[From: A.Zampolli, N.Calzolari (eds.) *Computational and mathematical linguistics*. Proceedings of the International Conference on Computational Linguistics, Pisa, 1973, vol.II (Firenze: Olschki, 1980)]

ROLF A. STACHOWITZ

BEYOND THE FEASIBILITY STUDY: SYNTAX AND SEMANTICS

During the second part of the last decade, considerable advances had taken place in hardware and software development and - most important - in linguistic theory, through the incorporation of a semantic component into the theory of grammar. In view of these developments the Linguistics Research Center submitted a proposal to Rome Air Development Center with the purpose of ascertaining in some depth the then current opinions of the linguistic and computational linguistic community on the feasibility of fully automatic high quality translation. In particular, we were interested in obtaining the views of Professor Bar-Hillel. Since the funds available for this study were limited, a large number of scholars from this country and abroad, whose opinion would surely have contributed to the effort, could not be invited; we considered mainly those linguists who, due to their theoretical background, regarded mechanical translation, as, at least theoretically possible, i.e., proponents of the universal base hypothesis.

The report on this study has been published. One of the conclusions is the statement that in spite of the progress that had been made in linguistic analysis, linguistic research had dealt primarily with syntactic analysis of individual sentences and less with semantic problems and discourse analysis. As a result, "current linguistic theory is inadequate for machine translation". The recommendations indicate that for improved machine translation, research in the areas of descriptive linguistics, theoretical linguistics, and comparative linguistics among others was necessary and should be supported. Moreover, research in discourse analysis and production of coherent discourse should be encouraged and various grammatical models different from transformational grammar should be investigated.

There was general consensus among the various participants about these points. The main differences pertained to the extent to which pragmatic information, that is, knowledge of the world, had to be incorporated into a mechanical device to produce an acceptable translation, as well as to the amount of time needed to produce a comprehensive grammar for any language.

As a result of this study, Rome Air Development Center has expressed an interest in obtaining opinions of Western and Eastern European linguists on the feasibility of MT, especially of the Soviet linguists, who had given an estimate similar to ours about the length of time required to produce acceptable MT.

We have proposed a continuation of the study, this time with a shift of emphasis to those areas for which support has been recommended in the first study. The participants would be adherents of operator grammars, categorial grammars, string analysis grammar, dependency grammars, also logicians interested in the formalization of the semantics of natural language, such as the collaborators of Richard Montague in this country or members of the Konstanz Research Group in Germany, and finally, members of those groups here and abroad who had begun to work on problems in discourse analysis.

This second study will concentrate on three topics:

1) on the applicability of these grammatical models to MT;

2) on the results of intermediate studies in discourse analysis, in particular, the results of contrastive studies of scientific and non-scientific prose, to evaluate whether scientific language might indeed have a simpler syntax and pose fewer linguistic problems than ordinary prose and;

3) on the possibilities of applying the findings of the logical semanticists to problems of MT, in particular, to the formal representation of meanings and disambiguation of sentences in textual co-text.

We proposed the following procedure: each research group should submit two statements: a critique of the views in the feasibility study report and its appended papers, and a representation of their own model, its application to MT, and their estimate of the length of time required to obtain acceptable MT.

Copies of these statements would then be distributed to the other participants, each of whom was expected to react in writing to both statements. Statements and reactions were to be discussed in a final conference. A common statement of all participants on the feasibility of MT, their conclusions and recommendations would appear in a final report prepared at the University of Texas at Austin.

During the sessions of the first study, there had been considerable disagreement among the participants as to the extent to which pragmatic information has to be provided to a mechanical system for quality translation. The most extreme opinions were on one hand, that such an MT device had to be equipped with an information retrieval system containing the knowledge of the world, on the other, that it would need no pragmatic information whatsoever since the human reader could provide this information. Due to the lack of empirical data the issue could, of course, not be decided. We, therefore, also proposed as part of the second study, a two-year research effort on the necessity and amount of pragmatic information for purposes of automatic analysis and MT. Though administered by UT, Austin, this study is to be performed at The University of Jerusalem. Professor Bar-Hillel has already expressed his consent.

