DOCUMENT RESUME

ED 061 784 FL 002 937

AUTHOR	Lehmann, Winifred P.; Stachowitz, Rolf
TITLE	Feasibility Study on Fully Automatic High Quality Translation: Volume II. Final Technical Report.
INSTITUTION	Texas Univ., Austin. Linguistics Research Center.
SPONS AGENCY	Rome Air Development Center, Griffiss AFB, N.Y.
REPORT NO	RADC-TR-71-295
PUB DATE	Dec 71
NOTE	263p.
EDRS PRICE	MF-$0.65 HC-$9.87
DESCRIPTORS	*Computational Linguistics; *Computers; Computer Science; Data Processing; *Descriptive Linguistics; Dictionaries; Information Storage; *Language Research; Linguistic Theory; *Machine Translation; Semantics; Syntax; Transformation Generative Grammar; Transformation Theory (Language)

ABSTRACT

This second volume of a two-volume report on a fully automatic high quality translation (FAHQT) contains relevant papers contributed by specialists on the topic of machine translation. The papers presented here cover such topics as syntactical analysis in transformational grammar and in machine translation, lexical features in translation and paraphrasing, requirements for machine translation, current status of hardware and software as it affects FAHQT, bilingual computer dictionaries, and the shape of the dictionary for machine translation. Volume 1 (FL 002 936) includes papers as well as specific consideration of the FAHQT inquiry. (VM)

RADC-TR-71-295, Volume II
Final Technical Report
December 1971

FEASIBILITY STUDY ON FULLY AUTOMATIC HIGH QUALITY TRANSLATION

University of Texas

Rome Air Development Center
Air Force Systems Command
Griffiss Air Force Base, New York

FEASIBILITY STUDY ON FULLY AUTOMATIC HIGH QUALITY TRANSLATION

Dr. Winifred P. Lehmann
Dr. Rolf Stachowitz

University of Texas

Syntactic Analysis for Transformational Grammars

by

S. R. Petrick

IBM T. J. Watson Research Center

If one wishes to obtain a syntactic analysis algorithm for some class of grammars, it is, of course, essential to characterize that class of grammars completely and precisely. If we merely tie down those details for which there exists abundant linguistic justification and leave unspecified those aspects of a linguistic theory which have not been so thoroughly worked out, it may be possible to propose small sets of rules to generatively account for certain linguistic phenomena but it is more difficult to give meaningful consideration to the problem of syntactic analysis. It is likely that the existence of any general algorithm for syntactic analysis depends upon the specification of the incomplete aspects of the model in question. It should be noted in this regard that the requirement that a language be recursive can offer some help in making certain decisions with respect to the construction of a linguistic theory which would otherwise be arbitrary.

As is well known, transformational theory has been changing rapidly from its inception up to the present time. There is disagreement as to the basic mechanisms that should be allowed (e.g., conventions on transformational applicability, allowable structures, primitive transformations, etc.) and as to the use to which those mechanisms should be put (e.g., lexical or transformational treatment of certain sentences).

A person who wishes to produce an algorithm for transformational syntactic analysis is faced then with a difficult task; he must on the one hand completely specify a class of transformational syntactic components for which an analysis procedure can be found, and he must on the other hand so define that class as to make it a reasonable model of contemporary transformational descriptive practice. After reading this paper the reader can judge for himself the extent to which I have met the latter requirement while at the same time satisfying the former requirement.

There are several alternatives to the course I have chosen that are open to anyone wishing to work on "transformational syntactic analysis". First, he can talk about the theoretical requirements of transformational analysis without actually working out the complete details of an analysis algorithm for any class of grammars. Such work can be valuable, especially if it contributes to our knowledge of the precise nature of transformational rules and conventions. The more assumptions we can build in, and consequently the tighter we can make our model without impairing its facility to describe language, the better that model is, and the closer we are to saying something about a discovery procedure.

A second alternative (followed by the MITRE Corporation <1>) is to seek an analysis algorithm for a particular grammar rather than for a class of grammars. There are several objections I would raise to such a course of action. First of all, even though linguists tend to be quite tentative and cautious about the properties and details of a class of grammars they propose as models of natural languages, they are certainly even more tentative about endorsing the likelihood that the particular fragmentary grammars they produce will stand up with the passage of time. My second objection to the consideration of particular grammars concerns the difficulty of producing an analysis procedure for a particular grammar. While it would appear an easier task than for a class of grammars, it is very hard to extract the necessary properties that one needs for a proof from a particular grammar without in the process specifying a class of grammars that have those properties. The situation is not unlike that in mathematics where a generalization is often easier to prove than a more restricted result. I had an excellent professor in Theory of Functions (William Ted Martin) who delighted in providing such examples. Whenever we would bog down in obtaining a needed proof we would hear his familiar advice to "Ask for more when the required result is too specific."

A third alternative which has been followed by most people who characterize their work in syntactic analysis as "transformational" in nature is to define a "transformational-like" linguistic theory based upon some algorithm for syntactic analysis, not upon the usual generative transformational apparatus. The deep structures produced by such programs often appear to be very close to those that are assigned to the same sentences by current transformational grammars, and the rules, which are variously called "transformations" or "inverse transformations" or sometimes just "rewriting rules", often bear names and functions similar to the transformations of generative transformational grammar theory. Efforts I would classify as being of this type have been undertaken by Kay <2>, Simmons <3>, Moyne <4>, Thorne <5>, Fraser and Bobrow <6>, Woods <7>, Winograd <8>, and Kellogg <9>, to name a few. The most compelling argument for such systems is their efficiency for natural language processing projects relative to existing parsers for generative transformational grammars. Surprisingly, relatively little is made of this by the proponents of these systems. The argument often given, on the other hand, namely that of suitability as a perceptual model, has been totally unconvincing in my opinion. The most important thing to note is that whatever the merits or shortcomings of such systems, they are linguistic theories which are unrelated to generative transformational grammar theory and as such their proponents face the task of independently establishing their adequacy for linguistic descriptive and explanatory purposes - their capacity for expressing significant linguistic generalizations. Unfortuantely, many of the proponents of such systems have not given enough attention to this task, basing the justification of their linguistic theories not upon their ability to account for specific linguistic data, but rather upon their tenuous relationship to generative transformational theory.

Rather than discussing these alternatives further I will instead discuss my own work on transformational syntactic analysis. Let us begin by sketching briefly the transformational analysis algorithm of my thesis.

The model of transformational theory in question is roughly that which was in vogue prior to Aspects of the Theory of Syntax <11>. The base component is a context-free grammar with certain restrictions on recursiveness and sentence embedding. Transformational applicability is specified by a structural index. This structural index is satisfied by a proper analysis which is a sequence of subtrees that constitutes a single cut through a tree. The modification to a tree by a transformation is specified by a structural change. For our simple model this is limited to substituting strings of trees for each of the trees of the proper analysis that satisfies the structural index. A number of restrictions on derivations are necessary to guarantee that the language generated is recursive.

Our analysis algorithm is based upon a reversal of the procedure used for generating a given string. This reverse procedure makes use of inverse transformations, which are mechanically computed--one for each generative transformation. In analyzing a sentence S, inverse transformations are applied in the reverse order of that in which the corresponding generative transformations are applied in deriving S.

To understand inverse transformations let us examine their generative counterparts. We observe first of all that the structural change of a transformation references a sequence of nodes that occur in the structural index, interspersed possibly with additional morphemes. We call this sequence the <u>inverse structural index</u> of the transformation in question. The structural change of course gives more information than is contained in the inverse structural index, but the latter provides the basis for our analysis algorithm.

To give an example of an inverse structural index, let us consider a passive transformation whose structural index is (NP AUX V X NP X BY PASS) and whose structural change is (5 2 (BE EN 3) 4 0 6 7 1). The inverse structural index is (NP AUX BE EN V X X BY NP) because 5 denotes NP, 2 denotes AUX, etc.

