.
Meanings of sentences are represented by expressions in
Planner, the language that also underlies the inferential
power of the system. There is much in the system to interest
linguists, computer programmers, and workers in artificial
intelligence as well as psychologists and psycholinguists. In
addition to the description of the system, there are useful
introductions to Lisp, Programmer, and Planner, and an account
is presented of systemic grammar that is almost unrivalled in
bringing out the simplicity of the basic ideas. M. Guck
Descriptors: COMPUTATIONAL LINGUISTICS; SEMANTICS; RESEARCH
DESIGN AND INSTRUMENTATION; SYNTAX; SYNTHETIC LANGUAGES
Identifiers: computer program for understanding natural
language; book review;

7804043 7804043
On Natural Language Based Computer Systems
Petrick, S. R.
IBM Thomas J. Watson Research Center, Yorktown Heights NY
10598
IBM Journal of Research and Development~ 1976, 20, 4, July,
314-325. CODEN: ibmj-a
International Business Machines Corporation, Armonk NY 10504
Section Heading Codes: 5113
Arguments for & against the use of natural langs in
question-answering & programming systems are discussed.
Several natural lang-based computer systems are considered in
assessing the current level of system development. The first
system is the LSNLIS (Lunar Sciences Natural Lang Information
System), containing information about lunar rock & soil
derived from Apollo missions. It was able to answer 78% of
the queries posed by lunar geologists but only a much smaller
% of follow-up queries. The REL (Rapidly Extensible Lang)
system has been applied to questioning of anthropological
data, class scheduling & Fortune 500 data question-answering.
The core Eng lang is extensible by means of definition based
on string substitution. The SHRDLU system developed by T.
Winograd demonstrates that it is possible to bring together
syntactic, semantic, inferential & graphical capabilities in a
single system. It has a more highly developed response
generator than the above system. NLP (Natural Language
Processing) was used to develop an automatic programming
system for queueing systems (see Hiedorn, G. E. "Automatic
Programming through Natural Language Dialogue: A Survey," IBM
Journal of Research & Development, 1976, 20, 4, 302-313.).
The REQUEST (Restricted English QUESTion-Answering system) is
based on a transformational grammar of Eng. Certain pervasive
difficulties in developing natural lang based systems are
identified, & the approach taken to overcome them in the
REQUEST system is described. Modified HA
Descriptors: COMPUTATIONAL LINGUISTICS; EXPERIMENTAL DATA
HANDLING
Identifiers: natural language based computer systems;

An Overview of OWL, A Language for Knowledge Representation :

Massachusetts Inst of Tech Cambridge Lab for Computer Science
409648)
AUTHOR: Szolovits, Peter; Hawkinson, Lowell B.; Martin, William A.
DDD13G3 Fld: 5G, 9B, 92D, 62B GRAI7719
Jun 77 29p
Rept No: MIT/LCS/TM-86
Contract: N00014-75-C-0661
Monitor: 18
Presented at Workshop on Natural Language for Interaction with Data Bases held by the International Institute for Applied Systems Analysis at Schloss Laxenburg, Austria, Jan 77.

Abstract: The motivation and overall organization of the OWL language for knowledge representation is described. OWL consists of a memory of concepts in terms of which all English phrases and all knowledge of an application domain are represented, a theory of English grammar which tells how to map English phrases into concepts, a parser to perform that mapping for individual sentences, and an interpreter to carry out procedures which are written in the same representational formalism. The system has been applied to the study of interactive dialogs, explanations of its own reasoning, and question answering.

Descriptors: *Programming languages, Artificial intelligence, Symbols, Computational linguistics, Phrase structure grammars, Computer applications, Parsers, Mapping, Man machine systems, Linguistics, Taxonomy, Nodes, Words(Language), Semantics, Reasoning, Psycholinguistics, Indexing

Identifiers: Knowledge Representation, Sentences, Interactive systems, Question answering systems, *OWL programming language, Linguistic memory system, Dialogues, NTISDODXA

AD-A041 372/4ST NTIS Prices: PC A03/MF A01

62-07606 DOC YEAR: 1979 VOL NO: 62 ABSTRACT NO: 07606
Computer understanding of natural language.
Schank. Roger C.
Yale U
Behavior Research Methods & Instrumentation 1978 Apr Vol 10(2) 132-138
LANGUAGE: Engl CLASSIFICATION: 21. 60
Describes the state of research in automatic natural language processing at Yale University. Five programs are discussed: SAM. FRUMP. PAM. TALESPIN. and POLITICS.
SUBJECT TERMS: LANGUAGE. COMPUTER SOFTWARE. COMPUTER APPLICATIONS: 27740. 10960. 10900
INDEX PHRASE: automatic natural language processing. discussion of SAM & FRUMP & PAM & TALESPIN & POLITICS programs

# THEORETICAL LINGUISTIC
# MODELS AND PARSERS
# SECTION 2

?300833 7300833
Theoretical and methodological considerations on automatic
yntactic analysis
Geens, Dirk
Inst. Applied Linguistics. Louvain, Belgium
ITL- 1972. 15, 47-66. CODEN: itlg-a
Institute of Applied Linguistics. Vesaliusst. 2. 3000
ouvain, Belgium:
Section Heading Codes: 062

Automatic analysis (AA) is one of the most disputed
isciplines in applied linguistics. Not only a computer
rogram but also, for example, a transformational grammar
ries to obviate the endless number of grammatical sentences
ith a finite set of rules. Applying the theory is only one
tep further than defining the theory itself. Thus, if the
nalysis turns out to be wrong, the linguist will first
ttempt to correct the theoretical model because it has been
hown to be wrong, whereas those who disagree with the applied
ethod can only adjust their theory in a haphazard fashion.
pplication and theory are thus dependent on each other. As a
esult, the linguist will benefit most by a combination of
heory and practice. Any AA should have a double aim: (1) as
speculum for the model used; and (2) as being applicable in
ields other than pure linguistics, e.g., in the description
f language. The ASA program is divided into paradigmatic and
yntagmatic parts. Whereas in the paradigmatic analysis the
elations that exist in words between actualized and potential
alency are indicated, the syntagmatic analysis will indicate
he relations between the words that constitute the sentence.
The detailed ASA program can only be evaluated by the extent
o which it can now live up to expectations in actual
ractice. The AA program has a double aim: (1) the
erification of the model used for linguistic descriptions;
nd (2) if this model seems to satisfy present needs, the
ctual application of the model. Grammar can indeed be
ormalized, and as a result must be made machine-applicable,
ecause this would seem to be the only way in which
rammatical theories can be examined in order to avoid
isleading interpretations made by the "understanding reader."
he next step must be to evaluate the model used here with
egard to the large group of existing theories. To this end,
fforts in the field of the formalization of grammar and hence
utomatic analysis must be increased.

Descriptors: DATA PROCESSING AND RETRIEVAL; APPLIED
INGUISTICS; SYNTAX; THEORETICAL LINGUISTICS
Identifiers: automatic syntactic analysis; theory,
ethodology;

7700081 7700081
Human Associative Memory
BOOK AUTHOR: Anderson, John R. & Bower, Gordon H.
Keenan, Janice M.
U Denver, University Park CO 80210
Language Sciences- 1976, 39, Feb, 30-32. CODEN: lasc-b
Series: REVIEW
New York: Wiley, Halsted Press, 1973.Research Center for the
Language Sciences, Indiana University, 516 E. 6th St.,
Bloomington IN 47401
Section Heading Codes: 4016 LANGUAGE: Engl

In recent years it has become apparent that the distinction
between linguistic competence & linguistic performance is
quite fuzzy. What is needed is a model that unites the 2 -- a
model that represents a speaker/hearer's knowledge of the
anguage in terms of the rules or processes required to change
from 1 mental state to the next. This is an impressive
attempt at such a model. While the model suffers from its
reliance on the traditional, yet questionable, tenets of
associationism, the book does an excellent job of presenting &
analyzing the problems involved in constructing a natural
anguage processing system. It provides many insights for
readers interested in the interface between competence &
performance. AA

Descriptors VERBAL LEARNING; MEMORY; COMPETENCE AND
PERFORMANCE; PSYCHOLINGUISTICS
Identifiers: human associative memory; competence vs.
performance; book review;

Computational Understanding: Analysis of Sentences and Context

Stanford Univ Calif Dept of Computer Science*Advanced Research Projects Agency, Arlington, Va.*National Inst. of Mental Health, Rockville, Md. (094120)

Technical rept.
AUTHOR: Riesbeck, Christopher Kevin
C4262A3 FLD: 5G, 92D* USGRDR7508
May 74 250p
REPT NO: STAN-CS-74-437, AIM-238
CONTRACT: DAHC15-73-C-0435, PHS-MH-06645
PROJECT: ARPA Order-2494
MONITOR: 18

ABSTRACT: The goal of this thesis was to develop a system for the computer analysis of written natural language texts that could also serve as a theroy of human comprehension of natural language. Therefore the construction of this system was guided by four basic assumptions about natural language comprehension. First, the primary goal of comprehension is always to find meanings as scon as possible. Other tasks, such as discovering the syntactic relationships, are performed only when essential to decisions about meaning. Second, an attempt is made to understand each word as soon as it is read, to decide what it means and how it relates to the rest of the text. Third, comprehension means not only understanding what has been seen but also predicting what is likely to be seen next. Fourth, the words of a text provide the cues for finding the information necessary for comprehending that text.

DESCRIPTORS: *Computational linguistics, Natural language, Data processing, Speech recognition, Semantics, Syntax

IDENTIFIERS: NTISDODA

AD/A-005 040/1ST NTIS Prices: PC$7.50/MF$2.25

7802550 7802550
The Computer and Literary Studies
BOOK AUTHOR: Aitken, A. J., Bailey, R. W., Hamilton-Smith, N. (Eds)
Greenblatt, Daniel L.; Tallentire, D. R.; Martin, W.
Style- 1976, 10, 3, summer, 281-295. CODEN: styl-b

ED107151 FL006923
Detecting Syntactic Ambiguity: Three Augmented Transition Network Techniques.
Herman, L. Russell, Jr.
21 Mar 75 21p.; Paper presented at the Southeastern Conference on Linguistics (SECOL) (13th, Vanderbilt University, March 1975)
EDRS Price MF-$0.76 HC-$1.58 PLUS POSTAGE
When a grammar is expressed in augmented transition network (ATN) form, the problem of detecting syntactic ambuguity reduces to finding all possible paths through the ATNs. Each successfully terminating path through the ATN generates an acceptable parsing of the input string. Two ATN forms, minimal-node and pseudo-tree, are described along with the conventions for traversing each. The two forms are compared in regard to efficient use of computer time and space and in regard to appropriateness for each of the three path-finding techniques. Three techniques are discussed for finding all acceptable paths through ATNs. The techniques are "Backtracking," "Simultaneous Parallel Analysis," and "Amputate And Re-enter." Relative merits of the three techniques are discussed in terms of computer execution time, required data storage, programmer time, and amenability of the program to modification. A rudimentary ATN-based parser for English has been written in SPITBOL to test the implementation of these techniques. (Author)

STRUCTURAL SIGNS OF CERTAIN CLASSES OF COMPLEX SENTENCES (IN CONNECTION WITH THE QUESTION OF HOMONYMOUS CONJUNCTIONS) (STRUKTURNYE PRIZNAKI NEKOTORYKH KLASSOV SPOZHNOPODCHINENNYKH PREDLOZHENII) (V SVYAZI S VOPROSOM OB OMONIMII SOYUZOV)

Foreign Technology Div Wright-Patterson AFB Ohio (141600)
AUTHOR: Kaplan, L. I.
5002C1 FLD: 5G USGRDR6820
25 Aug 67 39p
REPT NO: FTD-TT-65-1893
Unedited rough draft trans. of Nauchno-Tekhnicheskaya Informatsiya (USSR) n3 p36-43 1964.