Apart from providing one of the models to be discussed the Center hopes to contribute to this study the results of an investigation on the type and amount of textual information necessary for the disambiguation of individual sentences in context (cf. p. 10, below).

Our proposal is still being negotiated; the second study will materialize as soon as funds are available.

The main topics of the second feasibility study reflect the main requirements for MT. An MT system needs grammatical models for which a recognition procedure is provided in order to perform analysis of input strings. It needs a discourse component in order to perform co-textual disambiguation of individual sentences and it needs a semantic component for at least two reasons:

a) to provide the discourse component with the meaning rules necessary for co-textual disambiguation, and

b) even for such similar languages as German and English, it is often necessary to translate from meaning representation into meaning representation since the posited deep structure of some German sentences have no corresponding English deep structures. This brings up the additional problem of how to generate surface structures from meaning representations.

Moreover, an MT system needs to be able to account for idiomatic expressions, lexical collocation and semi-sentences.

In spite of the research on the universals of language, contrastive analysis of pairs of languages has largely been neglected. Only a linguistic theory which incorporates a theory of reference, a theory of meaning and a theory of discourse analysis can provide MT with the tools it requires. We all know that currently no linguistic model exists which could fulfill the more important of these requirements.

Among the promising possibilities are extensions of the work that was begun by Richard Montague. Montague had assumed that there were “no important theoretical differences between natural languages and the artificial languages of logicians; indeed I consider it possible to comprehend the syntax and the semantics of both kinds of languages within a single natural and mathematically precise theory”. In a series of papers he developed a formalization with syntactic, semantic and interpretation rules which permitted him to present each sentence of a subset of English as one or (if ambiguous) more formulas of intensional logic and to associate each such formula with an interpretation. The interpretation represents the meaning of the English sentences and of their logical representation. Montague originally developed a theory of reference since he assumed he could do without the sense (the logical term for the linguistic term “meaning”) of an expression. In his two most recent papers, however, he revised his opinion. The sense of an expression can be defined by means of the totality of its referents in all logically possible models of his intensional logic.

Montague's semantics does not include a set of “meaning postulates” as Carnap had originally proposed, however, I believe that such a meaning rule component can be easily incorporated.

Klaus Brockhaus and Arnim v. Stechow of the Konstanz Research Group have described a formal semantics, in which particularly the implication rules of a meaning rule component are dealt with. They introduced semantic relators for hyponymy, negation, synonymy and incompatibility and show the various relations which can hold among them.

With the capability of representing English sentences as creative formulas of a logical calculus and with a meaning rule component disambiguation of sentences in co-text will clearly be possible.

Other proposals to get around the difficulties confronting a grammar which only generates individual sentences have been made by proponents of text grammars such as Petöfi and Ihwe in Konstanz. They proposed to enlarge the grammatical component of transformational grammar to generate sequences of sentences. The transformational component is to contain additional transformations to guarantee the coherence of sentences and the establishment of coreferentiality.

A semantic component with meaning rules which permit the atomization of meaning and a discourse component which permits the establishment of sentential reference is clearly a necessity for an information retrieval system which processes sentences of an unrestricted natural language. Considering the amount of information that needs to be handled by such components, it is certainly legitimate to ask what effect their incorporation into an MT system would have on the cost and speed with which MT can be performed.

An information retrieval system surely needs all the information in its data base, i.e., all the information in the co-text, if the data base is constructed based on textual information. It seems reasonable, however, to assume that immediate co-text is sufficient for disambiguation of sentences for purposes of MT. After all, the problem in MT is a carrying across of meaning, not an atomization of meaning. It is perfectly sufficient to translate an input string into what David Lewis calls "markerese", i.e., a meaning representation of Katz-Postal flavor. In such a representation, the meaning of a particular word can be regarded as being monolithic. Only those features need to be made explicit which are actually necessary for disambiguation in context. Moreover, the translation of sentences into creative formulas may be restrictable to the categorematic and referential terms of a sentence, excluding such syncategorematic terms as the quantifiers and others.