For a transformation to be applicable to a tree T there must be a proper analysis of T that satisfies the structural index of that transformation. The structural changes that may be performed by transformations as we have formalized them are limited to the substitution of a sequence of trees (including possibly the null sequence) for a single tree, a process which is followed by erasure of all nonterminal nodes that dominate no terminal symbol. Hence, for the tree resulting from application of a transformation, there must exist a proper analysis that satisfies the inverse structural index of that transformation.

We make use of this fact in the following way. If a string of morphemes s is the terminal string of a tree produced by the action of a transformation t, then it must be possible to segment s such that the ith segment can be analyzed as the ith node of the inverse structural index of t with respect to a context-free grammar consisting of the original base component rules augmented by rules reflecting structure that can be produced transformationally. It is possible to mechanically derive an augmented context-free grammar that includes all rules reflecting structure that might be formed in the transformational derivation of a given sentence. Hence, we have a necessary test that a given string of morphemes s was produced by a transformation t.

Sufficient information is given in a transformation to permit the computation of a function we will call the corresponding inverse transformation. This function maps a sequence of trees satisfying the inverse structural index of some transformation into a sequence of trees satisfying the structural index of that transformation. More precisely, if a transformation t performed on a tree T yields a tree T', we denote by P' the proper analysis of T' that satisfies the inverse structural index of t. Now the inverse transformation t' corresponding to t maps the proper analysis P' into a sequence of trees whose debracketization is the terminal string of T. For the previously considered transformation the inverse transformation can be specified in terms of the inverse structural index (NP AUX BE EN V X X BY NP) and the inverse structural change (9 2 5 6 1 7 8 PASS). Note that there is no requirement that the inverse structural index and the inverse structural change have the same number of terms. The inverse transformation, as we define it, is not a true inverse transformation for which t' t T = T.

Let us now see how an analysis procedure can be based upon our inverse transformations. We take up the analysis of a sentence S with respect to a given context-free grammar G and an ordered set of transformations $\underline{t}_1$, $\underline{t}_2$, ... $\underline{t}_n$. To simplify our exposition we begin with a grammar containing no binary transformations (i.e.,transformations are not applied in a cyclic fashion).

Using one of the methods given by Petrick <10> we determine an augmented context-free grammar G' that contains rules reflecting all structure that can be produced in the derivation of S. In reversing the generative procedure we first consider $\underline{t}_n$. If $\underline{t}_n$ was performed in deriving S, it must be possible to segment S such that with respect to G', the ith segment has an analysis as a tree dominated by the ith term of the inverse structural index of $\underline{t}_n$. If such a segmentation is possible, and if $\underline{t}_n'$ is performed on the sequence of trees provided by this segmentation, then the debracketization S' of the resulting sequence of trees must be the terminal string of the tree that existed just before application of $\underline{t}_n$. (Complete debracketization turns out to be unnecessary. Repeated debracketization of outermost structure until no derived constituent structure remains is all that is required.) If the analysis of S and S' (if it exists) are separately considered with respect to the original grammar with only transformations $\underline{t}_1$, $\underline{t}_2$, ..., $\underline{t}_{n-1}$, then the problem consists of one or more instances of essentially the original problem of analyzing S with respect to $\underline{t}_1$, $\underline{t}_2$, ..., $\underline{t}_n$. If we carry out this procedure for each of the remaining n-1 inverse transformations we obtain a set of debracketizations (S, S', ...). Further reversing the generative procedure, it remains only to determine which elements of this set are analyzable as the sentence symbol with respect to G. Every deep structure of S with respect to the given transformational grammar must be included in the set of trees thus obtained. With each tree it is also possible to associate the sequence of transformations used in obtaining it.

The analysis procedure becomes more complicated when binary transformations are included, as would be expected. Performance of an inverse binary transformation must always insert two occurrences of the sentence boundary symbol SENTB. The sequence of trees lying between these two SENTB markers corresponds to the constituent sentence, and the sequence of trees lying outside these two markers corresponds to the matrix sentence. Let us call the debracketization of the former sequence the constituent sentence continuation; and let us call the debracketization of the latter sequence, with the symbol COMP inserted to divide the left and righthand sections, the matrix sentence continuation.

It is clear that the constituent sentence continuation could arise from repeated application of the transformational cycle. Hence, the problem of determining the underlying deep structure of this derived string is another instance of the original problem. In other words, inverse transformations must be applied to the constituent sentence continuation in reverse generative order, as we already have discussed. If, eventually, no binary transformation applies on some inverse cycle, the recursion terminates; the structure thus found is of course dominated by the sentence symbol S1, and in the complete structural description of the given sentence this S1-dominated tree is attached under the COMP symbol of the matrix sentence continuation (by the rule COMP → SENTB S1 SENTB).

The generative transformational cycle works in such a way that no singulary transformations apply to the matrix sentence structure before a binary transformation has been applied to it and the constituent structure it dominates. Hence, the matrix sentence continuation resulting from an inverse binary transformation need not be subjected to the entire inverse transformational cycle. More than one embedded constituent sentence structure can be dominated by a single matrix sentence structure, however (as, for example, when both subject and object contain relative clauses), so the matrix sentence continuation must be subjected to repeated applications of the same or other binary transformations. The resulting matrix sentence continuation must finally be analyzed with respect to the base component G to see if an analysis as an S1 is possible. Every underlying structure assigned by the given grammar to S must be included in the set of structures thus obtained.

A brief reflection on why we begin the analysis of a sentence by applying inverse singulary transformations is in order. Although it is true that singulary transformations precede binary transformations in a given cycle, when the last binary transformation has been performed for the last time it is still possible for the singulary rules to apply to the result of this final embedding. Once the last singulary has been applied, however, generation is complete because no further binary transformation can be performed. The last singulary transformation for generation is therefore the first transformation whose corresponding inverse is to be applied in recognition.

As we have already observed, the exhaustive procedure we have describe must find all underlying structures assigned by a given transformational gramma to a sentence. It is possible, however, that one or more spurious structures will also be found. There are several sources of slack in our procedure that could cause such a situation to occur. One of these is related to our use of so-called "auxiliary" phrase structure rules, which reflect structure that can be transformationally derived. These rules are required in order to ensure the application of every inverse transformation necessary to reverse the generative derivation. The use of these rules, however, raises the possibility that an inverse transformation will be applied at some point where it should not apply, from the point of view of reversing a valid generative derivation. If the continuation resulting from this wrong application of an inverse transformation is not subsequently blocked, an invalid underlying structure may result.

Three other sources of incorrect "structural descriptions" are possible. All can be eliminated by suitable modification of the basic procedure we have presented. The first deals with the use of obligatory transformations. The procedure we have described finds all underlying structural descriptions of a sentence with respect to a grammar in which all the transformations are taken to be optional. It also yields all correct structural descriptions for a grammar in which only some of the transformations are obligatory, but it may in addition give erroneous structures.

The second source of unwanted structures is related to supplementary conditions that may be imposed on the applicability of a transformation. For example, the basic procedure we have described could not test to ensure that trees satisfying terms of a structural index are dominated by prescribed higher nodes.

The third source of error is related to trees that are reduplicated by a transformation. In recognition it is of course necessary to ensure that two trees are identical. It would be easy enough to mark such transformations and make the necessary tests for equality, thus eliminating this source of incorrect structures.

Incorrect structures arising from any source may be discarded during the final synthesis phase. In this phase, structures produced by the analysis process are viewed as instructions to produce sentences. The base tree and the list of transformations constitute commands defining the appropriate phrase structure and transformational rules to apply to generate a sentence. The resultant string is compared with the original input string. The structural descriptions that yield strings matching the input string are those that constitute structural descriptions of the input string s. All other structural descriptions, which yield either nonmatching strings or no strings at all, are discarded. In practice, bogus recognitions are rare. In theory, their possible occurrence renders synthesis a necessary part of an effective recognition procedure of the type we have sketched.

