[From: *TA Informations* 1967]

COMPUTATIONAL LINGUISTICS : RESEARCH IN PROGRESS AT THE RAND CORPORATION

David G. HAYS*

The RAND Corporation, Santa Monica, California

Although the subject of machine translation was considered at The RAND Corporation in 1950, the workable scheme developed then was not carried out. Its authors became involved in other matters, and it was not until the late fall of 1957 that a continuing research project was established. One of the first themes of the new group's doctrine was that linguistic research would be necessary before automatic translation could be made real ; another, concelved soon after, was that a parsing program could be independent of the grammar of any particular natural language. A modestly successful system for Russian-to-English translation was assembled by 1960 (the dictionnary, grammar, and rules of translation were all very small), but by then it had become clear that the broader field of computational linguistics deserved general attention, and that machine translation should either be pursued as a developmental task, using existing techniques for a practical purpose, or set aside to await progress of a fundamental nature. The latter course was followed at RAND.

The current work of the linguistics group includes theoretical studies, descriptive studies of English and Russian, and development of basic computing tools. These have been the areas of interest to project members for a long time, and it seems likely that they will be for some time to come.

1. Linguistic theory.

Dependency grammar plays a central role in work on linguistic theory at RAND, but it is not the only model ever considered. Elaborations, using

* Any views expressed in this paper are those of the author. They should not be interpreted as reflecting the views of The RAND Corporation or the official opinion or policy of any of its governmental or private research sponsors. Papers are reproduced by The RAND Corporation as a courtesy to members of its staff. This paper was prepared, in august 1966, for publication in *T.A. Informations*.

dependency structures as a basis for transformational or stratificational systems, are being studied. How to represent the content of natural-language documents for manipulation in information retrieval, question answering, and other procedures is presently regarded, at RAND as elsewhere, as an urgent question for which a practical answer is obtainable. Three possible answers are provided by transformational grammar (base P-markers), by stratificational grammar (sememic networks), and by mathematical logic (for example, a predicate calculus). One way to approach the question is to examine the consequences of different decisions with respect to a relatively narrow subject area. Jacquelyn De Meire will be examining some English verbs from this point of view during 1966-67 ; she will work in close collaboration with Hays and Martin Kay.

The group's interest in transformational grammar is reflected in two papers delivered at the 1966 summer meeting of the Linguistic Society of America. Kay and Meyer Wolf examined the general organization of the transformational model, in which syntactic structures are interpreted by semantic and phonological components. They find its general structure ill-suited to the design of models for expression and understanding of meanings, i.e., models for the performance of the human being as speaker and as hearer. The speaker must produce a syntactic structure that expresses his intended meaning ; the hearer must obtain one that matches what he perceives. In either case, interpretation of a given syntactic structure is the inverse of the operation required. According to Kay and Wolf, therefore, it is misleading to say that the current transformational model is a model of the human being's competence, if it be understood that his competence is put to work during his performance.

Hays reviewed some details of tranformational theory, recommending consideration of alternative treatments at several points. He spoke about the use of complex symbols at all nodes in P-markers ; this treatment, which is familiar enough in the practise of computational linguistics, but not part of the standard formalizations of grammatical theory, would reduce the motivation to allow unary branching in P-markers. Distinguishing among production, recognition, and characterization, Hays suggested study of mechanisms for characterizing sets of P-markers ; the set of strings characterized by such a mechanism would be the set of terminal strings associated with the P-markers. Mechanisms of this kind would replace those using rewrite rules, and turn the attention of grammarians away from the study of strings. A third problem is that of stating structural descriptions in transformation rules ; the current formalism allows only strings, but grammarians writing transformational rules sometimes violate the formalism by showing two or more levels of constituency. To meet their needs, a formalism allowing structural descriptions in the form of trees should be explored. Moreover, if P-markers are labeled at all nodes with complex symbols, the structural descriptions might well be able to specify the values of individual features or sets of them, without giving a full description of the labeling at a node.

2. Descriptive studies : Russian.

With support from Rome Air Development Center, an agency of the United States Air Force, the RAND group is constructing from several sources a large file of Russian text on magnetic tape. RAND's contribution is merely to assemble tape, standardize it with respect to encoding and format, and impose bibliographic controls. Initially, the file is to contain about 50 millions words ; it will be available very widely.

Dean S. Worth is preparing a dictionary of Russian words, arranged by derivational history. The first publication will be a list of words grouped by base morphemes. Other arrangements of the material may be published thereafter. In support of research on derivational morphology, Worth is also preparing a deep index to the literature of the field. The index will be stored as a catalog (see below) on magnetic tape. Worth's exploitation of these materials is being reported in a sequence of studies published as RAND Memoranda ; one has appeared, and another is in preparation.

Several years ago, a file of Russian text was annotated with syntactic descriptions and rough translations. That file has served as the basis for several studies conducted by Kenneth E. Harper. Most recently, he has examined the differential tendencies of Russian nouns to govern adjectives or nouns in the genitive case. He finds great diversity among nouns ; some are almost always modified, others almost never. Furthermore, he notes that nouns modified about half the time are modified almost always when they occur for the first time in an independent publication. Although further study is needed to clarify the meaning of this result, it seems apparent that the interpretation of an unmodified noun late in a text must be strongly influenced by modifiers used with it much earlier.