ABSTRACT: The author deals with the subject of complex subordinate clauses within a sentence in which homonymic connecting words are used. The relationship between the main and subordinate clauses, and the function of words within the sentence (i.e., how a word tends to govern, or is governed by other words, the presence of certain grammatical forms in words, etc.) are discussed. (Author)

DESCRIPTORS: (*Machine translation, Russian language), (*Russian language, *Syntax), Semantics, Algorithms, Analysis, Computational linguistics, USSR

IDENTIFIERS: Translations, Homonyms

AD-673 454 CFSTI Prices: PC$6.00 MF$0.95

TRANSFORMATIONS AND DISCOURSE ANALYSIS PAPERS. 69. COMPUTABLE AND UNCOMPUTABLE ELEMENTS OF SYNTAX

Pennsylvania Univ., Philadelphia. (278 950)
AUTHOR: Hiz, Henry
6843A4 FLD: 5G, 917 USGRDR6924
1967 18p
GRANT: NSF-557

ABSTRACT: A syntax of a language may be said to be computable in a different sense when it assigns, in a computable way, for each given usable text, all its relevant structures. One also may call a syntax computable if all its rules are decidable, in the sense that for each pair of texts it is decidable whether they are linked by the rule. (Author)

DESCRIPTORS: (*Linguistics, Analysis), (*Syntax, Mathematics), Computational linguistics, English language

IDENTIFIERS: Generative grammars, Strings(Linguistics)

PB-186 473 CFSTI Prices: HC$3.00 MF$0.95

EJ198840 TM504029
The Sausage Machine: A New Two-Stage Parsing Model.
Frazier, Lyn; Fodor, Janet Dean
Cognition, v6 n4 p291-325 Dec 1978 Dec78
Language: ENGLISH

The human sentence parsing device assigns phrase structure to sentences in two steps. The first stage parser assigns lexical and phrasal nodes to substrings of words. The second stage parser then adds higher nodes to link these phrasal packages together into a complete phrase marker. This model is compared with others. (Author/RD)

Descriptors: *Language Processing/ *Linguistic Theory/ Models/ Phrase Structure/ Psycholinguistics/ Sentence Diagraming/ *Sentence Structure/ Syntax
Identifiers: *Parsing

ED037734 AL002368

An Approach to the Semantics of Verbs.
von Glasersfeld, Ernst
Georgia Inst. for Research, Athens.
Apr 70 18p.; Paper delivered at the Southeastern Conference on Linguistics, Chapel Hill, North Carolina, April 1970
Sponsoring Agency: Air Force Office of Scientific Research, Arlington, Va. Directorate of Information Science.
EDRS Price MF-$0.76 HC-$1.58 PLUS POSTAGE

This paper explains a method of semantic analysis developed in the course of a natural-language research project that led to the computer implementation of the Multistore Parser. Positing an interlinguistic substratum of semantic particles of several different types (e.g. substantive, attributive, developmental, relational), a method is illustrated which makes it possible to map the meaning of activity words in context; the resulting mappings, on the one hand, incorporate much of what, hitherto, has been considered "pragmatics," and on the other, they furnish an exact definition of the semantic "deep structure" underlying the grammatical surface structure of a phrase or sentence. The mappings are here used to demonstrate semantic similarities and discrepancies between an English verb and the German verbs which are required for its translation in various contexts. (Author/FWB)

Descriptors: Computational Linguistics/ *Deep Structure/ *English/ *German/ Mathematical Linguistics/ *Semantics/

75021511 v3n2

Transformations & inference of tree grammars for syntactic pattern recognition
Bhargava, B.K.
Purdue U, West Lafayette, Ind.
IEEE Systems, Man and Cybernetics Society 1974 International Conference A744295 Dallas, Tex 2-4 Oct 74
IEEE Systems, Man and Cybernetics Society
Conference Record No. 74CH0908-4 SMC; inquire: Order Dept., Institute of Electrical and Electronics Engineers, 345 East 47 St., New York, N. Y. 10017.
Descriptors: TRANSFORMATION; TREE; PATTERN; RECOGNITION
SECTION HEADING: MATHEMATICS
Section Class Codes: 6500

75021505 v3n2

On inference of tree grammars for syntactic pattern recognition
Gonzalez, R.C.
U Of Tennessee, Knoxville, Tenn.
IEEE Systems, Man and Cybernetics Society 1974 International Conference A744295 Dallas, Tex 2-4 Oct 74
IEEE Systems, Man and Cybernetics Society
Conference Record No. 74CH0908-4 SMC; inquire: Order Dept., Institute of Electrical and Electronics Engineers, 345 East 47 St., New York, N. Y. 10017.
Descriptors: TREE; PATTERN; RECOGNITION
SECTION HEADING: MATHEMATICS
Section Class Codes: 6500

7720188 77-3-000103

Theoretical Issues in Natural Language Processing
Papers from an Interdisciplinary Workshop in Computational Linguistics, Psychology, Linguistics, Artificial Intelligence, 10-13 June, 1975, Cambridge, MA
Nash-Webber, Bonnie; Schank, Roger
Cambridge, MA: Yale Univ. Mathematical Soc. Sciences Board. 1975. 219 pp.
Doc Type: festschrift
Descriptors: linguistics - collections, analyzed
Descriptor Codes: 0301000000

73062239 v1n7

Natural language processing
Joshi, A.K.
1973 National Computer Conference A732237 New York, N Y 4-8 Jun 73
American Federation of Information Processing Societies
Proceedings, 9 Jun 73; $40.00; Mr. T. C. White, American Federation of Information Processing Societies, 210 Summit Ave., Montvale, N.J. 07645.
Descriptors: LANGUAGE; PROCESSING
SECTION HEADING: GENERAL ENGINEERING AND TECHNOLOGY
Section Class Codes: 5000

7804059 7804059

Observations on Context Free Parsing

Sheil, B. A.

Statistical Methods in Linguistics- 1976, 71-109. CODEN: smln-a

Spra'kforlaget Skriptor, P.O. Box, 104 65 Stockholm 15, Sweden (Name changed to Journal of Linquistic Calculus after 1976 Volume)

Section Heading Codes: 5113

The principles underlying context free parsing are investigated. The use of a well-formed substring table is sufficient to achieve polynomially bounded parsing. On the basis of its presence in all known polynomial parsers, such a device may also be necessary to achieve this bound. The desirability of a parser automatically achieving tighter bounds for various subclasses of the context free grammars is examined & found to be dependent on the subclass concerned. It is argued that use of a transformed grammar by the parser is not necessarily a disadvantage, as has been previously claimed. As an illustration of these ideas, a variant of recursive descent parsing is developed & its behavior analyzed. This algorithm, when equipped with a well-formed substring table, is shown to be as efficient as any known general purpose context free parser, while its simple structure makes it easier to understand & prove correct. Modified HA

Descriptors: CONTEXT FREE GRAMMAR; STRUCTURALIST LINGUISTIC THEORY

Identifiers: context free parsing;

7502857 7502857

Pattern-matching rules for the recognition of natural language dialogue expressions

Colby, Kenneth Mark; Parkison, Roger C.; Faught, Bill

Computer Science Stanford U CA 94305

American Journal of Computational Linguistics- 1974, 1, Microfiche 5, 1-82. CODEN: ajcl-d

Center for Applied Linguistics, 1611 N. Kent St., Arlington VA 22209 (Including The Finite String as of 1974, Vol. 11, No. 1)

Section Heading Codes: 116

Man-machine dialogues using everyday conversational English present difficult problems for computer processing of natural language. Grammar-based parsers which perform a word-by-word, parts-of-speech analysis are too fragile to operate satisfactorily in real time interviews allowing unrestricted English. In constructing a simulation of paranoid thought processes, an algorithm capable of handling the linguistic expressions used by interviewers in teletyped diagnostic psychiatric interviews was designed. The algorithm uses pattern-matching rules which attempt to characterize the input expressions by progressively transforming them into patterns which match, completely or fuzzily, abstract stored patterns. The power of this approach lies in its ability to ignore recognized and unrecognized words and still grasp the meaning of the message. The methods utilized are general and could serve any "host" system which takes natural language input. Appendices contain a sample interview, the dictionary, and a list of simple patterns. HA

Descriptors: DYADIC INTERACTION; DATA PROCESSING AND RETRIEVAL; ENGLISH; MEANING; SPEECH RECOGNITION BY MACHINE

Identifiers: algorithm for pattern-matching rules for computer recognition of natural English dialogue;

# MACHINE TRANSLATION

## SECTION 3

7602860 7602860
Junction Grammar as a Base for Natural Language Processing
Lytle, Eldon G.; Packard, Dennis; Gibb, Daryl; Melby, Alan K.; Billings, Floyd H., Jr.
Brigham Young U, Provo UT 84601
American Journal of Computational Linguistics- 1975, 3, 77.
CODEN: ajcl-d
Center for Applied Linguistics, 1611 N. Kent St., Arlington VA 22209 (Including The Finite String as of 1974, Vol. 11, No. 1)
Section Heading Codes: 065

Junction Grammar, a model of language structure developed by Eldon Lytle, is being used to define the interlingua for a machine-assisted translation project. Junction Grammar representations, called junction trees, consist of word-sense information interrelated by junctions, which contribute syntactic & semantic information. The 1st step of the current translation system is interactive analysis, during which the program interacts with the human operator to resolve ambiguities & then produces a junction tree representation of the meaning of the input text. The 2nd & 3rd steps of the translation process are automatic transfer & synthesis into 1 or more target languages. For each target language the transfer step makes adjustments on each junction tree, if needed, before sending it to the synthesis program for that language. This translation system is currently under development at Brigham Young U. Present lexicons for English analysis, & Spanish, German, French, & Portuguese synthesis contain about 10,000 word-senses each. HA

Descriptors: COMPUTATIONAL LINGUISTICS; MACHINE TRANSLATION; INTERNATIONAL LANGUAGES; AMBIGUITY; MEANING; ENGLISH; SPANISH; GERMAN; FRENCH; ROMANCE LANGUAGES
Identifiers: junction grammar; model language structure for natural language processing, machine translation;

ENGLISH DICTIONARY CLASSIFICATION

Linguistics Research Center, Univ. of Texas, Austin. (208 250)
AUTHOR: Lee, Tuie Git,
0313F4 FLD: 5G USGRDR6603
Aug 65 29p
REPT NO: LRC-65-WD-1
GRANT: NSF-GN-308
See also PB-166 656. Distribution: No limitation.

ABSTRACT: The paper contains a description of the classification of English adjectives, nouns and verbs in the Linguistics Research System. Paradigms have been devised in chart form defining certain characteristics peculiar to subclasses to parts of speech for adjectives, nouns and verbs. Concise explanations of each subclass with examples are also given. All subclasses are ordered with the most frequently used subclasses listed first.

DESCRIPTORS: (*English language, Classification), Computational linguistics, Semantics, Syntax, Machine translation, Dictionaries

IDENTIFIERS: Adjectives, Nouns, Verbs

PB-168 758 CFSTI Prices: PC$6.00 MF$0.50

77030885 v5n4
Parsing of natural language sentences containing unknown words
Dankel, D.D.
U Of Illinois, Urbana, Il.
Association for Computing Machinery North Central Regional Conference A771149 Urbana, Illinois 25-26 Mar 77
Association for Computing Machinery (North Central Region) Proceedings, 26 Mar 77, $5 plus mailing costs: Student ACM, Dept of Computer Science, Univ. of Illinois, Urbana, IL 61820.
Descriptors: LANGUAGE; UNKNOWNS; WORD
SECTION HEADING: MATHEMATICS
Section Class Codes: 6500

AN AUTOMATIC PHRASE STRUCTURE ANALYSIS OFA SPANISH TEXT

Linguistics Research Center, Univ. of Texas, Austin. (208 250)
AUTHOR: Thomas, Carolyn Beth,
0444E1 FLD: 5G USGRDR6610
Sep 65 131p
REPT NO: LRC-65-WD-2
GRANT: NSF-GN-308

ABSTRACT: A summary of morphological and syntactic classification is presented for a pilot description of Spanish in the Linguistics Research System. Sample displays are given for context-free phrase structure description and the resulting machine analysis. (Author)

DESCRIPTORS: (*Spain, Language), (*Language, Spain), Context free grammars, Computational linguistics, Syntax

PB-169 468 CFSTI Prices: PC$13.60 MF$1.00

Semantic Directed Translation of Context Free Languages

Ohio State Univ., Columbus. Computer and Information Science Research Center.*National Science Foundation, Washington, D.C. (407 586)

Technical rept.
AUTHOR: Buttelmann, H. William
C5C42K4 FLD: 05G, 92D USGRDR7519
Sep 74 39p
REPT NO: OSU-CISRC-TR-74-6
GRANT: NSF-GN-534.1
MONITOR: 18

ABSTRACT: A formal definition for the semantics of a context free language, called a phrase-structure semantics, is given. The definition is a model of the notion that it is phrases which have meaning and that the meaning of a phrase is a function of its syntactic structure and of the meanings of its constituents. Next the author gives a definition for translation on context free languages. He then studies a certain kind of translation on cfl's, which proceeds by translating on the phrase trees of the languages, and is specified by a finite set of tree-replacement rules. The author presents a procedure which, given a cfg and phrase-structure semantics for a target language, will (usually) produce the finite set of tree-replacement rules for the translation, if the translation exists. The procedure may be viewed as a computer program which is a translator generator, and which produces another program that is a translator.

DESCRIPTORS: *Phrase structure grammars, *Semantics, *Machine translation, Syntax, Computational linguistics, Recursive functions, Algorithms

IDENTIFIERS: Phrase structure semantics, *Context free grammars, NTISNSFSIS

PB-242 854/8ST NTIS Prices: PC$3.75/MF$2.25

Syntactic Analysis of the Russian Sentence

Ibm Watson Research Center Yorktown Heights N Y (349 250)

Final rept. May 65-May 67
AUTHOR: Plath, Warren J.; Andreyewsky, Alexander; Strom, Robert E.;
Lippman, Erhard O.
D1773B4 Fld: 5G, 5B d7709
Oct 67 170p
Contract: AF 30(602)-3782
Project: AF-4599
Monitor: RADC-TR-67-484
Distribution limitation now removed.

Abstract: The report describes results of a two year research effort in the field of automatic syntactic analysis of Russian within the framework of Russian-English machine translation R and D. The primary object of study and investigation consisted in design and development of the combinatorial syntactic analysis system, accompanied by an extensive linguistic research on Russian grammar. A concomitant small scale research on multiple path predictive syntactic analysis of Russian was conducted in parallel as an extension of the research effort initiated at Harvard University with the NSF support. Performance of the predictive analyzer on the test corpus of 160 Russian sentences is described.

Descriptors: (*Russian language, *Syntax), (*Machine translation, Russian language), Computational linguistics, Automatic, English language, Computer programs, Programming languages, Algorithms, Combinatorial analysis, Dictionaries, Subroutines, Linguistics

Identifiers: Syntactic analysis, NTISDODXD

AD-824 057/4ST NTIS Prices: PC A08/MF A01

Machine Translation (A Bibliography with Abstracts)

National Technical Information Service, Springfield, Va. (391 812)

Rept. for 1964-Feb 75
AUTHOR: Lehmann, Edward J., Young, Mary E.
C4654D3 FLD: 05G, 09B, 92D*, 88, 62, 86W USGRDR7513
May 75 132p*
MONITOR: 18
Supersedes COM-73-11717.

ABSTRACT: Studies on machine translation of various languages are presented as abstracts in this bibliography of Federally-funded research reports. Topics concerning syntax, computer programming, computer hardware, and semantics are included. (Contains 127 abstracts).

