American Journal of Computational Linguistics *Microfiche 81*

THE FINITE STRING

VOLUME 15 NUMBER 5

Released for publication March 24, 1979

With this issue, David G. Hays completes his term as Editor of AJCL and breathes a sigh of relief. Personal matters have made the last two issues of AJCL for 1978 excessively late. The next issues of AJCL will appear on paper; but the circumstances of the moment suggest that digital magnetic recording and direct wire transmission will be suitable for experimental use shortly.

AMERICAN JOURNAL OF COMPUTATIONAL LINGUISTICS is published by the Association for Computational Linguistics

EDITOR, 1974-1978: David G. Hays

EDITORIAL ASSISTANT, 1977-1978: William L. Benzon

MANAGING EDITOR, 1977-: Donald E. Walker, Artificial Intelligence Center, SRI International, Menlo Park, CA 94025

TECHNICAL ADVISOR, 1975-: Martin Kay, Xerox Palo Alto Research Center

CONTENTS

N -14
Computer Studies in
Formal Linguistics

Department of Computer and
Communication Sciences
THE UNIVERSITY OF MICHIGAN
Ann Arbor, Michigan 48109

July 1978

COMPUTATIONAL LINGUISTICS IN THE USSR

Joyce Friedman

ABSTRACT

As part of an official U.S./USSR Science Exchange on Applications of Computers in Management, a subgroup on natural language processing visited the Soviet Union from May 28 through June 11, 1978. The group met with scientists in Moscow, Novosibirsk, Leningrad, and Kiev. There were formal meetings and presentations of technical material, and also many informal discussions. This report presents a view of Soviet computational linguistics which emerged from these discussions.

Background

The U.S./USSR Science Exchange on Applications of Computers to Management includes many subtasks. The exchange in natural language processing is one task under the topic "theoretical foundations for software in applications in economics and management". The exchange in natural language processing was to have begun in June 1977. However, a scheduled trip by U.S. scientists was cancelled at the last minute by the USSR side; the reason given was that there were no hotel rooms available in Moscow. In spite of this initial disappointment the exchange began in November 1977 when three Soviet scientists visited the United States for two weeks. The visitors were Alexander Narin'yani of the Academy of Sciences Computing Center in Novosibirsk and Victor Briabrin and Dmitri Pospelov of the Academy of Sciences Computing Center in Moscow. The trip reported in this note is the rescheduled visit by the U.S. delegation. It took place May 28 to June 11, 1978.

The members of the U.S. delegation were: Donald Aufenkamp, N.S.F., U.S. Chairman of the U.S./USSR Joint Working Group on Scientific and Technical Cooperation in the Application of Computers to Management; Sue Bogner, H.E.W.; Joyce Friedman, Department of Computer and Communication Sciences, The University of Michigan; John Makhoul, Bolt Beranek and Newman, Inc., Cambridge; Stanley Petrick, Mathematics Department, IBM T. J. Watson Research Center, Yorktown Heights; Sally Sedelow, Departments of Linguistics and Computer Science, University of Kansas; and Walter A. Sedelow, Departments of Sociology and

Computer Science, University of Kansas. The U.S. delegation was accompanied throughout the trip by A. S. Narin'yani of Novosibirsk.

This report groups together similar work done in different locations. The main patterns of the natural language processing and theorem-proving systems can be viewed as based on (1) linguistics, (2) artificial intelligence, or (3) logic, although the distinctions are to some extent arbitrary. We also give an overview of the computers and programming languages available for work in computational linguistics. Work on lexicography, thesauri, and speech recognition was also discussed on the visit, but is not covered in this report.

(1) Linguistically-based Work on Natural Language

The main roots of the linguistically-based work are the meaning-text model of Mel'chuk, dependency grammar, and transformational grammar. They are variously interpreted by different systems.

Zoya Shalyapina, Laboratory of Machine Translation, Institute of Foreign Languages, described an English to Russian machine translation system under development since 1972 and based primarily on the meaning-text model. The representation is a dependency tree, with word order information, morphology and semantic/syntactic valencies. This structure preserves all the surface data but is also close to a semantic representation of the text. There is a dictionary and a grammar for each language.

The grammar rules are of the two forms: *If* <structure> *then* <condition>, and *if* <stucture> *then* <transformation>. Semantic information includes semantic descriptions of lexical and morphological units and the semantic acceptability of word pairs. There is a dictionary of 10,000 lexemes, described in terms of 30 semantic primitives The syntactic and semantic structures are compatible, so analysis goes only as deep as is necessary for a given sentence. Shalyapina's group works on linguistic aspecrs only; there is no computer implementation.