A thorough understanding of this analysis algorithm requires more precise definitions and some concrete examples. The reader is referred to references <10>, <12>, and <13> for these.

In the past three years an effort has been made to extend the class of grammars to more adequately reflect current linguistic theory. The principal extensions made have been provisions for handling: (1) complex symbols (nodes with features), (2) a generalized structural condition of transformational applicability, (3) stateable additional conditions of transformational applicability, (4) Chomsky adjunction, (5) the use of coordination-reduction rule schemata, (6) precyclic and postcyclic transformational components, and (7) obligatory as well as optional transformations.

Considering these extensions individually, the addition of complex symbols presents several problems. First, I have not faced the problem of lexical selection and its inverse so I have assumed that input sentences consist of strings of feature bundles. Second, I have restricted features to lexical items in the manner of Aspects <11>. The more general case of allowing features to be associated with nonterminal nodes has been considered, but there remain unsolved problems of deriving the features to be associated with the nonterminals of transformationally derived structure. Finally, certain feature-sensitive rules can give rise to nondeterministic inverse transformations. For example, if a transformation of the type [+A] → [-B] is used, it is indeterminate whether the reverse transformation should leave -B as it is, change it to +B, or delete it entirely. Separate continuations resulting from all three possibilities must be followed, and a rule of the type

$$[+A1] \rightarrow \begin{bmatrix} +A2 \\ +A3 \\ \cdots \\ +An \end{bmatrix}$$

would lead to 3^n independent continuations.

The generalized structural condition of transformational applicability and the additional conditions of applicability were allowed by making a basic change to the analysis algorithm. In the old algorithm, as we sketched it, intermediate derivational stages were not completely known; only proper analyses of intermediate trees were required and found. As an alternative to directly determining proper analyses that satisfy an inverse structural index it is possible to parse a continuation string resulting from the application of an inverse transformation up to the sentence symbol, using the augmented CF grammar; the resulting structures can be examined to insure that they satisfy the labelled bracketting structural condition and any other conditions of the forward transformation in question; and that forward transformation can actually be applied to insure the validity of the inverse transformational step. The mechanics of such a step are illustrated by the diagram which appears at the end of the paper.

It must be noted that even though a general labelled bracketting structural condition is allowed, a structural index must still be identified by means of that labelled bracketting, and a structural change is specified only through the use of that structural index, as before.

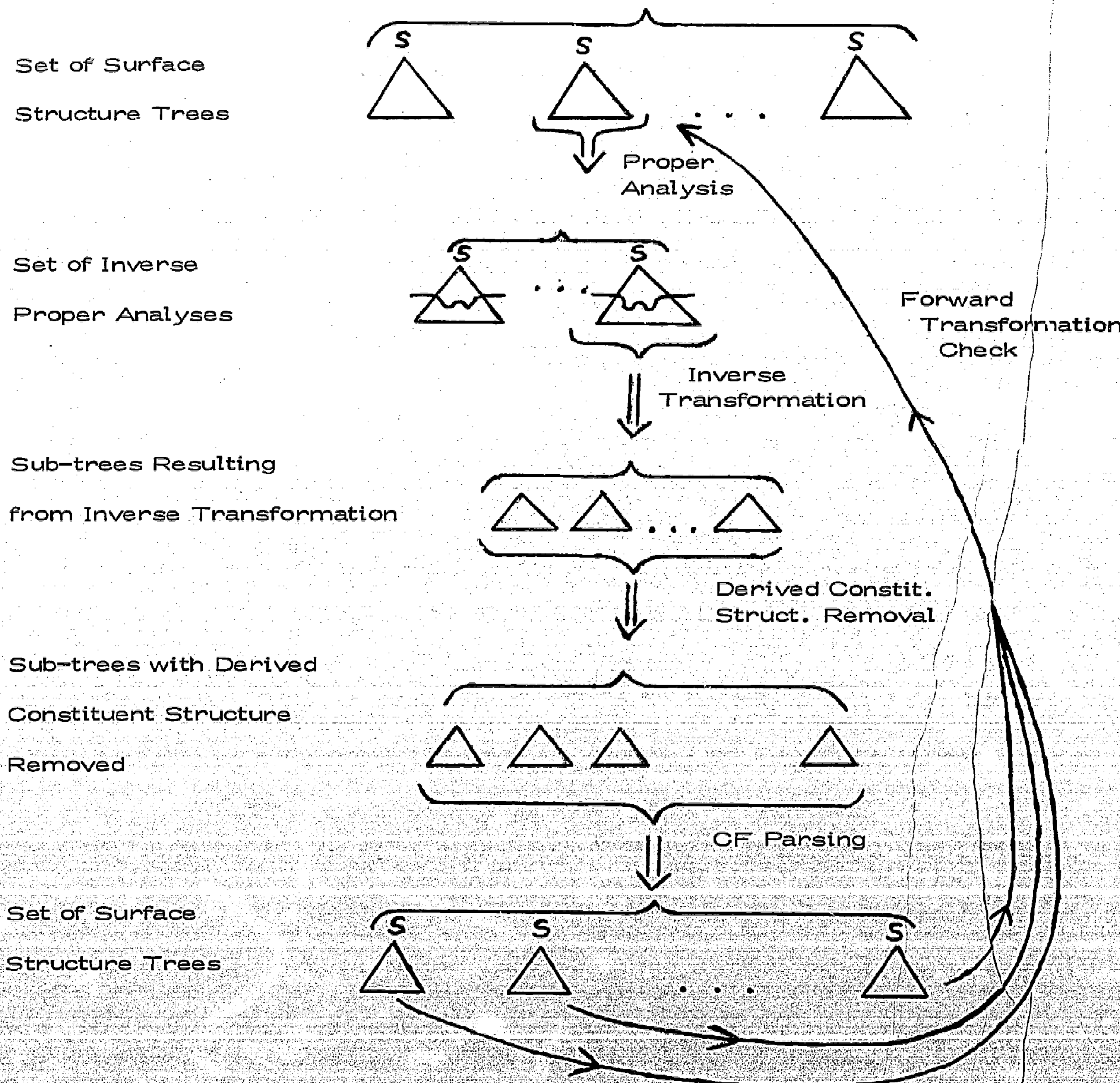
Chomsky adjunction presents no particular problem because an inverse structural index and inverse structural change may still be mechanically computed.

The enrichment of a transformational grammar to include the use of coordination-reduction rule schemata is discussed in reference <14>. Fortunately, this enrichment has an associated analysis algorithm which is deterministic.

Precyclic and postcyclic components present no new theoretical analysis problems. They do, however, offer enormous opportunities for the proliferation of spurious continuations. This, in turn, will probably require even more careful modification and tuning of a grammar to keep the analysis time within acceptable bounds.

Finally, the addition of obligatory transformations was fully discussed in reference <10> but was not programmed at that time. A presently existing transformational analysis program incorporates those considerations. This program has not yet been extensively tested and hence must be considered to be in a "debugging" state. For this reason it has not yet been described in the literature. Because it has not yet been tested on any sizeable grammar it is not possible to estimate running times. It is safe to say that the analysis of sentences with respect to a good-sized transformational grammar currently under development at the IBM T. J. Watson Research Center will undoubtedly require careful analysis-dictated modification of that grammar. In addition, the analysis procedure itself may well have to be significantly modified. The principal hope is that by actually performing forward transformational tests on the fly, spurious continuations can be avoided before they exponentially proliferate.