DESCRIPTORS: *Machine translation, *Bibliographies, Computational linguistics, Syntax, Semantics, Computer programming, Vocabulary, Translating

IDENTIFIERS: NTISNTIS

NTIS/PS-75/411/9ST NTIS Prices: PC$25.00/MF$25.00

Knowledge-Based Machine Translation

Yale Univ New Haven CT Dept of Computer Science (407051)

Research rept.
AUTHOR: Carbonell, Jaime G.; Cullinford, Richard E.; Gershman, Anatole V.
F069551 Fld: 5G, 92D, 957 GRAI7910
Dec 78 63p
Rept No: RR-145
Contract: N00014-75-C-1111
Monitor: 18
Availability: Microfiche copies only.

Abstract: This paper discusses knowledge-based machine translation research at Yale University Artificial Intelligence Laboratory. Our paradigm, illustrated by several working computer programs, is to analyze the source text into a language-free representation, apply world knowledge to infer information implicit in the input text, and generate the translation in various target languages. (Author)

Descriptors: *Machine translation, *Artificial intelligence, *Computational linguistics, Natural language, Information processing

Identifiers: Knowledge, NTISDODXA

AD-A062 691/2ST NTIS Prices: MF A01

Research on Chinese-English Machine Translation

California Univ Berkeley (071 850)

Final technical rept. 1 Jul 67-31 Jul 68
AUTHOR: Wang, William S-Y; Dougherty, Ching-Yi; Doughty, Herbert III; Johnson, C. Douglas; Lee, Sally H.
D032153 Fld: 5G d7702
Feb 69 46p
Contract: F30602-67-C-0347
Project: AF-4599
Monitor: RADC-TR-68-570
Distribution limitation now removed.

Abstract: The report documents results of a 13-month effort in Chinese-English machine translation R and D. Main emphasis was placed on design of automatic lookup system for segmentation of Chinese text into units of meaning, and design of automatic syntactic analysis system for recognition of Chinese sentence structure. The following tasks were progressing concurrently: further compilation of lexical data with refined grammar codes, and continuing sophistication of rules for automatic syntactic analysis. Completion of Syntactic Analysis System (SAS) and associated subroutines constitutes a major achievement. Continuation phase will be devoted mainly to interlingual transfer problem and synthesis in English, culminating in design of a prototype system for Chinese-English machine translation. (Author)

Descriptors: (*Chinese language, *Machine translation), Syntax, Computational linguistics, English language

Identifiers: NTISDODXD

AD-650 009/2ST NTIS Prices: PC A03/MF A01

Machine Translation (A Bibliography with Abstracts)

National Technical Information Service, Springfield, Va. (391 812)

Rept. for 1964-May 76
AUTHOR: Young, Mary E.
C6731E2 FLD: 05G, 09B, 92D*, 88, 62, 86% GRAI7615
Jun 76 141p*
MONITOR: 18
Supersedes NTIS/PS-75/411, and COM-73-11717.

ABSTRACT: Studies on machine translation of various languages are cited. Topics concerning syntax, computer programming, computer hardware, and semantics are included. (This updated bibliography contains 136 abstracts, 9 of which are new entries to the previous edition.)

DESCRIPTORS: *Bibliographies, *Machine translation, Computational linguistics, Syntax, Semantics, Computer programming, Vocabulary, Translating

IDENTIFIERS: Foreign languages, NTISNTIS

NTIS/PS-76/0434/1ST NTIS Prices: PC$25.00/MF$25.00

7305019 7305019
Automatic translation of natural languages
Kay, Martin
Information & Computer Science, U. California, Irvine
Daedalus- 1973, 102 (3), 217-230. CODEN: daed-a
280 Newton St., Brookline, Mass. 02146:
Section Heading Codes: 045

A consideration of attempts to build a translating machine for natural languages as well as a discussion of problems in the study of meaning. Although withdrawal of government funding has caused a loss of interest in automatic translation, some systems have been developed including: (1) the Mark II translator; (2) the "Georgetown program"; and (3) the Rand Corporation's MIND system. A fourth system is also proposed in which material would be translated into a language so constructed that esch foreign word and affix could be replaced by a counterpart in an artificial language (a one-to-one correspondonce) which would be much easier to learn than the foreign language itself. Computers are now being used to study meaning through programs that mimic human behavior. For processing of textual data, it was thought that different sets of requirements would demand different programs and that it would be necessary to design essentially different algorithms for basic linguistic processes. It seems now that the best algorithms will be variants of a single overall strategy. Three strategies have been proposed for obtaining deep structures for arbitrary sentences. Besides the problems of syntactic analysis, there are many problems in semantics, and the computational linguist is coming to see that it is in this field that his main contribution will be made.

Descriptors: MACHINE TRANSLATION: SEMANTICS: SYNTHETIC LANGUAGES: DATA PROCESSING AND RETRIEVAL

Identifiers: machine translation of natural languages: problems of meaning:

# CONCORDANCE PROGRAMS

## SECTION 4

A Short Concordance to Laurence Sterne's 'A Sentimental Journey Through France and Italy by Mr. Yorick.' Volume I. A-L

Illinois Univ., Urbana. Dept. of Computer Science.*Princeton Univ., N.J. Dept. of Statistics.*National Science Foundation, Washington, D.C. Div. of Computer Research. (176 011)
AUTHOR: Pasta, Betty B., Pasta, David J., Pasta, John R.
C3793I4 FLD: 5B, 88E USGRDR7426
Sep 74 227p
REPT NO: UIUCDCS-R-74-676-Vol-1
MONITOR: 18
See also Volume 2, PB-236 233. Prepared in cooperation with Princeton Univ., N.J. Dept. of Statistics and National Science Foundation, Washington, D.C. Div. of Computer Research.

ABSTRACT: The concordance to Laurence Sterne's last work, A Sentimental Journey through France and Italy by Mr. Yorck, employs a KWIC (Keyword-in-Context) form which centers the word on the page and includes the words of text immediately preceding and following. The keyword types are in alphabetic order listed with each token given in order of appearance in the text. In the listing, special symbols precede the alphabet and numerals follow the alphabet. A word-frequency list containing all the words in the Journey is included. Some high frequency function words were blocked in the concordance, and this reduced its size from 40,635 to 26,188 lines. Blocked words include certain articles, personal pronouns, parts of verbs to be and to have, and prepositions in, of, and to.

DESCRIPTORS: *Coordinate indexing, *Books, *Indexes(Documentation), Data processing, Computational linguistics, Information retrieval, Words(Language), Literature(Fine arts), English language

IDENTIFIERS: *Concordances, Permuted indexes, NTISIUU, NTISNSF

PB-236 232/5SL NTIS Prices: PC$7.50/MF$2.25

AUTOMATIC LINGUISTIC CLASSIFICATION

Linguistics Research Center, Univ. of Texas, Austin. (208 250)
AUTHOR: Pendergraft, Eugene D. , Dale, Nell,
0313F3 FLD: 5G USGRDR6603
Nov 65 46p
REPT NO: LRC-65-WAT-1
CONTRACT: DA-36-039-AMC-02162(E)
GRANT: NSF-GN-308
Distribution: No limitation.

ABSTRACT: The work plan of a long-range series of experiments in automatic linguistic classification is described, together with discussion of a first experiment. The latter is concerned with category identification. In particular, the data resulting from automatic syntactic analysis of English were used to identify syntactical categories which have similar membership. The series of experiments will combine the use of automatic linguistic analysis and automatic classification techniques. Automatic syntactic analysis, and in later experiments semantic analysis, will be performed within the Linguistics Research System (LRS). Automatic classification will be carried out within the Automatic Classification System (ACS). A programming interface is being constructed between the two systems so that their combined capabilities can be used for automatic linguistic classification and partial selforganization.

DESCRIPTORS: (*Linguistics, Classification), (*English language, Classification), Automatic, Syntax, Computational linguistics

A Short Concordance to Laurence Sterne's 'A Sentimental Journey Through France and Italy by Mr. Yorick'. Volume II. M-Z

Illinois Univ., Urbana. Dept. of Computer Science.*Princeton Univ., N.J. Dept. of Statistics.*National Science Foundation, Washington, D.C. Div. of Computer Research. (176 011)
AUTHOR: Pasta, Betty B., Pasta, David J., Pasta, John R.
C3793J1 FLD: 5B, 88E USGRDR7426
Sep 74 248p
REPT NO: UIUCDCS-R-74-676-Vol-2
MONITOR: 18
See also Volume 1, PB-236 232. Prepared in cooperation with Princeton Univ., N.J. Dept. of Statistics and National Science Foundation, Washington, D.C. Div. of Computer Research.

ABSTRACT: The short concordance to Laurence Sterne's A Sentimental Journey Through France and Italy by Mr. Yorick contains 26,188 words of the 40,635 word text. Blocked words include certain articles, personal pronouns, parts of the verbs to be and to have, and the prepositions in, of, and to. The text was divided into logical episodes, and each word was tagged with the number of the episode in which it appears.

DESCRIPTORS: *Coordinate indexing, *Books, *Indexes(Documentation), Data processing, Computational linguistics, Information retrieval, Words(Language), Literature(Fine arts), English language

IDENTIFIERS: *Concordances, Permuted indexes, NTISIUU, NTISNSF

PB-236 233/3SL NTIS Prices: PC$7.50/MF$2.25

A Computer-Aided Investigation of Linguistics Performance: Normal and Pathological Language

Iowa Univ Iowa City Dept of Mathematics (404511)

Technical rept.
AUTHOR: Wachal, Robert S., Spreen, Otfried
A1205A1 FLD: 5G, 56J USGRDR7101
Jul 70 22p
REPT NO: THEMIS-UI-TR-29
CONTRACT: N00014-68-A-0500
Report on the Theory and Applications of Automaton Theory.

ABSTRACT: A system of twenty FORTRAN and PL/1 programs, developed for an analysis of aphasic and normal speech transcripts, is described in detail. The programs aid in lexical, grammatical, paralinguistic, and statistical analyses as well as in data preparation and correction. They can also be used in schizophrenic and other kinds of pathological language and are adaptable to the analysis of written-language samples and the investigation of authorship and style. (Author)

DESCRIPTORS: (*Speech, *Computational linguistics), Performance(Human), Pathology, Computers, Psychiatry

IDENTIFIERS: PL/1 programming language, FORTRAN, Psycholinguistics, Themis project

AD-714 144 NTIS Prices: PC$3.00 MF$0.95

poken Language Vocabulary Structural Frequency Count - Japanese Data
nalyses

yracuse Univ Research Corp N Y (339750)

pecial rept. 1 Jul 72-30 Jun 73
UTHOR: Sukle, Robert J., Miron, Murray S., Pratt, Charles C.
259211 FLD: 5G, 92D USGRDR7410
un 73 465p
EPT NO: SURC-TR-73-228
ONTRACT: DAAG05-72-C-0574
ONITOR: 18

BSTRACT: The report is a frequency analysis of vocabulary and entence patterns in the Japanese language. The corpora used are a edia sample, a discussion session, elicited sentences, and words licited for frame sentences. The outputs are the following frequency ables: (a) semantic frequency of combined corpus (media, discussion, licited sentences) listed alphabetically with inflectional and erivational variants as subentries; (b) semantic frequency of ombined corpus listed by frequency; (c) sentence pattern frequency rom corpus of elicited sentences; (d) H-ranks and phi-coefficients or corpus of elicited words. (Author)

ESCRIPTORS: *Words(Language), *Vocabulary, Counting, Computational inguistics, Semantics, Speech

DENTIFIERS: *Japanese language, *Word frequency, Etymology, SD

D-775 925/1 NTIS Prices: PC$26.25/MF$1.45

oken Language Vocabulary and Structural Frequency Count: English
ta Analyses

racuse Univ Research Corp N Y (339750)

ecial rept. 1 Jul 72-30 Mar 73
THOR: Miron, Murray S.
2592H4 FLD: 5G, 92D USGRDR7410
r 73 322p
PT NO: SURC-TR-73-117
NTRACT: DAAG05-72-C-0574
NITOR: 18

BSTRACT: The report is a frequency analysis of vocabulary and ntence patterns in the English language. The corpora used are a dia sample, a discussion session, elicited sentences, and words icited for frame sentences. The outputs are the following frequency bles: (a) Semantic frequency of combined corpus (media, discussion, licitied sentences) listed alphabetically with inflectional and rivational variants as sub-entries; (b) semantic frequency of mbined corpus listed by frequency; (c) sentence pattern frequency om corpus of elicited sentences; (d) H-ranks and phi-coefficients r corpus of elicited words. (Author)

ESCRIPTORS: *Words(Language), *Vocabulary, Frequency, Computational inguistics, Speech, English language

DENTIFIERS: *Word frequency, Etymology, SD

D-775 924/4 NTIS Prices: PC$19.25/MF$1.45

ED132568 CS203088
Degrees of Syntactic and Rhetorical Fluency-Competency in Freshman Writing: A Computer-Assisted Study.
Chisholm, William
77 7p.; Paper presented at the Annual Meeting of the Midwest Modern Language Association (18th, St. Louis, Missouri, November 4-6, 1976)
EDRS Price MF-$0.83 HC-$1.67 Plus Postage.