We assume that the set of disambiguating features is fairly small and does not greatly exceed the features which we already indicate in the Center's lexicographic classification. We base this assumption on the observation that nouns are normally disambiguated by the fact that their properties match the presuppositions of verbs and adjectives with respect to particular argument positions. In discourse we normally speak about objects, their properties and the various relations which hold among the objects. Thus, if a noun remains ambiguous in a sentence, it will normally be disambiguated by means of another predication on its referent in the immediate co-text.

As the rarity of pro-forms for verbs may indicate, it will be more difficult to disambiguate verbs which remain ambiguous within a sentence. Thus, a sentence like the German *Sie erhalten das Denkmal* whose translations are *They received the monument* and *They preserved the monument,* cannot be disambiguated by means of information associated with the verb complements, though in the co-text, the monument was about to fall apart, or the city had wanted to give it away for a long time, a disambiguation is easily possible. However, a semantic classification of verbs based on their permissible adverbial environment will reduce the amount of information necessary for verb disambiguation. If we add the adverbial *for a long time* to the two English translations of the German example given, we obtain the non-sentence

* *They received the monument for a long time*

and the correct translation

They preserved the monument for a long time.

(The classification of adverbs has been a fairly neglected field of research; the few studies which have appeared are highly eclectic. We assume that this situation will change now that Renate Bartsch's Habilitationsschrift on adverbs has appeared. She has set up more than thirty syntactic and transformational criteria for the classification of adverbs).

Moreover, we hope that the establishment of an association component will result in a further reduction of the amount of information in a meaning rule component necessary for verb disambiguation. Such a component, in which relations between verbs are stored in an arrangement similar to that of onomastic dictionaries and thesauri might even avoid the necessity of establishing whether incompatibility or consequence relationships hold between any pair of the presuppositions, assertions and entailments of two verbs.

At the Center, we have started working on a mechanical sentence-by-sentence translation of a large-sized portion of German text into English in order to test our various conjectures on the amount of information needed for the future meaning rule component of the Linguistics Research System for MT.

This system has been described elsewhere. Suffice it to say here that currently it consists of three components, the surface component, the standard component, and the normal form component. The surface component analyzes input sentences and brings them into a standard form, a shallow deep structure, in which discontinuous surface elements are contiguous and constituents of lexical collocations occur in a predefined order. The standard component analyzes these standard forms and filters out those strings which are not well-formed according to the standard grammar. Lexical collocations are analyzed by standard dictionary rules. These rules are transformational rules. They recognize sequences of terminal strings; they rewrite, however, the top-most terminal symbols which dominate these terminal strings. Constituents which occur within a lexical collocation but do not belong to it are extraposed.

The output of the standard component is analyzed by the normal form component which assigns to individual and connected standard subtrees a semantic interpretation, the normal form reading, which corresponds to "markerese". The rules of the normal form component are in effect transformations. They assign the same normal form reading to synonymous sentences whose deep structures cannot be related transformationally in standard transformational grammar. Normal form readings will then be interpreted by the canonization component with its meaning rules and the discourse component, whenever ambiguous readings occur.

During production, the process is reversed. We generate the target surface structures from normal forms and standard forms.

In spite of the fact that we are using a rule schemata grammar with optional constituents which generates context-free phrase structure rules with complex symbols, the number of rules necessary for the analysis of German is still fairly large since, in German, sentence constituents may occur in almost any order. We plan to introduce set theoretical rules which will permit us to state for each term in a rule consequent whether it may permute freely or has to occur in a particular position.

A further reduction of rules will be obtained by permitting intermediate operations between some of the rule constituents. Depending on the outcome of these operations, the algorithm will construct different rules from the given rule schema.

We are confident that we will be able to perform quality mechanical translation within the next five years. The greatest amount of work will undoubtedly have to be performed in the area of lexicography; after all, all the rules of the meaning rule component are, in effect, lexical rules. Fortunately, we may be assisted in this effort by the Institut für Deutsche Sprache in West Germany. Our proposal to perform joint lexicographic research is currently being discussed.

The main difficulty in current linguistics is the general lack of confirmation of one's hypotheses. In general, linguists today are dealing with smaller and smaller problems in syntax and semantics. Many have completely withdrawn from semantic issues and are concentrating on problems of phonology. The number of papers pertaining to phonology during the forthcoming LSA meeting may be indicative of that trend. MT has the advantage that we cannot restrict the input language, that we cannot select the problems that we want to deal with. We have to accept language as it comes. It enforces upon us the confirmation or falsification of our hypotheses.