It is clear that at this time it is possible to produce transformational grammars (perhaps not uniformly well-motivated linguistically) which exceed the computational capabilities of the existing program. This, of course, limits the usefulness of the program for investigating the claims inherent in a given grammar (e.g., what structures are assigned to specific sentences?). It remains unknown, however, whether the current generation of transformational grammars can be modified so as to permit syntactic analysis in a reasonable time for experimental purposes. For the purpose of such applications as question answering systems, information retrieval systems and natural language programming systems this may be less of a problem than the problem of writing a grammar that specifies a sufficiently large subset of English.



THE OPERATION OF A SINGLE INVERSE TRANSFORMATIONAL STEP

PS 340

References

<1> Zwicky, A., Friedman, J., Hall, B., and Walker, D. The Mitre syntactic analysis procedure for transformational grammars, Proc. Fall Joint Computer Conference, 1965, Spartan Books, Washington, D. C., pp. 317-326.

<2> Kay, M., Experiments with a powerful parser, Proc. Deuxième Conference International sur le Traitement Automatique des Langues, Grenoble, Aug. 1967, Paper No. 10.

<3> Simmons, R. F., Burger, J. F., and Long, R. E., An approach toward answering English questions from text, Proc. 1966 Fall Joint Computer Conference, 1966, pp. 357-363.

<4> Moyne, J. A., Loveman, D. B., and Tobey, R. G. Cue: A preprocessor system for restricted, natural English, Proc. Symp. on Information Storage and Retrieval, Univ. of Maryland, April 1971, pp. 47-60.

<5> Thorne, J., Bratley, P., and Dewar, H. The syntactic analysis of English by machine, Machine Intelligence 3, D. Michie (ed.), American Elsevier, New York, 1968.

<6> Bobrow, D. G. and Fraser, J. B. An augmented state transition network analysis procedure, Proc. Internat. Joint Conf. on Artificial Intelligence, Washington, D. C., 1969, pp. 557-567.

<7> Woods, W. A. Transition network grammars for natural language analysis, Comm. ACM 13 (Oct. 1970), pp. 591-606.

<8> Winograd, T. Procedures as a representation for data in a computer program for understanding natural language, Rept. AI-TR1, Artificial Intelligence Laboratory, MIT, 1971.

<9> Kellogg, C., Burger, J., Diller, T., and Fogt, K. The converse natural language data management system: Current status and plans, Proc. Symp. on Information Storage and Retrieval, Univ. of Maryland, April 1971, pp. 33-46.

<10> Petrick, S. R. A recognition procedure for transformational grammars, Ph.D. thesis, MIT, 1965.

<11> Chomsky, N. Aspects of the theory of syntax, The M.I.T. Press, Cambridge, Mass., 1965.

<12> Petrick, S. R. A program for transformational syntactic analysis, Air Force Camb. Res. Labs. Research Paper No. 278, AFCRL-66-698, Oct. 1966.

<13> Keyser, S. J. and Petrick, S. R. Syntactic analysis, Air Force Camb. Res. Labs. Research Paper No. 324, AFCRL-67-0305, May 1967.

<14> Petrick, S. R., Postal, P. M., and Rosenbaum, P. S. On coordination reduction and sentence analysis, Comm. ACM 12 (April 1969), pp. 223-233.

Syntactic Analysis Requirements

of Machine Translation

by

S. R. Petrick

IBM T.J. Watson Research Center

342a

SYNTACTIC ANALYSIS REQUIREMENTS

OF MACHINE TRANSLATION

S. R. Petrick

In th [illegible] note I will confine my attention to machine translation (MT) systems whi [illegible] based upon an underlying formal generative grammar. This is not to [illegible] ne potential importance of various computational aids to human trans [illegible] ion, nor to deny the possibility of machine translation not based on a formal grammar. It is clear, however, that for fully automated MT any attempt to make use of presently existing linguistic theory or of that which is likely to exist in the foreseeable future requires a grammar-based approach.

A second assumption I wish to make is the existence of two distinct components of a grammar -- a syntactic component and a semantic component. The former assigns structure to sentences and the latter interprets those structures by translating them to a natural language (in the case of MT) or to an artifical language which has its own computer interpreter. It will not be assumed that the syntactic and semantic components necessarily interact in a simplistic fashion, i.e., every syntactic output is to have a distinct well formed semantic interpretation, and the final output of the syntactic component is the input to the semantic component. Instead, we will, for example, allow the syntactic component to generate structures which are rejected by the semantic component, and we will allow semantic analysis (and rejection) of fragments of a syntactic structure prior to the complete determination of that structure.

The importance of the syntactic component has been recognized for some time. For the purposes of MT it has two distinct ends to achieve: on the one hand it must specify a large enough subset of the source language to meet the operational requirements of the MT application in question. (The related function of ruling out syntactically ill-formed sentences is of limited importance in MT). On the other hand the structures it assigns must provide a reasonable basis for semantic interpretation. These two requirements are closely related, i.e., it is relatively easy to satisfy one at the expense of the other, but much harder to adequately meet them both.

A not uncommon attitude which has been expressed both in the computational linguistic literature and orally at symposia and conferences is that syntax in general and syntactic analysis in particular has been well worked over, is thoroughly understood, and presents no serious problems — in contrast to the situation in semantics where little has been done and not much is understood. I submit that such remarks reflect the experience of one who has chosen a class of grammars, in most cases context-free grammars, which permits a reasonable coverage of a source language at the expense of assigning structural descriptions which bear little relationship to underlying meaning and which, therefore, provide an inadequate basis for semantic interpretation. It is not just because large-coverage context-free grammars have been found to often assign 100 or more structural descriptions to unambiguous sentences that makes them inadequate. Rather, this is just symptomatic of a more deep-seated inability to relate form to underlying meaning.

This shortcoming is not limited to the class of context-free grammars. If the rewriting system is extended to encompass context-sensitive grammars and/or rewriting rules with whose constituents complex features can be associated then economies and linguistic generalizations are realized, but the fundamental problem of relating form to meaning appears intractable for any system which attempts to interpret the surface form of sentences. It was this realization that prompted Chomsky to propose as the basis for semantic interpretation deep structures which were in many cases far removed from surface structures. Chomsky made use of a transformational component to relate corresponding deep and surface structures, but the acceptance of the deep-surface structure distinction is a matter which is independent of any consideration of the most appropriate means for making explicit that correspondence. Accordingly, a host of models (each of which is a proposed linguistic theory even if not called such) have been proposed for mapping surface structures into corresponding deep structures, or (in some cases) for directly assigning deep structure to sentences without explicitly producing surface structure.

It is my contention that linguistic models which do not provide the deep structure of sentences (at least implicitly if not explicitly) fail to provide a basis for the semantic analysis of all but a small class of sentences, a class so restricted that its use is precluded for most applications including MT. Hence, for the remainder of my discussion I will focus my attention on the problems of syntactic and semantic analysis associated with some type of deep structure model.

As pointed out previously, there is a trade-off possible between syntax and semantics. If more is done by the syntactic component the task of the semantic component is lightened and vice-versa. Contemporary linguistic theory has been much concerned with this question of where to draw the line, and even though the questions of overall simplicity considered have not been motivated by any concern for MT, it is nevertheless instructive to consider the applicability to MT of models of present-day deep structure complexity. There is, of course, no general agreement among linguists as to the type and complexity of deep structures and of the related transformational component required by those deep structures. Even though these decisions loom large and important to linguists, however, they are not so large as to preclude assessment of the suitability of a rather large class of deep structure models for MT.

Let us begin then by considering the requirements of the semantic component. It is, of course, possible to produce sentences whose semantic analysis and/or translation requires not only a number of deep structure distinctions but also a large amount of information about the world, about logical deduction, and about the context of discourse in which the sentence appears. I am resigned to the prospect that these obstacles preclude for the foreseeable future extremely high quality translation.