An exploratory study of quantitative measurement of syntactic and rhetorical fluency examined students' writing near the beginning and near the end of a two-quarter, freshman English program. The syntactic analysis focused on the clause, which was classified according to basic syntactic type and elaborating syntactic structures. The rhetorical analysis concentrated on the orthographic unit and included counts of selected rhetorical features and counts of logical relationships between successive units of thought. Preliminary results are reported, though in general the measures chosen did not discriminate between the 20 compositions written at the beginning of the program and the 20 written at the end. (AA)UI

Descriptors: College Freshmen/ *Composition Skills (Literary)/ Higher Education/ Language Fluency/ *Language Patterns/ Language Research/ *Rhetoric/ *Syntax

7502755 7502755
A literary analysis by computer
Waltman, Franklin M.
Foreign Languages State U New York Coll Cortland 13045
Hispania- 1974, 57, 4, Dec, 893-898. CODEN: hisn-b

7304688 7304688
A computer-assisted study of the vocabulary of young Navajo children
Spolsky, Bernard; Holm, Wayne; Holliday, Babette; Embry, Jonathan
Linguistics, U. New Mexico
Computers and the Humanities- 1973, 7 (4), 209-218. CODEN: cohu-a

7935243 79-3-000653
Semi-Automatic Construction of Semantic Concordances
Fraenkel, A. S.; Raab, D.; Spitz, E.
Computers and the Humanities, US ISSN 0010-4817, Flushing, NY, 1979, 13:283-88

ED108633 IR002150
Design Document: KWIC Module: L.A.P. Version I.
Porch, Ann
Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, Calif.
26 May 72 9p.
Report No.: SWRL-TN-5-72-37
EDRS Price MF-$0.76 HC-$1.58 PLUS POSTAGE

The Language Analysis Package (LAP) was developed by the Southwest Regional Laboratory (SWRL) to assist researchers in the analysis of language usage. The function of the KWIC (Keyword-in Context or Concordance) Module of the LAP is to produce keyword listings from the input text being analyzed. Such listings will contain location information broken ( in by document identifier, page, paragraph, and line. Other design features are presented in this document together with the file

Spoken Language Vocabulary and Structural Frequency Count - Swahili Data Analyses

Syracuse Univ Research Corp N Y (339750)

Special rept. 1 Jul 72-30 Jun 73
AUTHOR: Rubama, Ibrahim, Miron, Murray S., Pratt, Charles C.
C2592I2 FLD: 5G, 92D USGRDR7410
Jun 73 301p
REPT NO: SURC-TR-73-229
CONTRACT: DAAG05-72-C-0574
MONITOR: 18

ABSTRACT: The report is a frequency analysis of vocabulary and sentence patterns in the Swahili language. The corpora used are a media sample, a discussion session, elicited sentences, and words elicited for frame sentences. The outputs are the following frequency tables: (a) semantic frequency of combined corpus (media, discussion, elicited sentences) listed alphabetically with inflectional and derivational variants as subentries; (b) semantic frequency of combined corpus listed by frequency; (c) sentence pattern frequency from corpus of elicited sentences; (d) H-ranks and phi-coefficients for corpus of elicited words. (Author)

DESCRIPTORS: *Words(Language), *Vocabulary, Counting, Computational linguistics, Speech

IDENTIFIERS: *Swahili, African languages, Word frequency, SD

AD-775 926/9 NTIS Prices: PC$18.25/MF$1.45

anual for the Development of Language Frequency Counts

yracuse Univ Research Corp N Y (339750)

pecial rept. 1 Jul 72-30 Jun 73
UTHOR: Miron, Murray S., Pratt, Charles C.
2592H3 FLD: 5G, 92D USGRDR7410
un 73 58p
EPT NO: SURC-TR-73-235
ONTRACT: DAAG05-72-C-0574
ONITOR: 18

BSTRACT: As part of a continuing project of language analysis, SURC
resents its final manual. This manual is an explanation of the
rocedures used to collect and analyse data for this project. After
kplaining the theory and application of the methodology, the manual
iscusses specific problems encountered in the design, administration
id analysis of the language data collected. (Modified author
ostract)

ESCRIPTORS: *Vocabulary, *Words(Language), Computational linguistics,
emantics, Manuals

ENTIFIERS: Word frequency, Etymology, SD

-775 923/6 NTIS Prices: PC$6.00/MF$1.45

04072 7804072
Trends in Computer Applications to Literature
Widmann, R. L.
Computers and the Humanities- 1975. 9. 5. Sept. 231-235.

User's Guide to the SOLAR Semantic Analysis File

System Development Corp Santa Monica Calif*Advanced Research Projects Agency, Arlington, Va. (339900)

Technical rept.
AUTHOR: Pye, Tom, Diller, Timothy, Olney, John
C4643K4 FLD: 5G, 9B, 92D, 62B* USGRDR7513
31 Apr 75 39p
REPT NO: SDC-TM-5292/001/00
CONTRACT: DAHC15-73-C-0080, ARPA Order-2254
MONITOR: 18

ABSTRACT: The document contains a general explanation of the semantic analysis file of SOLAR (a Semantically-Oriented Lexical Archive). It is intended as an introduction and reference manual for the on-line user, the casual reader, or the data collector. The document indicates the design concepts, the resulting file structure, the intended file content, retrieval procedures, and data collection procedures.

DESCRIPTORS: *Semantics, *Speech recognition, English language, Information retrieval, Data processing, Computational linguistics, Natural language, Manuals

IDENTIFIERS: NTISDODA

AD-A009 328/6ST NTIS Prices: PC$3.75/MF$2.25

Phrase Dictionary Distribution Analysis and Growth Prediction Report

Cryptanalytic Computer Sciences Inc Cherry Hill N J (406482)

Final rept. 26 Jan-26 Apr 74
AUTHOR: Waite, J. H., Boehm, R., Fisher, J. G., Epstein, S. D., Stewart, D. J.
C3114K4 FLD: 5G, 5B, 92D, 88B USGRDR7417
26 Apr 74 56p
CONTRACT: DAAA21-74-C-0269
MONITOR: 18

ABSTRACT: The report describes a study of the DDC Phrase Glossary. It includes a computer program to tabulate word frequencies for blocks of phrases of optional sizes. On the basis of these distributions, empirical and statistical analyses are made including two prediction models. Two-word distributions are also included. Based upon the available distributions, a two-word Phrase Glossary size of 320,000 two-word phrases was determined. Also included are analyses of various techniques, such as suffix truncation, imbedded phrases, and query effectiveness. Comparisons are made of the DDC system to other plain language machine retrieval systems. (Author)

DESCRIPTORS: *Information retrieval, *Dictionaries, Words(Language), Occurrence, Models, Predictions, Computational linguistics, Computer applications

IDENTIFIERS: Phrase structure, NTISDODA

AD-780 957/7 NTIS Prices: PC$3.75/MF$1.45

AUTOMATIC LINGUISTIC ANALYSIS

OUTSIDE THE U.S.A.

SECTION 5

The SQAP Data Base for Natural Language Information

Research Inst. of National Defense, Stockholm (Sweden). (402 800)
AUTHOR: Palme, Jacob
C5112J2 FLD: 05G, 92D USGRDR7520
Jul 75 79p
REPT NO: FOA-P-C8376-ME(E5)
MONITOR: 18

ABSTRACT: The Swedish Question Answering Project (SQAP) aims at handling many different kinds of facts, and not only facts in a small special application area. The SQAP data base consists of a network of nodes corresponding to objects, properties and events in the real world. Deduction can be performed, and deduction rules can be input in natural language and stored in the data base. This report describes the data base, specially focusing on problems in its design, both problems which have been solved and problems which are not yet solved. Specially full treatment is given to the data base representation of natural language noun phrases, and to the representation of deduction rules in the data base in the form of data base patterns.

DESCRIPTORS: *Computational linguistics, Computer programming, Artificial intelligence, Semantics, Words(Language), English language, Sweden

IDENTIFIEPS: Swedish question answering project, NTISSWRIND

PB-243 783/8ST NTIS Prices: PC$4.75/MF$2.25

7704731 7704731
Automatische Lemmatisierung -- Zielsetzung und Arbeitsweise eines linguistischen Identifikationsverfahrens(Automatic Lemmatization -- Goals and Procedures of a Linguistic Identificational Program)
Weber, Heinz Josef
U Saarlandes, 6600 Saarbrucken Federal Republic of Germany
Linguistische Berichte- 1976, 44, Aug, 30-47. CODEN: lgbr-a
Friedrich Vieweg & Sohn, P. O. Box 5829, D-6200 Wiesbaden, Federal Republic of Germany
Section Heading Codes: 4610 LANGUAGE: Ger

The goals of this project are identifying & specifying word forms within a text by means of a large dictionary (about 100.000 stems with syntactic & semantic specifications) & a grammatical component. Word forms within u text are to be specified with regard to their lexical codification & linguistic context. The procedure has 5 steps: (1) analysis of inflectional variants & retrieval of stems -- in case of lexical ambiguity, detection of the various readings as offered by the dictionary, (2) detection of discontinuous verb constituents -- a special problem of German (e.g., er ging vor vielen Jahren in der Fremde verloren. . . (he was lost abroad for many years)) -- & reconstruction of the compound stem (e.g., verlorengehen (to be lost)), (3) disambiguation of syntactic homographs (e.g., English "leaves" -- verb/noun or German billige (just/equitabl-) verb/adjective) by distributional analysis, (4) idwntification of idiomatic expressions consisting of several verbel units (e.g., English "to kick the bucket" or German die Kurve kratzen) -- in this case a special dictionary component is used, & (5) disambiguation of semantic homographs by means of selectional restrictions in connection with a rough specification of the syntactic structure of a sentence. AA

Descriptors: COMPUTATIONAL LINGUISTICS; DATA PROCESSING AND RETRIEVAL; GERMAN; DICTIONARY; AMBIGUITY; DISCOURSE ANALYSIS
Identifiers: automatic lemmatization of German word forms;

7603928 7603928
Toward a Generative Dependency Grammar
Vater, Heinz
U Cologne, Federal Republic of Germany
Lingua- 1975, 36, 2-3, Jun, 121-145. CODEN: ling-a
North Holland Publishing Company, P. O. Box 211, Amsterdam, The Netherlands
Section Heading Codes: 050

The notion of valence & the relation of dependency connected with it were introduced into the theory of grammar by Tesniere. Later, D. Hays ("Dependency Theory: A Formalism and Some Observations" Language 1964, 40, 511-525.), Gaifman, & K. Baumgartner ("Konstituenz und Dependenz" (Constituents and Dependence) in Steger, H. (editor) Vorschlage fur eine strukturale Grammatik des Deutschen (Project for a Structural Grammar of German) Darmstadt: Wissenschafliche Buchgesellschaft, 1970.) showed that dependency & constituent grammar are not only complementary but (at least weakly) equivalent. Robinson worked out a model of a generative grammar with a deep structure built on dependency relations rather than on phrase structure relations (A Dependency Based Transformational Grammar (Research Report RC-1889) Yorktown Heights, NY: IBM Watson Res Ctr.). Robinson argues that the concept of head cannot be formalized within the framework of a phrase-structure categorial component, but that it can be formally specified for each phrase, if dependency rules gene ate the structural strings of categories, thus supplying additional information needed for some of the transformations.

In this paper, an attempt is made to overcome the shortcomings in Robinson's model by modifying her dependency rules & adding semantic specifications to the dependents of V, taking into account some of the considerations that led Fillmore to make up his "cases." HA

Descriptors: TRANSFORMATIONAL AND GENERATIVE GRAMMAR; TESNIE
Identifiers: theory of generative dependency grammar; valence; dependency vs. phrase structure grammar, Tesniere, Fillmore;

7890024 7890024
A Swedish Lexical Data Base
Allen, Sture; Ralph, Bo
Sprakdata Goteborgs U, Norra Allegatan o S-413 01 Sweden
Series: AILA 1978 0007

A lexical data base for present-day Swedish is in the process of being developed at the department of natural-language processing, U of Goteborg. The lexical material is drawn from authentic texts. Large samples of words with their contexts still traceable are available through the Swedish Logotheque, which maintains word & text banks in machine readable form. The linguistic analysis is carried out interactively, using an adapted form of case grammar. Linguistic information includes grammatical constructions; semantic definitions; morphotactic properties of the items; phonetic/phonological, graphonomic, stylistic, & statistical data; & a brief etymological note. The definitions contain words reducible to a minimal list of defining words. These defining units are regarded as indivisible primitives. A controlled defining vocabulary is used to avoid circularity in the definitions. This data base may have a number of uses. The sophisticated form of storage employed allows the material to be approached in several ways: the material can also be immediately restructured in the way the linguist chooses. The data base's most obvious use, however, is for dictionary production. The first thing generated from the data base will be an unconventional monolingual Swedish dictionary which will reflect the distinguishing features of the data base.

Descriptors: LEXICOLOGY; GERMANIC LANGUAGES; VOCABULARY; DICTIONARY
Identifiers: Swedish lexical data base;

75008184 v3n1
Technique for parsing ambiguous languages
Koster, C.H.
4th Annual Meeting of Society for Informatics B744204
Berlin, Ger (FR) 9-12 Oct 74
Society for Informatics
Papers (Eng or Ger) in "Lecture Notes in Computer Science," end 1974; approx. DM40; inquire: Springer Verlag, 175 Fifth Ave., New York, N. Y.
Descriptors: LANGUAGE
SECTION HEADING: GENERAL ENGINEERING AND TECHNOLOGY
Section Class Codes: 5000

7828995 78-3-003428
Dependency Grammar as Syntactic Model in Several Procedures
of Automatic Sentence Analysis
Kunze, Jurgen
Linguistics: An Interdisciplinary Journal of the Language
Sciences, Cambridge CB2 3EB, England. ·195(1977):49-62
Ooc Type: journal article
Descriptors: linguistics - linguistics, theoretical -
linguistics, descriptive - grammar - syntax; linguistics -
linguistics, general - linguistics, computational -
mechanolinguistics - Automated Analysis
Oescriptor Codes: 0303050004; 0302020003

43-12950 DOC YEAR: 1969 VOL NO: 43 ABSTRACT NO: 12950
FINOSIT: A computer program for language research.
Pylyshyn, Zenon W.
U. Western Ontario, London, Canada
Benavioral Science 1969, 14(3), 248-251.