Uri Apresyan also works with the meaning-text model and with machine translation as the goal. His work is primarily on French to Russian translations, but he also works on English. His English grammar is said to be the most complete ever published: the Russian grammar will soon appear. The linguistic model will have four parts: morphology, deep syntax, surface syntax, and semantics; however, the current reduced model lacks semantics. A dictionary gives for each ord its morphology, its syntactic and semantic features (there are 150 syntactic features; 500 semantic features), the semantic criteria for possible governing words, and selectional restrictions. Rule schema or "syntagmas" go from morpheme structure to a surface syntactic structure that is an unordered dependency tree. There are about 200 syntagmas for Russian, each representing 20 rules. A syntagma allows a tree with X over Y to be constructed from a string containing X and Y under various complex conditions. The lexical information and the syntagmas determine the transformation from word strings to surface-syntactic structure. A deep structure is then defined by "paraphrastic" rules. which convert. for example, *strike* to *deliver* when the object is *a blow*. The deep structure is no longer language-specific but is universal, and serves as the basis for translation between languages. Apresyan stressed the value of continuing to work on the same linguistic model in order to complete its development; he contrasted this with the attitude of some current American linguists.

The linguist Iakalev, of the Economics Institute is developing a natural language interface for a data base system. This work has computer support and is to be running soon in a large factory. The natural language subset has sentences such as "what is the number of workers of <type> in <place>?" and is said to be easy for economists to learn. The system is based on very recent models of transformational grammar: Iakalev mentioned "traces" and some of Jackendoff's theories. The system goes from input to a deep structure from which it constructs a formula for the computation of a numerical result.

(2) Artificial Intelligence Work in Natural Language

AI-based systems are being developed at the Computing Center of the Academy of Sciences at Moscow, under the direction of Victor Briabrin and at the Computing Center of the Siberian Division of the Academy of Sciences, Novosibirsk under the direction of Alexander Narin'yani, in Ershov's group.

The system demonstrated to us in Moscow was DILOS (Dialogie Information Logical System). This work is heavily influenced by artificial intelligence work in the U.S. (Briabrin spent seven months at M.I.T. working with William Martin and with Carl Hewitt.) DILOS is written in LISP and runs on the BESM-6 computer in Moscow, as well as on a PDP-11/45 at the International Institute for Applied Systems Analysis in Laxenburg, Austria. The system is intended both to test various approaches to natural language processing and for practical applications. It contains an ATN linguistic processor and a semantic processor based on frames. The current applications area is airline ticket reservations; the demonstration was however on a very small data base of AI Natural Language Systems (including DILOS, GUS, RPL, OWL, and LUNAR). The system was able to answer simple natural language questions from the data base but it was not possible from the demonstration to get a good feeling for the actual range of language accepted.

Narin'yani's group in Novosibirsk has 17 people, including 6 linguists and 9 mathematicians and programmers. Until a few years ago, the work followed Mel'čhuk's model. This has now been abandoned here and work proceeds along four lines, so far relatively independently: (1) Narin'yani is develping a *formal linguistic model* which combines dependency and constituent structure in a mixed multi-level representation. Analysis proceeds by local modification of the graph structures, expanding and compressing case frames at various levels. The linguistic model so far includes formal description of adverb groups and adjective groups. This formal model has now been written up, but so far is not implemented. (2) The semantic question-answering system VOSTOK-0 contains a formal model of time. On the basis of texts of sentences such as "From the 3rd up to the 10th of March

Mike was in Moscow" it answers questions like Where was Mike at noon on the 17th of March?". The system is coded in SETL and was demonstrated to us. While the natural language fragment is still small, even for a model of time, (e.g. no time adverbials), the inferencing scheme worked successfully. (3) Several "applicational" systems are being devel ped. The first of these, the PL-1 "mini" or "toy" system ZAPSIB-0 uses essentially no syntactic analysis (though it relies heavily on word order). It has a well-defined subject domain, a data base of personnel information, and answers questions such as "who under 30 earns more than average?" (Salary information is public in the USSR.) In this very limited subject domain, the approach works well. The "midi" applicational system is under development and is more syntactically oriented. It will contain a nondeterministic bottom-up parser for a binary context-sensitive grammar with discontinuous constituents. (4) The final subgroup is the programming language group; it has implemented SETL on the BESM-6.