My own experience with semantic interpretation has been with translation to a formal language which, although not a programming language in the sense of having an existing hardware or software interpreter, is close enough to a programming language that the task of translating it to an existing programming language is an easy one. The problem of translating a given structure to a functional programming language appears to me to be greater than that of translating that structure to another natural language. This follows from two considerations: First, deep structures of different languages which have been proposed to date are remarkably similar. In those cases where differences have been argued, they have seldom exceeded differences in subject, verb, object ordering. Deep structures differing so slightly are easily related through the use of such standard translation mechanisms as the Irons Translator[1].

Second, the task of using a transformational grammar to convert deep structures into surface structures is not conceptually difficult. Hence, it would appear that for a very large class of sentences, the translation sequence shown below should provide the basis for translation:

Source Sentence S	$\longrightarrow$	Deep Structure of S	$\longrightarrow$	Corresponding Target Language Deep Structure	$\longrightarrow$	Target Language Translation of S

Indeed, it has been my experience that semantic interpretation of deep structures through the use of the Irons or more generally the Knuth[2] translation mechanism even provides a reasonable basis for a natural language question answering system. This has also been argued by Kellogg and Thompson among others. For more discussion see reference 3.

I have argued that the use of a deep structure (read semantic structure if you prefer) generative grammar does provide a reasonable basis for MT. It does so, however, by throwing a considerable burden on the syntactic component. We have seen that structures can be assigned which appear adequate for the purposes of MT. But what of the coverage requirement, i. e., that a sufficiently large subset of the source language be specified? In addition, we must concern ourselves with the theoretical and practical requirements of syntactic analysis for a class of grammars that is capable of assigning adequate deep structures.

I will discuss these two considerations with respect to generative transformational theory and also, more briefly, with respect to other deep structure-based linguistic theories.

Let us first consider the matter of coverage. It is, of course, the case that most transformational studies of syntax do not supply completely specified base and transformational component rules in discussing syntactic phenomena. There have been, however, a few attempts to write a completely specified set of rules within a well-defined transformational framework[4, 5, 6, 7, 8]. These efforts establish a lower bound on coverage which can be achieved without sacrificing structural adequacy. It is somewhat difficult to characterize the coverage achieved by any means short of exhibiting the grammars in question. There are, however, at least two ways to give a feel for the coverage attained by a specific grammar. The first is to give a list of "representative" sentences and the second is to list the syntactic constructions and phenomena provided for. Thus, for example, Rosenbaum[7] gives derivations of the 22 sentences:

1. the boys like the girl

. . . .

21. the pajamas of a king are colorful
22. the people who approve of him think that John is smart

He also lists 79 representative sentence types and includes transformations for handling verb phrase complements, pronominalization, preposition segmentalization and raising, indirect objects, relatives, genitives, negatives, certain time and place adverbials, etc. Similarly, in a more recent effort at the IBM Thomas J. Watson Research Laboratory a transformational grammar has been produced which generates such sentences as:

> what companies had a profit which was more than ten million dollars?

and print the one element of the set which contains M which is atomic and provides such construction types as: yes-no and Wh-questions, passives, prepositional phrases, nominal structures formed from underlying abstract verbs, restrictive relatives, possessive genitives, and certain types of negatives, comparatives, and coordinate structures.

Now just as existing grammars establish a lower bound on coverage attainable there are several considerations which suggest upper bounds for at least the foreseeable future. For example, many syntactic phenomena may be identified which have not yet been studied by anyone. Many other phenomena have been studied, but the results have served more to show the existence of substantial problems than to offer compelling and widely accepted solutions. Examples here are plentiful and include coordination, gapping, and pronominalization as well as almost every syntactic phenomenon which has been studied to some extent. And finally, experimental work conducted to date shows that it is far from trivial to put together and test grammars that provide for such relatively well understood constructions as yes-no questions, WH-questions, restrictive relatives, imperatives, etc.

The large number of unexplored and little understood syntactic phenomena suggest difficulty in achieving sufficient coverage for practical application, but an even more instructive exercise in illustrating this difficulty is provided by producing a set of sentences thought to be useful and representative for some application and comparing their syntactic requisites with the facilities offered by any existing or proposed grammar. I have seen this operation carried out at the MITRE Corporation with respect to a command and control question answering application and have myself undertaken the same task for a formatted file question answering facility. The results were the same. Very low coverage was observed; certainly less than 10% of the sentences studied were covered even allowing for lexical addition and extension by including some rather obvious additional transformations. The saving feature in the case of natural language question answering systems or natural language programming systems, however, is that they need not process unconstrained input sentences. Instead the user can be constrained to and instructed as to how to limit his input in terms of both lexicon and allowable constructions. All that is required is that natural subsets provided must be learnable by human speakers and must be rich enough to permit expressing that which must be expressed in a convenient fashion. The attainability of even these requirements remains to be established but at least offers some hope of success. On the other hand the usual situation with MT is that the input is not produced with the limitations of a particular formal grammar in mind. This, more than any other single factor, convinces me that grammar-based MT offers little hope for practical usage for at least the next ten years. This is not to say that MT is not an interesting and productive vehicle for keeping linguistic research in both syntax and semantics tied to reality. Others might disagree with this assessment, of course.

There may be a few MT applications where time and economic considerations permit the phrasing or rephrasing of source sentences by speakers cognizant of a system's grammatical constraints. Such an example is the preparation of technical manuals in one language for translation into another language. This is, however, not the usual situation in MT.

When we leave the (at least for me) familiar grounds of transformational theory and consider the coverage problem for such analysis-based linguistic theories as those of Woods[9], Winograd[10], Bobrow and Fraser,[11] Thorne,[12] Moyne,[13] Kellogg,[14] Kay[15] and Simmons,[16] we are faced with a difficult task for a number of reasons. Many of these models have been used only sparingly for the specification of any natural language. Hence, there is little to go on in assessing the coverage of these models. In addition, those models for which one or more large grammars have been written have not been documented in a way and to an extent which makes the determination of coverage feasible. Alternative clarification of coverage via sample sentences and listed construction types presents the same problem as we observed for transformational grammars, but whereas most linguists are by this time familiar with transformational formalism, this is not true of the aforementioned analysis-based models. Therefore, their coverage can at present be estimated only by their originators. It is far from clear to this observer that these approaches offer the same independence of construction types as is achieved by transformational theory. In any case, none of these models have supported claims of greater coverage than that afforded by current transformational theory. It is important to note that although these models are often described as "transformational" by their originators, they have not been related to transformational theory and hence must be judged on the usual grounds of linguistic adequacy just like any other proposed linguistic theory.

The remaining consideration is the theoretical and practial requirements of syntactic analysis for a deep structure - specifying class of grammars. For those analysis-based grammars previously mentioned there are few theoretical syntactic analysis problems. In addition, the computation time required for parsing, although generally not known, could reasonably be expected to be less than that required for parsing with respect to a transformational grammar. (Whether it is sufficiently small to satisfy economic considerations is, of course, another story.) This is to be expected for analysis-based linguistic theories whose principal motivation is to facilitate syntactic analysis. It is descriptive adequacy, not syntactic analysis considerations which are most likely to preclude the practical use of analysis-based grammars.

The situation is quite different with respect to transformational grammars. There is no shortage of work in linguistic description through the use of transformational grammars, although it must be noted that most efforts are directed toward determining the allowable class of transformational grammars rather than toward developing in detail any one comprehensive grammar. Syntactic analysis for any class of transformational grammars is a very complex and time-consuming proposition. It is probably for this reason that most workers in computational linguistics have chosen to forego conventional transformational theory in favor of an analysis-based alternative.

There have been only two computer implemented efforts on transformational grammar syntactic analysis. One, carried out by the MITRE Corporation, was limited to a particular grammar; a syntactic analysis program was tailored to this grammar. The program appeared to be successful in producing desired structures in a reasonable time, but it was never established that this program invariably found all of the structures assigned to a sentence by the particular transformational grammar in question (i. e., that it was, in fact, an analysis program for that grammar).