7600433 7600433
COCOA: A Wordcount and Concordance Generator
Berry-Rogghe, G. L. M.
Instit deutschespr 12 Freidrich-Karlstr 6800 Mannheim 1,
Federal Republic of Germany
Association for Literary and Linguistic Computing Bulletin-
1973, 1, 2, Sum, 29-33. COOEN: allc-b

53-08536 OOC YEAR: 1975 VOL NO: 53 ABSTRACT NO: 08536
COCOA: A FORTRAN program for concordance and word-count
processing of natural language texts.
Corcoran, Paul E.
U Adelaide, South Australia
Behavior Research Methods & Instrumentation 1974 Nov Vol
6(6) 566

7405529 7405529
COCOA: A word count and concordance generator
BOOK AUTHOR: Berry-Rogghe, G. L. M., & Crawford, T. D.
Gamberini, Spartaco
U Coll Cardiff CF1 XL Wales United Kingdom
Language and Style~ 1974, 7, 2, Spr. 146-148. COOEN:
lgns-a
Series: REVIEW

7615335 76-3-000551
Observations on Context Free Parsing
Shail, B. A.
Statistical Methods in Linguistics, Stockholm, 1976,
71-109
Ooc Type: journal article
Oescriptors: linguistics - linguistics, general -
linguistics, computational - mathematical models
Oascriptor Codes: 0302020001

7827990 78-3-000651
A Partial-Parsing Algorithm for Natural Language Text Using
a Simple Grammar for Arguments
Sallis, Philip J.
Association for Literary and Linguistic Computing Bulletin,
PLACE UNKNOWN, 1978, 6:170-76
Doc Type: journal article
Descriptors: linguistics - linguis:ics, general -
linguistics, computational - mechanolinguistics
Oescriptor Codes: 0302020003

7600434 7600434
Publishing Computer Output of Processed Natural Language
Texts-I
Last, R. W.
German U of Hull, England
Association for Literary an Linguistic Computing Bulletin-
1973, 1, 3, Michaelmas, 5-7 .N: allc-b

7600448 7600448
PASP: Some Views on Automated Syntactical Parsing of Large Language Corpuses
Boot, M.
Rijksuniversiteit, Utrecht The Netherlands
ITL, Review of Applied Linguistics- 1974, 23, 23-38.
CODEN: itlg-a
Institute of Applied Linguistics, Blijde Inkomststr. 21, 3000 Louvain, Belgium
Section Heading Codes: 065

A discussion of some system analysis problems. The problems, never mathematically defined, concern the syntactic parsing of large language corpora not artificially restricted (PASP). Already developed strategies for PASP are discussed, & a more complete strategy is proposed. Major characteristics of this strategy are: (1) the ad hoc character of some parts of it; (2) use of a linear string grammar; (3) definition of probability rules; (4) translation of probability rules into a priori rules for string grammar; (5) context sensitivity; & (6) flexibility of the system. HA

Descriptors: COMPUTATIONAL LINGUISTICS; DESCRIPTIVE LINGUISTICS; THEORETICAL LINGUISTICS; SYNTAX; GRAMMATICAL ANALYSIS; DATA PROCESSING AND RETRIEVAL; MATHEMATICS; CONTEXT SENSITIVE GRAMMAR; TRANSFORMATION RULES

Identifiers: automated syntactical parsing in system analysis of PASP; linear string grammar, context sensitivity, probability rules;

49-11199 DOC YEAR: 1973 VOL NO: 49 ABSTRACT NO: 11199
Models for automatic translations.
Vauquois, Bernard
National Center for Scientific Research, Paris, France
Mathematiques et Sciences Humaines 1971 Sum Vol. 9(34) 61-70
LANGUAGE: Fren CLASSIFICATION: 11
Discusses the steps necessary to arrive at a model which can
be implemented on a computer. 3 existing models and their
characteristics are presented: (a) a model for morphological
analysis; (b) a model for syntactic analysis; and (c) a model.
actually operational, for higher level surface syntax.
SUBJECT TERMS: LINGUISTICS, COMPUTER APPLICATIONS.
MORPHOLOGY (LANGUAGE), SYNTAX; 28450, 10900, 32080, 51220
INDEX PHRASE: computer implementation, morphological &
syntactic & higher level surface syntax analyses models

7603975 7603975
Linguistic Data Processing and ALLC Activities in Germany
Lenders, W.
Instit Communication Theory Res & Phonetics U Bonn, 5300 Liebfrauenweg 3 Federal Repubiic of Germany
Association for Literary and Linguistic Computing Bulletin- 1974, 2, 1, 24-27. CODEN: allc-b
6 Sevenoaks Ave., Heaton Moor, Stockport, Cheshire SK4 4AW, England
Section Heading Codes: 060

(Presented at the Association for Literary and Linguistic Computing (ALLC) International Meeting, 1973.) Scientific research in the field of literary & linguistic data processing has been intensified in the last few years in Germany. Specialists in text-oriented data processing have met with specialists concerned primarily with the elaboration of new methods of text analysis. Various projects are being carried out at the universities of Saarbrucken, Marburg, Bonn, & at the Instit for German Language at Mannheim & Bonn. The projects concern natural language communication between man & computer, syntactic analysis, machine translation, statistics & stylistic analysis, automatic language cartography, automatic lexicography, morphology, syntactical analysis, new methods in stylistic & mathematical linguistics, new textual editing techniques, & computer translation. The ALLC has set up regional branches & improved information sharing among different projects. The Specialist Group for Medieval German Texts has also intensified its activities. D. Burkenroad

Descriptors: DATA PROCESSING AND RETRIEVAL; EXPERIMENTAL DATA HANDLING; SYNTAX; MACHINE TRANSLATION

Identifiers: linguistic data processing, Germany;

7602861 7602861
Computer Translation with Paired Grammars
Green, T. R. G.
Sheffield U, S10 2TN England
Behavior Research Methods and Instrumentation- 1975. 7. 6. Nov. 557-562. CODEN: brmi-a
The Psychonomic Society, 1108 W. 34th St., Austin TX 78705
Section Heading Codes: 065

In certain types of experiments, the S controls an on-line computer by giving commands in a simple source language -- possibly a subset of English or a high level computer language. The commands must then be decoded before they can be obeyed. In 1 method an ad hoc program is written for the specific purpose. An alternative is to write a general purpose translator to decode the source language into a more primitive target language. A suitable translator is described, driven principally by "paired" context-free grammars of the source & target languages but also able to accommodate context-sensitive rules. The technique used could be called paired-grammar translation. It is based on a context-free phrase-structure with a top-down, left-to-right parsing system. Backus-Naur form is used for the grammar notation. The target grammar is paired with the source grammar in such a way that every non-terminal symbol in the source grammar is associated with the same non-terminal symbol in the target which, by definition, is its translation. The method is simple; context-sensitivity is handled by special-purpose subroutines written as needed. With the programming medium, it is assumed that the language used has facilities for list processing, recursion, & representation of strings. If a language is not available, FORTRAN would be adequate. Using the translator has several advantages. It is obviously much easier to write an ad hoc recognizer for a very primitive language than for a subset of English. Also, for small languages it is very easy to write & check grammars: minor modifications are a trivial job, & the finished product is unlikely to contain hidden bugs. An example is given which takes into consideration the problem of translating a string of commands, some of them conditional, out of a language that uses nested conditionals & into a language that uses jumps to labels. Modified HA

Descriptors: COMPUTATIONAL LINGUISTICS; MACHINE TRANSLATION; CONTEXT FREE GRAMMAR; CONTEXT SENSITIVE GRAMMAR
Identifiers: computer translation with paired grammars; context-free phrase structure. Backus-Naur form notation;

7704196 7704196
The Use of the Computer in Linguistic and Literary Research
Pester, A. R.
The Polytechnic, Wolverhampton England WV1 1LY
Association for Literary and Linguistic Computing Bulletin-
1976. 4. 3. 245-250. CODEN: allc-b

7930714 79-3-000654
Knowledge-Based Parsing

Gershman, Anatole Vitali
Dissertation Abstracts International, Pt. A US ISSN 0419-4209, Pt. B US ISSN 0419-4217. Ann Arbor, MI. 1979. 40:2751B
Doc Type: journal article
Descriptors: linguistics - linguistics, general linguistics, computational - mechanolinguistics - Automated Analysis
Descriptor Codes: 0302020003

7704196 7704196
The Use of the Computer in Linguistic and Literary Research
Pester, A. R.
The Polytechnic, Wolverhampton England WVI ILY
Association for Literary and Linguistic Computing Bulletin-
1976, 4, 3, 245-250. CODEN: allc-b
6 Sevenoaks Ave., Heaton Moor, Stockport, Cheshire SK4 4AW, England
Section Heading Codes: 4110

Contributions to the Fourth International Symposium of the Association for Literary and Linguistic Computing (Oxford, England 5-9 April, 1976) are reviewed. Briefly described are the salient issues of each of the 43 papers given. These relate to current work in: authorship studies stylistics, cluster analysis, concordances, software, transliteration, syntactic analysis, text editing, thematic analysis, & photocomposition. The literary bases of the contributions range from early Greek & Hebraic texts to Braille, modern French poetry, & dialects of Upper Michigan. AA

Descriptors: APPLIED LINGUISTICS; COMPUTATIONAL LINGUISTICS; SYNTAX; ADOLESCENT LANGUAGE; READING AIDS FOR THE BLIND; FRENCH; POETRY; DIALECTOLOGY; STYLISTICS; STATISTICAL ANALYSIS OF STYLE; EXPERIMENTAL DATA HANDLING; RESEARCH DESIGN AND INSTRUMENTATION

Identifiers: computer use in linguistic/literary research;

ED036783 AL002062
Applied Computational Linguistics.
Hays, David G.
Sep 69 19p.; Paper delivered at the International Conferencu Congress of Applied Linguistics. Cambridge, England, September 1969
EDRS Price MF-$0.76 HC-$1.58 PLUS POSTAGE

Much work in computational linguistics, e.g. the preparation of concordances and text files, has dealt strictly with the surfac* of language, treating it as nothing more than strings of characters or phonemes. The "classical" scheme, developed as a result of dissatisfaction with the inability of such surface systems to deal with problems such as ambiguity, consists of surface processing, syntactic processing and semantic processing, with the object of obtaining an expression for the content of the input text: work with programming systems for generation of sentences with transformational grammar is representative of this tradition. It must be recognized, however, that the essential characteristic of language is its connection with information and that language is the external manifestation of the human capacity to process symbols in such ways that information is retained. This capacity should be the object of linguistics, and rules of grammar should describe those "action patterns" which underlie human symbol processing. Recent work in applied computational linguistics recognizes the importance of this conception and should therefore lead to wider computer applications, perhaps even to real man-machine conversations and the concomitant use of the computer as an imaginative consultant for a wide renge of problems. (FWB)

Descriptors: Analog Computers/ *Applied Linguistics/ *Communication (Thought Trensfer)/ *Computational Linguistics/ Computer Assisted Instruction/ *Computer Programs/ Digital Computers/ Information Retrieval/ *Information Storage/ Linguistics/ Machine Translation/ Surfece Structure

Identifiers: *Action Patterns

# AUTOMATIC INDEXING
# AND TEXT ANALYSES
## SECTION 6

Automatic Informative Abstracting and Extracting. Part I. Experiments in the Use of Syntactic Information in Automatic Extracting and Indexing

Lockheed Missiles and Space Co Inc Palo Alto Calif Palo Alto Research Lab (210118)

Final rept.
AUTHOR: Earl, Lois L.
C1174L1 FLD: 5B, 5G, 88B* USGRDR7315
May 73 199p*
REPT NO: LMSC-D350104
CONTRACT: N00014-70-C-0239
MONITOR: 18

ABSTRACT: The report summarizes a 9-year study of English morphology, phonetics, syntax, and semantics, and the experiments in automatic indexing and extracting completed. Five main topics are discussed: an algorithm for assigning parts of speech from morphology; an algorithm for automatic syntactic analysis; an experiment in construction of a 'structure dictionary' for extracting purposes; experiments in using frequency and/or syntactic criteria for indexing and extracting purposes; development of word government tables as the basis of a semantic component of an automated text analysis system.

DESCRIPTORS: (*Subject indexing, Automatic), (*Computational linguistics, Subject indexing), Abstracts, Data processing systems, Syntax, English language, Algorithms, Semantics, Phonetics

IDENTIFIERS: *Automatic extracting(Documentation), *Automatic indexing, N

AD-762 456 NTIS Prices: PC$6.00/MF$0.95

ED048911 LI002720
Automatic Content Analysis: Part I of Scientific Report No. ISR-18, Information Storage and Retrieval...
Cornell Univ., Ithaca, N.Y. Dept. of Computer Science.
Oct 70 169p.; Part of LI 002 719
Sponsoring Agency: National Library of Medicine (DHEW), Bethesda, Md.; National Science Foundation, Washington, D.C.
Report No.: ISR-18 Part I
EDRS Price MF-$0.76 HC-$8.24 PLUS POSTAGE

Four papers are included in Part One of the eighteenth report on Salton's Magical Automatic Retriever of Texts (SMART) project. The first paper: "Content Analysis in Information Retrieval" by S. F. Weiss presents the results of experiments aimed at determining the conditions under which content analysis improves retrieval results as well as the degree of improvement obtained. The second paper: "The 'Generality' Effect and the Retrieval Evaluation for Larger Collections" by G. Salton assesses the role of the generality effect in retrieval system evaluation and gives evaluation results for the comparisons of several document collections of distinct size and generality in the areas of documentation and aerodynamics. In the third paper: "Automatic Indexing Using Bibliographic Citations" by G. Salton citations are used directly to identify document content and an attempt is made to evaluate their effectiveness in a retrieval environment. The final paper: "Automatic Resolution of Ambiguities from Natural Language Text" by S. F. Weiss discusses the evolutionary process by which ambiguities are created and classifies ambiguities into three classes: true, contextual and syntactic. (For the entire SMART project report see LI 002 719, for parts 2-5 see LI 002 721 through LI 002 724.) (NH)

Descriptors: *Automatic Indexing/ Automation/ Bibliographic Citations/ *Content Analysis/ Electronic Data Processing/ *Evaluation/ Indexing/ *Information Retrieval/ Lexicology/ Programing Languages/ *Relevance (Information Retrieval)/ Vocabulary

Identifiers: Automatic Content Analysis/ On Line Retrieval Systems/ *Saltons Magical Automatic Retriever of Texts/ SMART

ED084281 TM003289

On the Uses of the Computer for Content Analysis in Educational Research.