(3) Logic-based work in Natural Language,

In Moscow, at VINITI, the linguist E. B. Paducheva and the mathematician T. D. Korelskaya are developing jointly an approach to natural language processing based on both transformational grammar and first-order logic. The current domain is converse theorems in geometry. The system is able to process geometry theorems and produce their "converse theorems". In this system the semantic representation language is first-order logic. Algorithmic procedures for analysis and synthesis have been developed, as well as processing procedures within the logic. The linguistic part of the method is based on transformational grammar. As is the case with most of the Soviet work on transformational grammar, the deep structure uses dependency grammar rather than constituent structure grammar. The transformations are originally written in the forward direction, i.e. from deep to surface structure. Analysis is done using a "reversed" version of each transformation (not obtained automatically). While the forward transformations are independent of order the reversal rules are strictly ordered, for efficiency. There are 30-35 transformations, each expressed as a structural description, given as a template, and a structural change, given as a sequence of elementary operations. The work is developed in detail, but has no computer implementation. The system is said to contain interesting solutions to problems of quantification, negation, and conjunction reduction. The authors reported, with some amusement, that the description of the work was printed in 42,000 copies.

The current work at the University of Leningrad under G Tseitin, Faculty of Engineering and Mathematics, was described to us by others as based on logic, but Tseitin himself took a philosophical approach in his discussions with us. His remarks were more suggestive than descriptive. He indicated that his approach to natural language was by analogy to programming languages, using macros as in operating systems. He claimed "that there is no such thing as meaning", but said that his approach did use procedural semantics. His previous work on complexity and theorem-proving is not related to his work on natural language. However, he'd argue that a natural language system for computers should reflect the fact that natural language performance by people does not require exponential time. Tseitin's own current work is not on natural language, as he is busy writing an ALGOL68 implementation.

Tseitin and Liakina, formerly of the Faculty of Philology, also talked about several earlier natural language systems which I am unable to distinguish. They are described in a number of publications from 1966 on. In general, they employ dependency grammars, and use transformations during syntactic analysis. Restrictions on the grammar are stated in the predicate calculus and resolution theorem-proving is used. The goal is English to Russian translation of scientific prose.

The system of J. Kapitonova, Head of the Laboratory of Applied Cybernetics at the Institute of Cybernetics at Kiev, is an interactive theorem-proving system for mathematical texts. The objective is to be able to fill in the standard gaps in proofs, as indicated by "it is obvious that" or "as in the proof of the previous Theorem". The text is first processed manually into a highly stylized mathematical language. Only the formal material, theorems and proofs, is analyzed; discussion is treated as comment and is ignored by the programs. Several large texts, including Curtis and Reiner *Algebraic Theory of Groups*, have been preprocessed. The theorem-prover is tailored to the specific mathematical domain. It uses resolution theorem-proving, heuristic techniques, as well as special mathematical and logical techniques. The system has been programmed and is about to be tried out on a recent thesis. This project is of ten years duration, and has had a minimum of 12 people.

Interest in Montague grammar was considerable. My talk in Moscow was very well attended and there were many good questions. The audience was generally familiar with Montague's work and with recent papers on the topic in *Artificial Intelligence* and *Theoretical Linguistics*. The interest seemed to come from a more general interest in logic as a knowledge representation in natural language systems. Agafanov in Novosibirsk is also interested in the possible applications of Montague grammar to programming languages.

Computing in Computational Linguistics

Computer access appears to be much more difficult to obtain for computational linguists in the Soviet Union. Many of the projects had no computer support, even though they were in areas where computer testing of grammars or theories could be very useful. Most of the computing was on the second-generation computer BESM-6, although there are more recent computers, e.g., the ES-EBM (Ryad), series, available for other purposes. U.S. computers were on order from Hewlett-Packard, CDC, and Burroughs. The terminals we saw were mainly graphics terminals from Eastern Europe, with both Roman and Cyrillic character sets[6], and seemed fine in use.

There is much interest in advanced programming languages.

SETL is implemented in Novosibirsk. (This is with the aid of the U.S./USSR Science exchange.) In Moscow, PASCAL is implemented. In Leningrad, Tseitin is implementing ALGOL68 for the Ryad series of computers, compatible with the IBM 360.

We did have occasion to see some interactive systems in operation. The languages were impressive, but the programmer support was not. There seemed to be few error diagnostics. When there were crashes it was not possible to tell which were due to the computer and which to the programs.

Conclusions

Work onnatural language processing in the USSR seems to be along three major lines. The work by linguists is motivated by machine translation. It relies on versions of Mel'chuk's meaning-text model, with some type of transformations on a dependency base. It is characterized by a great deal of sophisticated development of large grammars, by large groups of linguists, but is without computer support. The artificial intelligence work is directed toward data base information systems, is at an earlier state of development, and is heavily based on U.S. work. It is carried out in Computing Centers and has good programming and computer support. The logic-based work in carried out by individuals or small groups in several locations without computer support, and by one large group with computers.

CONTENTS

CONTENTS