In contrast to the MITRE approach, Petrick[17] defined a class of transformational grammars and found a syntactic analysis algorithm that is valid for members of this class. The extremely nondeterministic nature of this algorithm made unfeasible the treatment of grammars as written by a linguist unfamiliar with the analysis procedure. However, Kirk and Keyser[6] showed that by suitable recasting, a substantial portion of an existing grammar (due to Rosenbaum) could be used for syntactic analysis.

In addition to the problem of computing time, there is another serious difficulty in transformational grammar syntactic analysis. The class of grammars for which syntactic analysis algorithms have been devised does not include many of the facilities currently being used by descriptive grammarians. Indeed, transformational theory is far from static, and at any given time there is little agreement on just what should constitute an allowable class of transformational grammars. In reference 18 we give an account of the current status of syntactic analysis for transformational grammars. In summary, it can be stated that although the class of grammars for which syntactic analysis is possible has been significantly extended, the introduction of new variants of transformational theory has more than kept pace with theoretical and programming efforts to cope with them. Consequently, any given linguist would undoubtedly find that his rules and assumptions do not correspond perfectly with the formulation of the allowable class of grammars. Nevertheless, it is hoped that this class is now extensive enough to permit recasting of current transformational grammars into an acceptable form without seriously compromising their linguistic integrity.

REFERENCES

1. Irons, E. T. A syntax directed compiler for ALGOL 60. Comm ACM 4 (Jan. 1961), pp. 51 - 55.

2. Knuth, D. E. Semantics of context-free languages, Math. Sys. Theory 2 (1968), pp. 127 - 145.

3. Petrick, S. R. On the use of syntax-based translators for symbolic and algebraic manipulation, Proc. Second Symp. on Symbolic and Algebraic Manipulation, Los Angeles, Calif., March 1971, pp. 224 - 237 (Also IBM RC3265)

4. Zwicky, A., Friedman, J., Hall, B., and Walker, D. The MITRE syntactic analysis procedure for transformational grammars, Proc. Fall Joint Computer Conference, 1965, Spartan Books, Washington, D. C., pp. 317 - 326.

5. Rosenbaum, P.S. and Lochak, D. The IBM core grammar of English, Specification and Utilization of a Transformational Grammar, Scientific Report No. 1, (IBM Corp., Yorktown Heights, N. Y., 1966)

6. Keyser, S. J. and Kirk, R., Machine recognition of transformational grammars of English. Air Force Cambridge Res. Labs. final report No. 67-0316, Jan. 1967.

7. Rosenbaum, P. S., IBM English grammar II, Specification and Utilization of a Transformational Grammar, Scientific Report No. 2, (IBM Corp., Yorktown Heights, N. Y., Oct. 1967)

8. Stockwell, R. P., Schachter, P., and Partee, B. H. Integration of transformational theories of English syntax, USAF Electronic Systems Division Report ESD-TR-68-419, Oct. 1968, Vols. I, II.

9. Woods, W. A. Transition network grammars for natural language analysis, Comm. ACM 13 (Oct. 1970), pp. 591 - 606.

10. Winograd, T. Procedures as a representation for data in a computer program for understanding natural language, Rept. AI-TRI, Artificial Intelligence Laboratory, MIT, 1971.

11. Bobrow, D. G. and Fraser, J. B. An augmented state transition network analysis procedure, Proc. Internat. Joint Conf. on Artificial Intelligence, Washington, D. C., 1969, pp. 557 - 567.

12. Thorne, J., Bratley, P., and Dewar, H. The syntactic analysis of English by machine, Machine Intelligence 3, D. Michie (Ed.), American Elsevier, New York, 1968.

13. Moyne, J. A., Loveman, D. B. and Tobey, R. G., Cue: A preprocessor system for restricted, natural English, Proc. Symp. on Information Storage and Retrieval, Univ. of Maryland, April 1971, pp. 47 - 60.

14. Kellogg, C., Burger, J., Diller, T., and Fogt, K. The Converse natural language data management system: Current status and plans, Proc. Symp. on Information Storage and Retrieval, Univ. of Maryland, April 1971, pp. 33 - 46.

15. Kay, M., Experiments with a powerful parser, Proc. Deuxieme Conference International sur le Traitement Automatique des Langues, Grenoble, Aug. 1967, Paper No. 10.

16. Simmons, R. F., Burger, J. F., and Long, R. E., An approach toward answering English questions from text, Proc. 1966 Fall Joint Computer Conf., 1966, pp. 357 - 363.

17. Petrick, S. R., A recognition procedure for transformational grammars, Ph.D. thesis, MIT, 1965.

18. Petrick, S. R., Syntactic analysis for transformational grammars, Proc. of the Conference on Linguistics, The University of Iowa, Iowa City, Iowa, Oct. 1970.