Hiller, Jack H.; And Others

Feb 73 21p.; Revised version of paper presented at national conference of Association for Computing Machinery (San Francisco, August 1969)

EDRS Price MF-$0.76 HC-$1.58 PLUS POSTAGE

Current efforts to take advantage of the special virtues of the computer as an aid in text analysis are described. Verbal constructs, category construction, and contingency analysis are discussed and illustrated. Mechanical techniques for reducing human labor when studying large quantities of verbal data have been sought at an increasing rate by researchers in the behavioral sciences. Whatever the purpose of research, if it is to have a scientific character, ti must involve an attempt to reduce natural language da a, by formal rules, to measures reflecting theoretically relevant properties of the text, its source, or its audience effects. At the present time, there is no one theory or method dominating the field of natural language analysis. Although much work is currently being expended to implement a finite set of rules on the computer, little has been accomplished that is directly useful to researchers in the social sciences. (Author/CK)

Descriptors: Audiovisual Aids/ Classification/ *Computer Programs/ *Content Analysis/ Educational Research/ *Measurement Instruments/ *Scoring/ Social Sciences/ *Structural Analysis/ Technical Reports

Automatic Indexing: A State-of-the-Art Report

National Bureau of Standards, Washington, D.C. Center for Computer Sciences and Technology.*National Science Foundation, Washington, D.C.
(400 468)
AUTHOR: Stevens, Mary Elizabeth
D2653G4 Fld: 5B, 88A, 86V GRAI7715
Feb 70 298p
Rept No: NBS-Mono-91
Monitor: 18
Sponsored in part by National Science Foundation, Washington, D.C. Revision of report dated 30 Mar 65. Library of Congress catalog card no. 65-60023.

Abstract: A state-of-the-art survey of automatic indexing systems and experiments has been conducted by the Research Information Center and Advisory Service on Information Processing, Information Technology Division, Institute for Applied Technology, National Bureau of Standards. Consideration is first given to indexes compiled by or with the aid of machines, including citation indexes. Automatic derivative indexing is exemplified by key-word-in-context (KWIC) and other word-in-context techniques. Advantages, disadvantages, and possibilities for modification and improvement are discussed. Experiments in automatic assignment indexing are summarized. Related research efforts in such areas as automatic classification and categorization, computer use of thesauri, statistical association techniques, and linguistic data processing are described. A major question is that of evaluation, particularly in view of evidence of human inter-indexer inconsistency. It is concluded that indexes based on words extracted from text are practical for many purposes today, and that automatic assignment indexing and classification experiments show promise for future progress.

Descriptors: *Automatic indexing, Indexes(Documentation), Computational linguistics, Machine translation, Subject index terms, Thesauri, Reviews

Identifiers: NTISCOMNBS, NTISNSFG

7703418 7703418
The Computer in Literary and Linguistic Research: Papers from a Cambridge Symposium
BOOK AUTHOR: Wisbey, R. A.
Colby, Benjamin
U California, Irvine 95664
American Anthropologist- 1974. 76. 2. Jun. 447. CODEN: amat-a
Series: REVIEW

Studies and Design Specifications for Computerized Measurement of Textual Comprehensibility

Applied Psychological Services Inc Wayne Pa (031800)

Final rept. Mar 75-Jun 76
AUTHORS: Siegel, Arthur I.; Williams, Allan R.; Lapinsky, Walter J.; Warms, Tom A.; Wolf, J. Jay
D301114 Fld: 5G, 9B, 5J, 92B, 92D GRAI7719
Oct 76 255p
Contract: F41609-75-C-0037
Project: 1121
Task: 04
Monitor: AFHRL-TR-76-77

Abstract: A previous report (AD-A001 537) defined a series of 14 novel measures for determining the comprehensibility of English text on the basis of current psycholinguistic and Structure-of-Intellect oriented concepts. That report not only suggested the potential usefulness of the measures, but also conjectured the feasibility of automating the calculation of these measures. The present report takes the next logical steps in implementing these measures for computer application. First, these measures are analytically defined and described. Then, selected measures are subjected to 'laboratory' experimental investigation using Air Force Manuals, Career Development Course materials, and USAF Technical Orders as sample texts. Results of these experiments are presented. An automatic calculation method is then developed for each of the 13 selected measures. The structure of the programming specifications is modular and is intended to calculate the measures for variable size blocks of texts. Flow charts and summary descriptions of the program attributes are also presented, together with explanations of run request syntax, sample measures calculations, and output formats. This report then constitutes a complete definition of the program suitable for future implementation on an automatic data processing system.

Descriptors: *Psycholinguistics, *Reading, *Intelligibility, *Information processing, Computer programming, Computational linguistics, English language, Instruction manuals, Courses(Education), Text processing, Comprehension, Measurement, Syntax, Semantics, Assessment, Computer programs, Flow charting

Identifiers: *Comprehensibility, Cognition, Structure of intellect theory, NTISDODXA

AD-A041 285/8ST NTIS Prices: PC A12/MF A01

ED038159# LI001909
Machine-Aided Indexing. Technical Progress Report for Period January 1967-June 1969.
Klingbiel, Paul H.
Defense Documentation Center for Scientific and Technical Information, Alexandria, Va.
Jun 69 28p.
Report No.: DDC-TR-69-1
Available from: Clearinghouse for Federal Scientific & Technical Information, Springfield, Va. 22151 (AD-696 200. MF $.65, HC $3.00)
Document Not Available from EDRS.

Working toward the goal of an automatic indexing system which is truly competitive with human indexing in cost, time and comprehensiveness the Machine-Aided Indexing (MAI) process was developed at the Defense Documentation Center (DDC). This indexing process uses linguistic techniques but does not require complete syntactic analysis of sentences by the computer. The individual words are read into the computer and are either held for further consideration or eliminated. Lexical items (comma, periods and special symbols) are recognized. The output is a list of candidate index terms and a screened exception list of terms and phrases for human review. Eventually the list of candidate terms will enter an Integrated Language Data Base which is capable of posting terms directly to the data base, switching synonyms to postable terms or listing unrecognized terms for technical consideration. The step-by-step indexing procedure follows an overview of the entire process. (NH)

Descriptors: *Automation/ Computer Programs/ *Electronic Data Processing/ *Indexing/ *Information Retrieval/ *Program Design/ Program Development
Identifiers: *Machine Aided Indexing/ MAI

DOCUMENT RETRIEVAL THEORY, RELEVANCE, AND THE METHODOLOGY OF EVALUATION. REPORT NO. 3. MICROCATEGORIZATION FOR TEXT-PROCESSING

Lehigh Univ., Bethlehem, Pa. Center for the Information Sciences. (077 370)
AUTHOR: Reed, David M., Hillman, Donald J.
0585H2 FLD: 5B, 9B, 5G USGRDR4120
7 Jul 66 44p
GRANT: NSF-GN-451
MONITOR: 18
See also PB-170 970.

ABSTRACT: A computational approach to syntactic analysis is developed to meet the demands of the specific automatic indexing scheme described in PB's 170 969 and 170 970. A programmed analyzer is presented which employs a limited dictionary look-up procedure and a context-sensitive computational grammar. The dictionary contains less than three hundred functor word and suffix entries. The heuristically developed grammar is written in LECOM, a programming language similar to COMIT. The analyzer assigns categories to all words in an input text and identifies nominal, prepositional and infinitive phrases. Relative pronouns and the pronoun 'it' are replaced by antecedents. It is shown that this computational approach to syntactic analysis is economically feasible for automatic indexing systems which require minimal syntactic analysis and can tolerate minor errors. The economy of the system results from its limited dictionary, relatively small number of computational rules and restriction to technical English. (Author)

DESCRIPTORS: (*Information retrieval, Subject indexing), (*Computational linguistics, Information retrieval), Linguistics, Programming languages, Programming(Computers), Documentation

IDENTIFIERS: LECOM

CFSTI Prices: PC$6.00 MF$0.50

ED027915# LI000736
Semantic Tools in Information Retrieval.
Rubinoff. Morris: Stone. Don C.
Pennsylvania Univ.. Philadelphia. Moore School of Electrical
Engineering.
May 67 21p.
Sponsoring Agency: Air Force Office of Scientific Research.
Washington. D.C.: Army Research Office. Durham. N.C.
Contract No.: AF-49-638-1421
Available from: Clearinghouse for Federal Scientific and
Technical Information. Springfield. Virginia 22151 (AD 660
087. MF-$0.65. HC-$3.00).
Document Not Available from EDRS.
This report discusses the problem of the meansings of words
\used in information retrieval systems. and shows how semantic
tools can aid in the communication which takes place between
indexers and searchers via index terms. After treating the
differing use of semantic tools in different types of systems.
two tools (classification tables and semantic expansions) are
investigated in some detail. Finally. experiments now in
progress are described which involve statistical techniques
for semi-automatic generation of a vocabulary and a set of
classification tables for an area of specialization. Thes
techniques enable the construction or updating of semantic
aids with far less intellectual effort than now required. but
still retain a consensus of expert opinion through the
literature produced by experts. (Author/JB)
Descriptors: Automation/ Classification/ Computers/
Concordances/ Correlation/ •Indexing/ Information Retrieval/
•Information Systems/ •Semantics/ Sentences/ Thesauri/
•Vocabulary/ Word Lists

Evaluation of Automated Natural Language Processing in the Further
Development of Science Information Retrieval

New York Univ., N.Y. Linguistic String Project.*National Science
Foundation, Washington, D.C. Div. of Science Information.

Final rept. 1 Aug 73-31 Jan 76
AUTHOR: Sager, Naomi
D2792F2 Fld: 5G, 5B, 92D, 88A GRAI7716
Jul 76 113p
Rept No: String Program-10
Grant: NSF-GN-33879
Monitor: 18.

Abstract: The report describes advances in computerized natural
language processing (NLP) and relates them to present and potential
functions of information systems. Section 1 summarizes developments in
the information field which have led to a renewed interest in NLP, and
sketches how NLP programs could be used to provide new information
services operating on natural language data bases. It describes the
basis for such programs in the inherent relation between information
and language structure. Section 2 describes the stages of processing
which take largely unrestricted natural language input of the type
encountered in scientific communications into data structures suitable
for advanced types of information processing. Section 3 describes a
newly developed clustering program for generating informationally
significant word classes from documents in particular subject areas.
Section 4 presents some examples and suggestions as to how NLP
techniques currently available or under development could be applied
in information systems. Section 5 suggests directions for further
research in NLP as a foundation for natural-language-based information
systems in the future.

Descriptors: *Computational linguistics, *Information retrieval,
Semantics, Syntax, Automatic language processing, Data processing,
Technical writing, Transformational grammars, Clustering

Identifiers: Natural language, NTISNSFSIS

7604366 7604366

Carlyle and the Machine: A Quantitative Analysis of Syntax in Prose Style

Oakman, Robert L.

U South Carolina, Columbia 29208

Association for Literary and Linguistic Computing Bulletin-1975, 3, 2, Sum, 100-114. CODEN, alic-b

6 Sevenoaks Ave., Heaton Moor, Stockport, Cheshire SK4 4AW, England

Section Heading Codes: 080

An analysis of a large selection of Carlyle's prose was done by means of a linguistic & quantitative method of syntactic analysis & a computerized parsing procedure. The study had 2 objectives to identify stylistically significant elements of Carlyle's syntax & to determine the profitability of large-scale automatic syntactic analysis in describing prose style. The initial syntactic analysis was performed by a computerized parsing routine developed by D. C. Clarke & R. E. Wall. Masses of quantitative information about syntactic features were analyzed with statistical methods of comparison & correlation. These quantitative stylistic features were discussed in conjunction with close critical analysis of specified passages. The stylistic habits known to be peculiarly Carlylean -- periodicity, accumulation, & irregularity -- were all revealed by the study. A growing tendency to omit important syntactic elements or to introduce irregularities into standard syntax was noted in the chronological development of his style. Carlyle stretched the capacities of English syntax to fit his own needs. This is the broadest-based study of its kind so far attempted, & the stylistic features discovered apply more generally than earlier impressionistic studies based on smaller more carefully selected passages. S. Karganovic

Descriptors: STYLISTICS; STATISTICAL ANALYSIS DF STYLE; SYNTAX; LITERARY GENRES; DATA PROCESSING AND RETRIEVAL

Identifiers: quantitative computer analysis of syntax in prose style; Carlyle;

COMPUTER OUTPUTS FOR SENTENCE DECOMPOSITION OF SCIENTIFIC TEXTS

New York Univ., N. Y. Linguistic String Project.
AUTHOR: Bookchin, Beatrice
532101 FLD: 5G, 9B USGRDR6901
Mar 68 410p
REPT NO: String Program-3
GRANT: NSF-GN-659
See also PB-178 391.