In 1965-66 a new file of text was annotated with much fuller syntactic descriptions than before. The new file, which is four times as large as its predecessor (about 1 100 000 words), is marked with syntactic functions as well as dependency connections. Morphological annotations are to be added. The file is intended to support studies of Russian syntax and semantics, at RAND and elsewhere. The design of a programming language to make requests for specialized concordances easy to write is already proceeding, and consideration is being given to methods for automatic classification of some of the data contained in the file. Roger M. Needham, of the Cambridge Mathematics Laboratory, is devoting some of his time to this problem during an extended visit to RAND. This work has been sponsored by Rome Air Development Center.

3. Descriptive studies : English.

At present, little work on English is being conducted at RAND. The work of De Meire was mentioned above. Adam Makkai continued his examination of idiom structure during his participation in a postdoctoral seminar, 1965-66 (see below).

Several dictionaries of English that had been put on magnetic tape elsewhere were merged and put into catalog format during the summer of 1966 by Ronald Jonas working with Marjorie Rapp.

The interests of the postdoctoral seminar participants for 1966-67 are likely to increase the level of work on English.

4. Computing tools.

Motivated by the belief that linguistic research requires large files, which can be useful for many purposes in many places once they are established, the RAND group has devoted a large proportion of its efforts to the development of encoding schemes and formats for the storage of textual material. Not only works of literature and scientific articles are envisaged as materials to be stored, but also dictionaries, concordances, library catalogues, and many other files in which much of the information is best recorded in natural language.

The text encoding scheme provides an unlimited set of characters and features of arrangement; it is thus suitable for text in any language. Many of the limitations that have troubled scholars who turned to the computer for language-data processing are limitations of input and output devices. The RAND group undertook to design an archival encoding, isolated from these limitations; there is no reason why the image of a text stored on magnetic tape should show the scars of passage through a paper-tape typewriter or card punch. Instead, it is proposed that the input device available for a given job be adapted to the text at hand in whatever way is convenient, without regard to the different requirements of other jobs and other texts. The encoding produced at the keyboard is then to be converted by a computer program into the universal encoding for archiving of the text. Thus, the special conventions adopted in one place for one kind of text and one kind of input device are not permanently installed in the archives, nor is the internal encoding limited to what can conveniently be produced at a keyboard. The linguist who goes to the archives need never know what the input conventions were, nor even what input device was used. He is free to combine, for his new purpose, texts originally typed in different ways.

To make the use of such a scheme practical, a conversion program is needed; one has been written and tested for the IBM 7040/44. A text is submitted to this program together with a statement describing the input device and the conventions governing its use in this one application. The program follows the implicit instructions contained in the description, which can be simple when simple conversion suffices, or complex when necessary, for example in the typing of dictionary or bibliography entries.

On the other side, printing of information stored in the archival format can raise problems, since the characters encoded may not be available on the printing mechanism that can be used. Schemes for transliterating and arranging textual information are therefore important, and a package of programs for this purpose has been designed.

The catalog system establishes a format for recording relationships among elementary units called data and provides programs for manipulating the data in accordance with their content and relationships. For example, a dictionary must include data of several classes, and for different purposes it may be necessary to present the contents of the dictionary arranged alphabetically, by grammatical category, and in other ways. It may be necessary to merge the contents of different dictionaries, to add new information or new categories of information, and so on. Similar statements could be made with respect to bibliographies and other large files.

At RAND, the text and catalog system is in use for storage of bibliographies, a dictionary of English, a dictionary of Russian, files of Russian text with and without annotations, and other materials. Naturally, the designers of the system hope that it will be adopted elsewhere; the benefits each user obtains from adoption increase in proportion to the number of users.

The Centre d'Études pour la Traduction Automatique at Grenoble, under the direction of Professor B. Vauquois, has contributed in important ways to the design of the system. At RAND, the leading designers have been Kay, Theodore W. Ziehe, Srederick D. Valadez, and Patricia A. Graves.

5. Other activities.

During 1965-66, a postdoctoral seminar was held at RAND under the sponsorship of the National Science Foundation; it is to be repeated in 1966-67. Meyer Wolf used the year to improve his ability to use programming languages for several purposes, and to experiment in the application of automatic-classification methods to dialect geography. His purpose is to explicate the determination of boundaries between dialect areas from evidence consisting of many different isoglosses. Adam Makkai devoted his time to extension of his work on English idioms. Taking the stratificational viewpoint set forth by Sydney M. Lamb, Makkai regards idioms as irregularities in either semolexemic or lexomorphemic conversions. His purpose is to identify idioms of the two kinds and to show criteria for assigning each to a particular class. For the second year of the seminar, five participants have been selected.

Bibliography is a particular interest of the RAND group; Hays, Rapp, and Bozena Henisz-Dostert survey current literature, submitting monthly lists to *The Finite String*. Two annual bibliographies have been published as RAND Memoranda, and a large collection of abstracts has been established.

A bibliography of the RAND group's publications, and those from other groups at RAND that bear on computational linguistics directly, was published in 1964, and revised in 1965 and 1966. The publications of the group, like most other RAND publications, are deposited for general use at several libraries in the United States, one each in Australia, Canada, and Japan, and at four European libraries: the National Lending Library for Science and Technology, Boston, Spa, Yorkshire, England; the Société Française de Recherche Opérationnelle, Paris; the Stadt- und Universitäts-Bibliothek, Frankfurt a. M.; and the Akiebolaget Atomenergi, Stockholm.

Except as noted in the text, the work described has been supported by United States Air Force Project RAND.