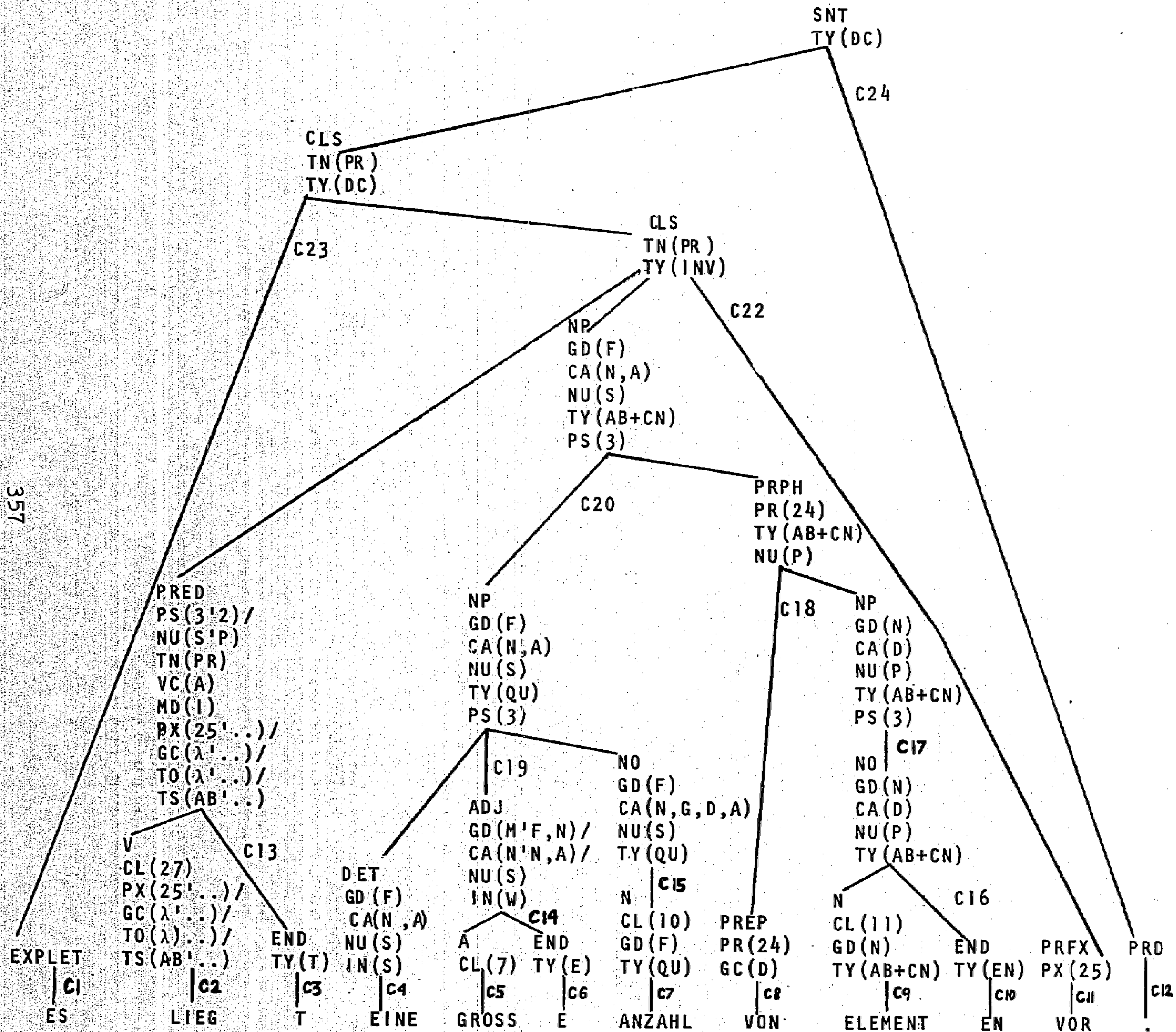
APPENDIX

Analysis of *Es liegt eine grosse Anzahl von Elementen vor.*

by Annette Stachowitz

Linguistics Research Center
The University of Texas at Austin

356


SNT
TY(DC)
C24
CLS
TN(PR)
TY(DC)
C23
CLS
TN(PR)
TY(INV)
C22
NP
GD(F)
CA(N,A)
NU(S)
TY(AB+CN)
PS(3)
C20
PRPH
PR(24)
TY(AB+CN)
NU(P)
C18
PRED
PS(3'2)/
NU(S'P)
TN(PR)
VC(A)
MD(I)
PX(25'..)/
GC(λ'..)/
TO(λ'..)/
TS(AB'..)
NP
GD(F)
CA(N,A)
NU(S)
TY(QU)
PS(3)
NP
GD(N)
CA(D)
NU(P)
TY(AB+CN)
PS(3)
C17
NO
GD(N)
CA(D)
NU(P)
TY(AB+CN)
C19
NO
GD(F)
CA(N,G,D,A)
NU(S)
TY(QU)
C15
V
CL(27)
PX(25'..)/
GC(λ'..)/
TO(λ)..)/
TS(AB'..)
C13
END
TY(T)
DET
GD(F)
CA(N,A)
NU(S)
IN(S)
ADJ
GD(M'F,N)/
CA(N'N,A)/
NU(S)
IN(W)
C14
A
CL(7)
END
TY(E)
N
CL(10)
GD(F)
TY(QU)
PREP
PR(24)
GC(D)
N
CL(11)
GD(N)
TY(AB+CN)
C16
END
TY(EN)
PRFX
PX(25)
PRD
EXPLET
C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
C12
ES
LIEG
T
EINE
GROSS
E
ANZAHL
VON
ELEMENT
EN
VOR
.

Rules:

```
C1      V EXPLET                    =     * ES

C2      V V                         =     * LIEG
        + CL(27)
        + PX(..'25'..)/
        + GC(..'λ'..)/
        + TØ(..'λ'..)/
        + TS(..'AB'..)

C2      V V                         =     * LAG
        + CL(9)
        + PX(..'25'..)/
        + GC(..'λ'..)/
        + TØ(..'λ'..)/
        + TS(..'AB'..)
```

(analogous C2 rules for the stems laeg, leg)

```
C3      V END                       =     * T
        + TY(T)

C4      V DET                       =     * EINE
        + GD(F)/
        + CA(N,A)/
        + NU(S)
        + IN(S)

C5      V A                         =     * GROSS
        + CL(7)

C6      V END                       =     * E
        + TY(E)

C7      V N                         =     * ANZAHL
        + CL(10)
        + GD(F)
        + TY(QU)

C8      V PREP                      =     * VON
        + PR(24)
        + GC(D)
```

```
C9    V N               =  * ELEMENT
      + CL(11)
      + GD(N)
      + TY(AB+CN)

C10   V END             = * EN
      + TY(EN)

C11   V PRFX            = * VOR
      + PX(25)

C12   V PRD             = * .

C13   V PRED            = V V             V END
      + PS(3'2)/          $ CL(..,27)     $ TY(T)
      + NU(S'P)                           B
      + TN(PR)
      + VC(A)
      + MD(I)
      ^ 2

C14   V ADJ             = V A             V END
      + GD(M'F,N)/        $ CL(..,7)      $ TY(E)
      + CA(N'N,A)/                        B
      + NU(S)
      + IN(W)

C15   V NO              = V N
      + CA(N,G,D,A)/      $ CL(..,10)
      + NU(S)
      ^ 2

C16   V NO              = V N             V END
      + CA(D)             $ CL(..,11)     $ TY(EN)
      + NU(P)                             B
      ^ 2

C17   V NP              = V NO
      + PS(3)             $ NU(P)
      $ 2.1NU
      ^ 2

C18   V PRPH            = V PREP          V NP
      ^ 2,3               . 3.1GC         $ CA
                                          $ GD

C19   V NP              = V DET           V ADJ        V NØ
      + PS(3)             $ GD/           . 4.1GD/     $ GD
      $*2.4GD/            $ NU/           . 4.2NU/     $ NU/
      $*2.5NU/            $ CA            . 4.3CA      $ CA
      $*2.6CA             . 3.1,W2.1/     $ IN
      ^ 4                 . 3.2,W2.2/
                          . 3.3,W2.3
                          .*3.4IN
```

```
C20  V NP            =   V NP           V PRPH
     ^ 2,3               $ TY(QU)       $ PR(54)
                                        $ NU(P)

C22  V CLS           =   V PRED         V NP        V PRFX
     + TY(INV)           $ PX           . 2.2PS/    . 2.1PX
     $ 3.4               $ PS/          . 2.3NU
     $ 2.5               $ NU           . 2.4TY
                         $ TS           ? PRN
                         $ TN           $ CA(N)
                         $ VC
                         $ MD
                         $ GC(λ)
```

The subscript PRN in the NP constituent is added to the clause label only if NP dominates a pronoun:

```
V NP       =     V PRN
+ PRN            $ TY(PS)
^ 2
```

```
C23  V CLS           =   V EXPLET       V CLS
     ^ 3                                $ TY(INV)
                                        * PRN
```

(This rule specifies that a clause with inverted word order may only be preceded by an expletive *es* if its subject is not a personal pronoun: *Es kommen drei Personen in Frage*. But: * *Es kommen sie in Frage*.)

```
C24  V SNT           =   V CLS          V PRD
     $ 2.1               $ TY(DC)       B
```

This analysis may show the difficulties that have to be accounted for in the analysis of surface strings with context-free phrase structure rules. Apart from the problems of discontinuity of elements in the surface structure and of phrasal dictionary elements, the amount of information in lexical elements which is relevant for correct analysis and translation is extremely large. Almost every verb can have different readings (and translations) depending on which one of a (sometimes very large) number of selection restrictions or feature packets it is associated with. (Feature packets may include separable prefixes, case government including prepositional objects governed, types of objects and subjects required, etc.). For example, the German verb liegen may be associated with 30 different feature packets, resulting in 30 different readings of which a few are shown here (these translations, with a few exceptions, are taken from Wildhagen and Héraucourt, German-English / English-German Dictionary, Vol. II German-English, Brandstetter Verlag, Wiesbaden, 1957):

1. liegen, intransitive, requiring a physical object as subject, with a locative adverb: to lie, to rest, to be located or situated;
2. liegen, governing a dative object which must be human and with a subject which must be abstract: to suit a.p., to appeal to sb.;
3. liegen, associated with the separable prefix an, with an inanimate concrete subject, governing a dative object or a prepositional object with the preposition an and an NP which must be concrete and inanimate: border on, be adjacent to;
4. liegen, with the separable prefix an, with a human subject and a human dative object: to entreat a.p.;
5. liegen, with the separable prefix bei, intransitive, with a concrete inanimate subject, with the auxiliary sein if used in the perfect tense: to be enclosed;
6. liegen, with the separable prefix danieder, intransitive, and with a human subject: to be lying ill;
7. liegen, with the separable prefix vor, intransitive, and with an abstract subject: exist.