ABSTRACT: This volume is the third in a series of detailed reports on a working computer program for string decomposition of sentences. This volume contains outputs obtained by the program for five short scientific texts. Each successive sentence of the text to be analyzed is entered into the computer without pre-editing. The program looks up each word of the sentence in a grammatical dictionary which gives for each word all its grammatical classifications without reference to the way the word is used in the given article. The program then decomposes each sentence into a very short elementary sentence which is the grammatical center of the original, plus various strings of words: each string has a fixed grammatical structure, and adjoins the elementary sentence or one of its adjoined strings. (Author)

DESCRIPTORS: (*Computational linguistics, Programming(Computers)), Dictionaries, Reports, Analysis, English language, Grammars

IDENTIFIERS: Strings(Linguistics), Parsing, Sentences, Computer analysis

PB-180 048 CFSTI Prices: PC$6.00 MF$0.95

ED051843# LI002903
Annual Report: Automatic Informative Abstracting and Extracting.
Earl, L. L.; And Others
Lockheed Missiles and Space Co., Palo Alto, Calif.
Mar 71 144p.
Sponsoring Agency: Office of Naval Research, Washington, D.C.
Report No.: M-21-71-1
Available from: National Technical Information Service, Springfield, Va. 22151 (AD-721 066, MF $.95, HC $3.00)
Document Not Available from EDRS.

The development of automatic indexing, abstracting, and extracting systams is investigated. Part I describes the development of tools for making syntactic and semantic distinctions of potential use in automatic indexing and extracting. One of these tools is a program for syntactic analysis (i.e., parsing) of English, the other is a dictionary of English word government patterns. Part II reports on the research program in describing and abstracting pictorial structures. This work is concerned with whether it is possible to construct a symbolic representation of a gray level picture which can provide essentially the same information as the picture itself. Based on a series of experiments using human subjects describing aerial terrain photographs, it was possible to make certain observations concerning deductive and metadescriptive aspects of description, i.e., the "set," contextual knowledge, and certainty of the subject. (Author/NH)

Descriptors: *Abstracts/ *Automatic Indexing/ *Automation/ Documentation/ Experimental Programs/ *Information Processing/ *Information Systems/ Linguistics/ Syntax
Identifiers: *Automatic Abstracting

Development of Language Analysis Procedures With Application to Automatic Indexing

Ohio State Univ., Columbus. Computer and Information Science Research Center. (407 586)
AUTHOR: Young, Carol Elizabeth
C2321I2 FLD: 5G, 88A* USGRDR7406
Apr 73 310p*
REPT NO: OSU-CISRC-TR-73-2
GRANT: NSF-GN-534.1
MONITOR: 18

ABSTRACT: The paper presents (1) a theoretical framework within which relationships among words are defined and (2) algorithms which have been developed to identify these relationships. The algorithms which have been developed effect four processes: the assignment of each word to a grammatical class, the identification of phrases and of clauses, and the assignment of case grammar roles. These linguistic analysis procedures are to be used to construct graphical representations of sentences. The graphs are proposed as the basis of a generalized indexing system. Portions of this document are not fully legible.

DESCRIPTORS: *Automatic indexing, *Phrase structure grammars, *Computational linguistics, *Syntax, Words(Language), Semantics, Schematic diagrams, English language

IDENTIFIERS: NSFSIS

PB-227 088/2 NTIS Prices: PC$7.25/MF$1.45

ED110048 IR002327

An Analysis of Methods for Preparing a Large Natural Language Data Base.

Porch, Ann

Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, Calif.

16 Feb 71 29p.

Report No.: SWRL-TM-5-71-02

EDRS Price MF-$0.76 HC-$1.95 PLUS POSTAGE

Relative cost and effectiveness of techniques for preparing a computer compatible data base consisting of approximately one million words of natural language are outlined. Considered are dollar cost, ease of editing, and time consumption. Facility for insertion of identifying information within the text, and updating of a text by merging with another text are given special attention. It is concluded that Magnetic Tape Selectric Typewriter (MTST) and Telterm2 (a cathode ray rube terminal) are two highly effective methods of text preparation. The decision of which to use on a particular project would depend on available funds and possible peripheral uses for the equipment. Criteria for making such a decision are discussed. (Author)

Descriptors: Computers/ *Cost Effectiveness/ *Data Bases/ Data Processing/ Electronic Data Processing/ *Equipment/ *Information Processing/ Information Storage/ *Input Output Devices/ Man Machine Systems/ Office Machines/ On Line Systems / Optical Scanners/ Typewriting

Identifiers: Administrative Terminal System/ ATS/ Cathode Ray Tube Terminals/ CRT/ Dataplex/ Flexowriter/ Keypunches/ Magnetic Tape Selectric Typewriter/ MTST/ Optical Character Scanning/ Teletypes

ED145829 IR005240

Evaluation of Automated Natural Language Processing in the Further Development of Science Information Retrieval. String Program Reports No. 10.

Sager, Naomi

New York Univ., N.Y. Linguistic String Project.

Jul 76 118p.

Sponsoring Agency: National Science Foundation, Washington, D.C. Div. of Science Information.

Grant No.: GN39879

EDRS Price MF-$0.83 HC-$6.01 Plus Postage.

This investigation matches the emerging techniques in computerized natural language processing against emerging needs for such techniques in the information field to evaluate and extend such techniques for future applications and to establish a basis and direction for further research toward these goals. An overview describes developments in the information field which have led to renewed interest in natural language processing, sketches of programs for processing natural language to fulfill language-based functions of information systems, and the relationship between information and language. The stages of processing unrestricted natural language input of scientific communication into data structures suitable for information procesaing--parsing, structural transformations of parse outputs, and arriving at an underlying semantically meaningful representation--are outlined. The report also describes research related to the computerized discovery of semantic structures in science subfields; this research is concerned with the problem of structuring a data base which is given in natural language. Examples and suggestions for the application of techniques currently available or under development to information problems, and suggestions for further research in the language area of information science are presented. (Author/KP)

Descriptors: Artificial Intelligence/ *Automatic Indexing/ *Computational Linguistics/ Evaluation/ Information Processing / *Information Retrieval/ Information Systems/ Language Classification/ Man Machine Systems/ *Science Materials/ *Semantics

Identifiers: *Natural Language Processing

# MISCELLANEOUS AUTOMATIC LANGUAGE PROCESSORS

## SECTION 7

### Research on Synonymy and Antonymy: A Model and Its Representation

Maryland Univ College Park Computer Science Center (403018)

Technical rept.
AUTHOR: Edmundson, H. P., Epstein, M. N.
A4631L2 FLD: 5G, 56J USGRDR7215
Mar 72 25p
REPT NO: TR-185
CONTRACT: N00014-67-A-0239-0004
PROJECT: NR-049-261

ABSTRACT: The paper describes a modified and extended version of an axiom system that constitutes a mathematical model of synonymy and antonymy. It also outlines the data structures used in the computer representation of the model. The intent of this research is to refine an axiomatic model previously proposed to better reflect the latent structure of synonym dictionaries and to influence their future compilation. Particular attention is given to providing a convenient computer representation for testing the current set of 13 axioms. The computer-based system provides an automated determination and verification of existing relations among dictionary entries and generates new relations among words to be included insuch a dictionary, as well as providing a measure of the binding power among related groups of words. (Author)

DESCRIPTORS: (*Semantics, Mathematical models), (*Computational linguistics, Semantics), Dictionaries, Data processing systems

IDENTIFIERS: Synonymy, Antonymy

AD-743 892 NTIS Prices: PC$3.00/MF$0.95

### PART-OF-SPEECH IMPLICATIONS OF AFFIXES

Lockheed Missiles and Space Co Palo Alto Calif (210110)
AUTHOR: Earl, Lois L.
3295L4 FLD: 5G USGRDR6711
4 Feb 66 7p
MONITOR: 18
Research supported in part by ONR.
Availability: Published in Mechanical Translation and Computational Linguistics v9 n2 p38-43 Jun 1966.

ABSTRACT: The paper describes a systematic investigation of the extent to which the part of speech of words can be identified from their prefixes and suffixes. The results indicate that it is possible to determine, with 95 per cent accuracy, the inclusive part of speech of an affixed word from a consideration of its prefixes, suffixes, and length. By 'inclusive' parts of speech we mean a string that will include all of the parts of speech assigned by both dictionaries considered but that may include one or two extraneous parts of speech. The extra parts of speech will differ according to the class of words, as adjectives may have an extra part-of-speech 'noun' or 'adverb,' while nouns may have an extra part-of-speech 'verb.' The part-of-speech implications of seventy-two prefixes and of eighty-seven suffixes are given. (Author)

DESCRIPTORS: (*English language, Computational linguistics), Grammars, Classification, Algorithms

7600347 7600347
String Transformations in the Request System
Plath, Warren J.
IBM Research Div, Yorktown Heights NY 10958
the Finite String- 1974. 11, 2. 8. CODEN: fnts-a
Center for Applied Linguistics, 1611 N. Kent St., Arlington
VA 22209 (Published as part of the American Journal of
Computational Linguistics as of The Finite String. 1974. Vol.
11, No. 1)
Section Heading Codes: 062
The Request System is an experimental natural language query
system based on a large transformational grammar of English.
In the original implementation of the system. the process of
computing the underlying structures of input queries involved
a sequence of 3 steps: (1) preprocessing (including dictionary
lookup): (2) surface phrase structure parsing: & (3)
transformational parsing. This scheme has since been modified
to permit transformational operations not only on the full
trees available after completion of surface parsing. but also
on the strings of lexical trees which are the output of the
preprocessing phase. Transformational rules of this latter
type. which are invoked prior to surface parsing, are known as
string transformations. Since they must be defined in the
absence of such structural markers as the location of clause
boundaries. string transformations are by necessity relatively
local in scope. Despite this inherent limitation. they have
so far proven to be an extremely useful & surprisingly
versatile addition to the Request System. Applications to
date have included homograph resolution. analysis of
classifier constructions, idiom handling. & the suppression of
large numbers of unwanted surface parses. While by no means a
panacea for transformational parsing. the use of string
transformations in Request has permitted relatively rapid &
painless extension of the English subset in a number of
important areas without corresponding adverse impact on the
size of the lexicon. the complexity of the surface grammar. &
the number of surface parses produced. HA
Descriptors: TRANSFORMATIONAL AND GENERATIVE GRAMMAR:
ENGLISH: EXPERIMENTAL DATA HANDLING: DEEP STRUCTURE AND
SURFACE STRUCTURE: TRANSFORMATION RULES: THEORETICAL
LINGUISTICS
Identifiers: string transformation in the Request System;
English;

User's Guide to the SOLAR KWIC File

System Development Corp Santa Monica Calif*Advanced Research Projects
Agency, Arlington, Va. (339900)

Special technical rept.
AUTHOR: Diller, Timothy C., Heath, Frank
C4873A4 FLD: 5G, 5B, 92D USGRDR7517
30 May 75 23p
REPT NO: TM-5292/008/00
CONTRACT: DAHC15-73-C-0080, ARPA Order-2254
MONITOR: 18

ABSTRACT: The document contains a general explanation of the KWIC file of SOLAR (a Semantically-Oriented Lexical Archive). It is intended as an introduction and reference manual for the on-line user, the casual reader, or the data collector.

DESCRIPTORS: *Semantics, *Words(Language), Speech recognition, English language, Information retrieval, Data processing, Indexes, Computational linguistics, Natural language, Manuals

IDENTIFIERS: KWIC indexes, SOLAR(Semantically Oriented Lexical Archive), Semantically oriented lexical archive, NTISDODA

AD-A011 178/1ST NTIS Prices: PC$3.25/MF$2.25

7403890 7403890
The annual meeting of the ACL
Moyne J. A.:
Queens Coll City U New York NY 10021
Computers and the Humanities- 1973, 7 (6), 413-415. CODEN: cohu-a
Queens College Press, Flushing NY 11367:
Section Heading Codes: 060

An outline report of the eleventh annual meeting of the Association for Computational Linguistics, held August 1 and 2, 1973 at the University of Michigan in Ann Arbor. Research on speech recognition and understanding continues to be a topic of major interest in computational linguistics (CL) around the country. Most of the speech projects are supported by ARPA and are intended to complement each other and run on the ARPA network. The traditional approach to speech recognition in the past was to rely on engineering developments and filtering devices for the segmentation of phonetic elements. The trend is toward more reliance on linguistic analysis and "understanding" of an utterance. Papers were presented which concern automatic parsing of Chinese, an automatic retrieval system with natural language communication, and a language developed for communication with computer by nonhuman primates. The four papers in the syntax session dealt with a computer model of Panini's grammar, semantic-directed translation of context-free languages, the testing of a grammar of English with no cycle, and a model of a "performance" grammar of English. The four papers in the lexical studies session were concerned with morphological, syntactic, and semantic analyses in lexicography and construction of dictionaries. One paper in this session reported the use of lexicostatistical devices for arriving at relationships among Indo-European languages. AA

Descriptors: COMPUTATIONAL LINGUISTICS: EXPERIMENTAL DATA HANDLING; DATA PROCESSING AND RETRIEVAL

Identifiers: computational linguistics: conference report; annual meeting of Association for Computational Linguistics:

The MIND System: A Data Structure for Semantic Information Processing

Rand Corp Santa Monica Calif (296600)
AUTHOR: Shapiro, Stuart Charles
A3314L3 FLD: 5G, 5B, 9B, 56J, 88B, 62B, 70C USGRDR7202
Aug 71 172p*
REPT NO: R-837-PR
CONTRACT: F44620-67-C-0045