We stated in our recommendations to the first study that MT should be sponsored as an intellectual pursuit contributing to our knowledge of language. I am convinced that the solution of problems which arise in machine translation will benefit general linguistics and linguistic theory, in addition to solving one of the major problems in communication today.

LRC LANGUAGE DATA PROCESSING PROGRAMS

1. *The LRC Glossary-Frequency Program.*

This program recognizes word units in any text. Any character string not containing a blank and enclosed in blank spaces is recognized as a word unit. The program produces two outputs: *a)* a glossary and *b)* a frequency count.

The first is an alphabetical listing of the distinct word units which occur in the text, each word unit is preceded by a number which indicates its number of occurrences in the text.

The second list is similar to the glossary list with the exception that the items are sorted on the number of occurrences first, the secondary sort is alphabetical.

2. *The LRC Index Program.*

The index program produces an output like the glossary program, i.e. the word unit and the number of its occurrences in the text. In addition, it prints out the location or locations of the word in the text. The location is represented by a 10-character alphanumeric string which is grouped according to the user's specification. The user may specify that certain groups are to be suppressed in the index display.

3. *The LRC Concordance Program.*

This program produces an alphabetical sort of each word occurring in the text plus its left and right environment which is specifiable by the user. Two sort options are possible: continue sorting to the right beginning with key word, and continue sorting to the. left beginning with key word. Thus identical word sequences will occur together.

These three programs operate with the following options:

a) punctuation stripping. This removes all sequences of punctuation marks following or preceding a blank space before treating a sequence of symbols as a word unit. Thus *absorb* and *absorb,* would be reduced to *absorb.*

b) Exclusion list. The programs only operate on those units which are not. identical to any word unit in the exclusion list. Any number of words can be input into the exclusion list. If *the* occurs in the exclusion list it will not appear in the output of these programs.

c) Inclusion list. Only those words in the text which are identical to the word unit in the inclusion lists are operated upon by the programs. If *the* does not occur in the inclusion list, it will not appear in the output of these programs.

4. *The LRC Dictionary Analysis Program.*

This program analyzes any sequence of symbols by means of the LRC dictionary grammar. This grammar contains about 106,000 English word stems, morphological endings, and punctuation marks and about 60,000 German word stems, morphological endings, and punctuation marks. Each word stem is classified according to the endings with which it may occur wellformedly. Stems which are different from their lemma contain a code by means of which the lemma can be constructed *(went → go).* During the mechanical analysis of words the rules used for its analysis are stored with it.

5. *The LRC Lemmatization Program.*

This program works on the output of the dictionary analysis program. It performs the following operations:

1) it ignores all final endings.

2) If the rule which analyzed the word stem contains a code for the lemma generation, it generates the lemma according to that code. If the rule does not contain such a code it constructs a lemma which is identical to the word stem and associates the whole word with that lemma.

The lemmata are then processed by the glossary-frequency program or, if an index was processed, by the index program.

The lemmatized index display has the following format:

countL Lemma count$_W$ Word form *(a)* n_1 *(x),* n_2 (*x*), ...
count$_W$ Word form *(b)* ... n_j *(x)*
count$_W$ Word from *(c),* etc.

where n_i refers to the location in the text, *(x)* is a symbol identifying the particular word form occurring in that position, in our case, *a, b*, or c; the locations are sorted in ascending order.

The capabilities of the lemmatization program can easily be extended to operate on information associated with other codes in dictionary rules. For example, if a code DER (for "derived from") with the appropriate information is added, derivational forms can be reduced to their root.

5.1. *The LRC Word List Comparison Program.*

This program compares entries in two arbitrary word lists, *A* and *B,* and produces three lists:

1) the list of all words occurring in both *A* and *B*;
2) the list of all words occurring in *A* only (but not in *B);*
3) the list of words occurring in *B* only (but not in *A).*

6. *Multiple Glossary-Frequency Program.*

This program produces a frequency count and/or a glossary of sequences of two-word units or three-word units, or two-word units separated by a one-word unit. It works with two options:

a) regular sort which sorts *engine rocket* before *rocket engine* and

b) special sort which rearranges sequences in alphabetical order thus producing *engine rocket* from both *rocket engine* and *engine rocket; rocket engine* will not appear in the output.

The frequency count of *engine rocket* is the sum of *rocket engine* and *engine rocket.*

The two following options are being added:

1) one-word inclusion or exclusion list. This will permit the program to accept or reject a multiple word unit depending on the fact whether one of its components occurs in the inclusion or exclusion list.

2) Multiple word inclusion or exclusion list. This permits the program to accept or reject multiple word sequences if they occur in the exclusion or inclusion lists.

7. *The LRC Catalogue Program.*

This program, also called the Library Program, permits a user to classify each unit in his data base according to up to 20 arbitrary descriptors, (author, title, publisher, year of publication, etc.). The values of each descriptor can be defined as simple or multiple, (one author vs. several co-authors). Sorts of any depth (up to 20) can be requested for any combination of descriptors. Thus, for example, all library units could be sorted in the sequence: publisher, year of publication, author, title. This would produce an alphabetical list of all units sorted first on publishers; all units with the same publisher would then be sorted according to year of publication, the tertiary sort would produce an alphabetical author sort for all units with the same publisher and year of publication, the fourth sort would finally arrange the titles of each such author alphabetically. Similarly, lists sorted on author first, then on year of publication, then on publisher, etc., could be produced. Values, which were defined as multiple, will appear once under each separate value; thus co-authors occur under each author.

Sorting can further be influenced by the “ignore option”. This option uses a list of words provided by the user. Words in this list are ignored during the sort when they occur at the beginning of a value. Thus the title *The LRC Programs* will be sorted under the letter *L* if *the* occurs in the ignore list.

Subsets of descriptor combinations (up to 20) can be displayed for each unit. It is thus possible to display a list of authors and their titles and to supress all additional information associated with the other descriptors. Sort and display options can be combined. The selection of sort descriptors is independent from the selection of display descriptors; thus a set of sort descriptors and display descriptors can be identical, overlap, or be disjoint.

Finally, the values of each descriptor can be input to the other LRC programs mentioned above. The Glossary and Index programs can treat a value as one word. It is thus possible to check values for correct spelling and consistency. Entries like *Wilson, Harry L.* and *Wilson, H. L.* can easily be found and corrected. The updating of entries currently requires re-encoding a whole line at a time. A program to permit the correction, insertion, and deletion of individual words is being prepared.

We are planning the following additional programs.

8. *The LRC Text Edit Program.*

This program is to process an input text *T* with the lemmatized index to produce a parallel text *T′* in which all words not occurring in the lemmatized index are removed and each remaining word is reduced to its lemma.

9. *Multiple Index Program.*

This program is to operate like the index program with the difference that it establishes sequences of word units and their location.

REFERENCES

K. BROCKHAUS, A. V. STECHOW, *On formal semantics: a new approach,* in « Linguistische Berichte», XI (1971), pp. 7-36.

T. A. VAN DIJK, J. IHWE, J. S. PETÖFI, H. RIESER, *Textgrammatische Grundlagen für eine Theorie narrativer Strukturen,* in «Linguistische Berichte», XVI (1971).

W. P. LEHMANN, A. STACHOWITZ, *Feasibility Study on Fully Automatic High Quality Translation,* University of Texas, RADC-TR-71-295, December 1971.

W. P. LEHMAN, R. A. STACHOWITZ, *Normalization of Natural Language for Information Retrieval,* University of Texas, AFOSR-69-1788, 1972.

D. LEWIS, *General Semantics,* in «Synthese », XXII (1970).

R. MONTAGUE, *Pragmatics and Intensional Logic,* in «Synthese», XXII (1970).

R. MONTAGUE, *Universal Grammar,* in «Theoria», XXXVI (1970) 3.

R. MONTAGUE, *The Proper Treatment of Quantification in Ordinary English,* in J. M. E. MORAVCSIK, P. SUPPES (eds.), *Approaches to Natural Language* (in press).