ABSTRACT: A description is given of the data structure used in the semantic file of the MIND system (Management of Information through Natural Discourse), and of the procedures for manipulating information stored in the file. The MIND system consists of nested and chained modules of high-level programming language statements; it is relatively easy to modify, either for improvement or for adaptation to specialized applications. The major features of the data structure are: It is a net whose nodes represent conceptual entities and whose edges represent relations that hold between entities; Some nodes of the net are variables, and are used in constructing general statements and deduction rules; Each conceptual entity is represented by exactly one node in the net from which all information concerning that entity is retrievable; Nodes can be identified and retrieved either by name or by a sufficient description of their connections with other nodes. The use of the system to experiment with various semantic theories is demonstrated by examining several questions of current linguistic theory. (Author)

DESCRIPTORS: (*Semantics, *Data processing systems), (*Information retrieval, Command + control systems), Programming(Computers), Computational linguistics, Management planning

IDENTIFIERS: MIND(Management of Information through Natural Discourse), Management of information through natural discourse, Natural language, Management information systems

AD-733 560 NTIS Prices: PC$3.00 MF$0.95

7302203 7302203
Automatic syntactic analysis
BOOK AUTHOR: Foster, J. M.
Wood, Derick
Applied Mathematics, McMaster U.
International Journal of Computer Mathematics- 1972, 3
(2/3), 189-191. CODEN: ijcm-a
Series: REVIEW
New York: American Elsevier, 1970 for the United States,
Gordon & Breach Science Publishers, 440 Park Ave., S., New
York, N. Y. 10016; and for all other countries, Gordon &
Breach Science Publishers, 42 William IV St., London WC2
England;
Section Heading Codes: 060 LANGUAGE: Engl.
A favorable review of a work which is the first in formal
language theory to deal solely with automatic syntactic
analysis. It is an introductory text, not a work that covers
the area of syntactic analysis exhaustively. It is in this
light that a reader should approach this book. It is lucidly
written with many worked examples that make it a joy to read.
Also contributing to this enjoyment is its size (a mere 65
pages), which means that it can be read at one sitting.
Topics covered include: (1) context-free grammars; (2)
parsing; (3) universal parsing methods; (4) special parsing
methods; (5) transformations on grammars; and (6) using
grammatical analyses for compilation.
Descriptors: SYNTAX; DATA PROCESSING AND RETRIEVAL;
GRAMMATICAL ANALYSIS
Identifiers: automatic syntactic analysis; book review;

7302244 7302244
Syntactic analysis in R E L English
Dostert, Bozena Henisz; Thompson, Fredrick Burtis
California Inst. Technology
Statistical Methods in Linguistics- 1972, 8, 5-38. CODEN:
smln-a
Spra'kforlaget Skriptor, P. O. Box, 104 65 Stockholm 15,
Sweden;
Section Heading Codes: 060
A discussion of refinements of R E L (Rapidly Extensible
Language) English. A description of elements of the system
includes a transformational grammar, features, case structure,
inclusion of pronouns, and parsing. The incorporation of
Fillmore's case grammar is new as is the inclusion of
pronouns.
Descriptors: DATA PROCESSING AND RETRIEVAL; ENGLISH; CASE
GRAMMAR
Identifiers: development of R E L English; computer
language;

SYNTACTIC ANALYSIS OF ENGLISH BY COMPUTER: A SURVEY

Bolt, Beranek and Newman, Inc., Cambridge, Mass. (060 100)
AUTHOR: Bobrow, Daniel G.,
0305G1 FLD: 5G USGRDR6602
1964 23p
Distribution: No limitation.

ABSTRACT: The review begins with a survey of the determination of classes among English words. Most programs doing syntactic analysis of English use a dictionary lookup operation to find possible classifications of words and then resolve ambiguities during the parsing operation. A survey is also given of those theories of grammar which serve as a basis for syntactic processing by computer. The forms of the rules for each grammar and a description of the syntactic structure associated with a sentence by each processor are given; reference is made to computer programs which have been written, and goals and present success of these programs are reviewed.

DESCRIPTORS: (*English language, Computational linguistics), (*Syntax, Analysis), Semantics, State-of-the-art reviews, Grammars, Transformational grammars

IDENTIFIERS: Words, Sentences, Tree diagrams(Linguistics)

PB-168 548 CFSTI Prices: PC$6.00 MF$0.50

302199 7302199
A syntax-directed parser for recalcitrant grammars
Abrahams, Paul W.
Courant Inst. Mathematical Sciences, New York U.
International Journal of Computer Mathematics- 1972. 3
(2/3). 105-15. CODEN: ijcm-a
for the United States. Gordon & Breach Science Publishers.
440 Park Ave., S., New York, N. Y. 10016; and for all other
countries. Gordon & Breach Science Publishers. 42 William IV
St., London WC2 England:
Section Heading Codes: 060
A syntax-directed parsing scheme being used in a PL/I
compiler for the CDC 6600 is discussed. It uses a highly
restricted grammar of the class LL(1) for efficiency, with an
escape hatch for those cases excluded by the grammar. These
cases are handled by oracles that can make decisions without a
full-scale syntactic analysis. The input to SYNDIPAR, the
SYNtax DIrected PARser, consists of syntax equations, semantic
routines, and token class definitions; the output consists of
a PARSE procedure in PL/I together with certain tables. The
PARSE procedure works in conjunction with a lexical scanner,
designed to allow look-ahead by oracles in a uniform fashion.
The actual parsing process takes place through the
interpretation of a program compiled by SYNDIPAR for a parsing
machine. The instruction set of the parsing machine is
described, and an example of the compilation of a syntax
equation is given.
Descriptors: COMPUTATIONAL LINGUISTICS; SYNTAX; DATA
PROCESSING AND RETRIEVAL; GRAMMATICAL ANALYSIS
Identifiers: syntax-directed parser; recalcitrant grammars;

7502566 7502566
The lexical subclasses of the Linguistic String Parser
Fitzpatrick, Eileen; Sager, Naomi
New York U NY 10003
American Journal of Computational Linguistics- 1974, 1.
Microfiche 2. 1-70. CODEN: ajcl-d
Center for [illegible]ed Linguistics, 1611 N. Kent St., Arlington
VA 22209 (I[illegible]. The Finite String as of 1974; Vol. 11, No.
1)
Section Heading Codes: 063
The New York University Linguistic String Parser (LSP) is a
working system for the syntactic analysis of English
scientific texts. It consists of a parsing program, a
large-coverage English grammar, and a lexicon. The grammar's
effectiveness in parsing texts is due in large part to a
substantial body of detailed well-formedness restrictions
which eliminate most incorrect syntactic parses which would be
allowed by a weaker grammar. The restrictions mainly test for
compatible combinations of word subclasses. The 109
adjective, noun, and verb subclasses, as well as others not
presented here, are defined in such a way that they can be
used as a guide for classifying new entries to the LSP lexicon
and as a linguistic reference tool. Each definition includes
a statement of the intent of the subclass, a diagnostic frame,
sentence examples, and a word list drawn from the present
dictionary. The subclasses are defined to reflect precisely
the grammatical properties tested for by the restrictions of
the grammar. Where necessary for clarifying the intent of the
subclass, three additional criteria are employed: excision,
implicit and coreference, and paraphrase. The subclasses have
been defined so as to be consistent with a subsequent stage of
transformational analysis which is currently being
implemented. HA
Descriptors: ENGLISH; DATA PROCESSING AND RETRIEVAL; SYNTAX;
GRAMMATICAL ANALYSIS; SPECIAL LANGUAGES; TRANSFORMATIONAL AND
GENERATIVE GRAMMAR; COMPUTATIONAL LINGUISTICS
Identifiers: Linguistic String Parser syntactic analysis for
English scientific texts:

ED162663 IR006668
A Computer-Assisted Language Analysis System (CALAS) and Its Applications.
Pepinsky, Harold B.
78 16p.; For related document, see ED 090 948
EDRS Price MF-$0.83 HC-$1.67 Plus Postage.
Language: English
Geographic Source: U.S./ Ohio

A Computer-Assisted Language Analysis System (CALAS) was developed as a syntactic and semantic analyzer of machine readable text in English. CALAS includes a set of computer programs, an algorithm for implementation, and human editors who assist the computer and its programmer in the processing of data. Data analysis is accomplished in three stages: (1) syntactic analysis of text, identifying each work in sequence in terms of its grammatical equivalent; (2) aggregation of the individual words into phrases identified in terms of their grammatical equivalents; and (3) aggregation of phrases into clauses, with component phrases identified in terms of the roles each plays and exhibited to display a main or independent clause. Discussion of the literature focuses on the relative frequencies with which the different types of verb phrases are used, and the measure of structural or stylistic complexity. (JEG)

Descriptors: Case (Grammar)/ Componential Analysis/ *Computational Linguistics/ Computer Programs/ Data Analysis/ Data Processing/ *Discourse Analysis/ Error Analysis (Language)/ Language Patterns/ *Linguistic Patterns/ Sentence Diagraming/ *Sentence Structure/ Speech Communication
Identifiers: *Computer Assisted Language Analysis System

ED024930 AL001582
The Multistore System: MP-2
von Glasersfeld, Ernst; Pisani, Pier Paolo
Georgia Inst. for Research, Athens.
Nov 68 72p.
EDRS Price MF-$0.76 HC-$3.32 PLUS POSTAGE

The second version of the Multistore Sentence Analysis System, implemented on an IBM 360/65, uses a correlational grammar to parse English sentences and displays the parsings as hierarchical syntactic structures comparable to tree diagrams. Since correlational syntax comprises much that is usually considered semantic information, the system demonstrates ways and means of resolving certain types of sentence ambiguity that are frequent obstacles to univocal analysis. Particular emphasis is given to the "significant address" method of programming which was developed to speed up the procedure (processing times, at present, are 0.5-1.5 seconds for sentences up to 16 words). By structuring an area of the central core in such a way that the individual location of bytes becomes significant, the shifting of information is avoided; the use of binary masks further simplifies the many operations of comparison required by the procedure. Samples of print-out illustrate some salient features of the system. (Author/MK)

Descriptors: *Computational Linguistics/ Computer Programs/ English/ Form Classes (Languages)/ Kernel Sentences/ Linguistic Patterns/ Machine Translation/ Phrase Structure/ *Programing/ Semantics/ *Sentence Structure/ *Structural Analysis/ Structural Grammar/ *Syntax
Identifiers: *Correlational Grammar/ Parsing

EJ068708 LI502664
The Resolution of Syntactic Ambiguity in Automatic Language Processing
Earl, Lois L.
Information Storage and Retrieval, 8, 6, 277-308, Dec 72

This paper describes how the problem of resolution of syntactic ambiguities is approached in the parser PHRASE, developed for use in experiments [illegible] automatic indexing and extracting. PHRASE is a mu[illegible] parser for declarative sentences, in which the synt[illegible]ture is built up in four stages. (10 references)

Descriptors: *Computer Proc[illegible]ronic Data Processing/ *Information Processing/ *L[illegible] Syntax
Identifiers: *Automatic Langu[illegible]ocessing

7920217 79-3-000651

Prediction and Substantiation A New Approach to Natural Language Processing

DeJong, Gerald

Cognitive Science A Multidisciplinary Journal of Artificial Intelligence, Psychology, and Language, US ISSN 0364-0213, New Haven, CT, 1979, 3:251-73

Doc Type: journal article

Descriptors: linguistics - linguistics, general - linguistics, computational - mechanolinguistics - Automated Analysis

Descriptor Codes: 0302020003

7804027 7804027

The Automatic Transformational Analysis of English Sentences: An Implementation

Hobbs, Jerry R.; Grishman, Ralph

City Coll City U New York, NY 10031 & New York U, NY 10003

International Journal of Computer Mathematics- 1976, 5, 4, 267-285. CODEN: ijcm-a

Gordon & Breach Science Publishers, 42 William IV St., London WC2, England; or Gordon & Breach Science Publishers, One Park Ave., New York NY 10016

Section Heading Codes: 5113

A system being developed for the transformational analysis of complex Eng sentences is described. The system is designed to serve as a "front end" for a variety of applications, such as question-answering, information retrieval, & command systems. This two-stage system has as its first stage the previously developed Linguistic String Parser. Unlike other systems, this system performs tests directly on surface trees, eliminating the need to perform grammatical decomposition before completing surface analysis. Major aspects of the target representation are outlined. Two types of operations were added to the previously obtained Restriction Lang: an operation for transforming trees, & one for sequencing the transformations. Three transformations are discussed in detail: passive right adjunct, gerundive nominal, & nominalization of Vs. Transformations remaining to be worked out include those yielding correct analyses of adverbials & those tracing adj functions. Modified HA

Descriptors: TRANSFORMATIONAL AND GENERATIVE GRAMMAR; ENGLISH; SENTENCE; DEEP STRUCTURE AND SURFACE STRUCTURE; COMPUTATIONAL LINGUISTICS

Identifiers: automatic transformational analysis, English sentences;

7811699 78-3-000679

Contextual Reference Resolution in Natural Language Processing

Lockman, Abe David

Dissertation Abstracts International, Pt. A US ISSN 0419-4209, Pt. B US ISSN 0419-4217, Ann Arbor, MI, 1978, 39:1863B

Doc Type: journal article

Descriptors: linguistics - linguistics, general - linguistics, computational - mechanolinguistics - Automated Analysis

Descriptor Codes: 0302020003

7811872 78-3-000569

Parsers for Indexed Grammars

Sebesta, Robert W.; Jones, Neil D.

International Journal of Computer & Information Sciences, Gainesville, FL, 1978, 7:345-59

Doc Type: journal article

Descriptors: linguistics - linguistics, general - linguistics, computational - mathematical models

Descriptor Codes: 0302020001

653165 ORDER NO: AAD65-03293

A HEURISTIC APPROACH TO NATURAL LANGUAGE PROCESSING 169 PAGES.

MANELSKI, DENIS MARTIN (PH.D. 1964 NORTHWESTERN UNIVERSITY).

PAGE 6446 IN VOLUME 25/11 OF DISSERTATION ABSTRACTS TERNATIONAL.