The subscript format, in which the rules for this analysis are written, makes surface analysis possible

because of the following two characteristics:
a) Rule constituents are only subconfigurations of work space configurations, i.e. only the features relevant in a particular rule are mentioned in that rule while all others are disregarded. For example, rule C13 (p. 3) only states the condition that a verb stem must be classified as belonging to the paradigmatic class 27 in order to be concatenable with the verb ending -t, thus forming a predicate with the indicated features. The remaining properties of the verb (prefix, case government, type of object and subject required) are irrelevant in this concatenation rule and are merely "carried up the structural tree" by means of the operation specified by the symbols $\wedge 2$ on the left side of that rule.
b) Agreement and government are specified as set theoretical operations between the values of rule constituents. For example, rule C19 (p. 3) very generally states that in a German sentence the sequence determiner-adjective-nominal should be analyzed as a noun phrase provided that they agree in gender, number and case, and that the adjective and the determiner must not agree in type of inflection (weak or strong). These conditions are expressed by the operations specified in the second and following lines of each constituent of this rule. (All other features of the nominal head are not specifically mentioned in the rule and are simply carried up the tree.) Thus, very large numbers of rules can be represented by one rule in this subscript format. This makes it possible to incorporate and refer to the large amount of information necessary for analysis and translation in the dictionary and syntax of a surface grammar. Access to this information available in the surface string would be practically impossible with a context-free phrase structure grammar with simple symbols because of the unmanageable number of lexical classes and morphological

and syntactic rules building on these classes.

In spite of the greater economy of subscript rules, however, problems resulting from permutations of elements of phrasal and idiomatic expressions cannot be easily solved in surface analysis. For this reason, the analysis of sentences containing such elements is, in practice, performed in two steps at the LRC: surface analysis and standard analysis. In standard analysis the elements of phrasal and idiomatic expressions are re-ordered to a pre-determined standard order and are then treated as one single dictionary item, possibly with internal variable slots. A detailed description of standard analysis may be found in <u>Research in German-English Machine Translation on Syntactic Level</u>, Final Technical Report, RADC-TR-69-368, Volume II, August 1970.

The following is an explanation of the symbols used in the structural tree. The symbols are defined going from left to right in the sentence and from the bottom to the top of the tree.

<u>Lexical level</u>:

EXPLET = Expletive <u>es</u>; not a pronoun but rather a syntactically empty placeholder for the subject of the sentence.

V
CL(27)
PX(25'...)/
GC(λ'...)/
TO(λ'...)/
TS(AB'...) = This <u>v</u>erb of paradigmatic <u>cl</u>ass <u>27</u> may be used with any of a number of specified separable prefixes, among them prefix 25, which is the German prefix <u>vor</u>. If it is used in conjunction with this particular prefix, it is intransitive (<u>g</u>overns <u>c</u>ase λ; semantic <u>t</u>ype of <u>o</u>bject λ) and takes a <u>s</u>ubject of the semantic class <u>t</u>ype <u>ab</u>stract.

END TY(T)	=	Ending of type -t
DET GD(F) CA(N,A) NU(S) IN(S)	=	Determiner, gender feminine, ambiguous with respect to case, i.e. it may be considered nominative or accusative, number singular, strongly inflected.
A CL(7)	=	Adjective of paradigmatic class 7.
END TY(E)	=	Ending of type -e
N CL(10) GD(F) TY(QU)	=	Noun of paradigmatic class 10, gender feminine, type quantifier, i.e. a quantifying noun which may be followed by a von PRPH and then constitutes a modifier of the head noun in that PRPH.
PREP PR(24) GC(D)	=	The preposition is identified as preposition number 24 (von) and has the feature "governs case dative".
N CL(11) GD(N) TY(AB+CN)	=	A noun of the paradigmatic class 11, gender neuter, and semantic type abstract and countable.
END TY(EN)	=	Ending of the type -en.
PRFX PX(25)	=	This prefix is identified as prefix number 25 (vor).
PRD	=	The period is marked as being a marginal symbol, i.e. it constitutes the boundary of a word and of a sentence.

Morphological level:

PRED PS(3'2)/ NU(S'P) TN(PR) VC(A) MD(I) PX(25'...)/ GC(λ'...)/ TO(λ'...)/ TS(AB'...)	=	The predicate (finite verb) has all the features of the underlying verb stem: prefix, case government, type of object and subject required. (The feature CL

- paradigmatic class - is dropped because it is no longer relevant.) In addition, it has the features person and number which mark it as either 3rd person singular or 2nd person plural. (The apostrophe and slash establish this relation between the individual features.) It is also marked as: tense present, voice active, and mood indicative.

ADJ GD(M'F,N)/ CA(N'N,A) NU(S) IN(W)	=	With respect to gender and case, the inflected adjective is characterized as masculine nominative; or feminine or neuter nominative or accusative. In number it is singular; the inflection is weak.
NO GD(F) CA(N,G,D,A) NU(S) TY(QU)	=	The inflected nominal has the same gender and type information as the dictionary noun entry and in addition has the tags number singular, case 4-way ambiguous, i.e. it is either nominative, genitive, dative, or accusative, depending on its environment.
NO GD(N) CA(D) NU(P) TY(AB+CN)	=	Inflected nominal with the gender and type of the underlying noun stem, case dative, number plural.

Phrase level:

NP GD(F) CA(N,A) NU(S) TY(QU) PS(3)	=	The noun phrase has the gender, case, and number characteristics in which the underlying determiner, adjective and noun agree, namely feminine nominative or accusative singular; the type is that of the head noun; the NP is marked as 3rd person.

NP
GD(N)
CA(D)
NU(P)
TY(AB+CN)
PS(3) = Noun phrase with all syntactic and semantic features of the underlying nominal, identified as 3rd person.

PRPH
PR(24)
TY(AB+CN)
NU(P) = This prepositional phrase is identified as dominating preposition 24, i.e. von, a d an NP with a head noun of type abstract and countable, number plural.

NP
GD(F)
CA(N,A)
NU(S)
TY(AB+CN)
PS(3) = This noun phrase, which dominates an NP followed by a von PRPH, has the syntactic features of the dominated NP: gender feminine, case nominative or accusative, number singular, and the semantic features of the head noun of the dominated PRPH: type abstract and countable. It is also marked as an NP in the 3rd person.

Clause and sentence level:

CLS
TY(INV)
TN(PR) = This clause is of the type with inverted word order; it may be followed by a "?" to form a question or, as in this sentence, it may be preceded by an expletive es to form a declarative sentence; its tense is present.

CLS
TY(DC)
TN(PR) = A clause of type declarative, tense present.

SNT
TY(DC) = A sentence of type declarative.

LEXICAL FEATURES IN TRANSLATION AND PARAPHRASING: AN EXPERIMENT

by

Rolf Stachowitz

Linguistics Research Center
The University of Texas at Austin

366a

LEXICAL FEATURES IN TRANSLATION AND PARAPHRASING: AN EXPERIMENT

I Introduction

It is obvious to any user of a monolingual dictionary that the meaning of a lexical item is not only dependent on the external form of the item but also on its syntactic or semo-syntactic properties.[1] The terms homonymy and polysemy reflect this knowledge. It is equally obvious for the user of a better than average bilingual dictionary that the meaning of a lexical item is also a function of each selection restriction associated with it. This observation is evident from the fact that different translations are associated with a particular lexical item dependent on the syntactic and/or semantic properties of the constituents in its environment. The verb *erinnern* provides an example for German: In the environment "reflexive pronoun" its translation is *remember*; in the environment "non-reflexive object" its translation is *remind*.

The observations are, of course, true for lexical items in a language independent of their translatability into some other language. Only a few monolingual dictionaries, however, make this observation explicit. Among the few notable examples are the German Woerterbuch der deutschen Gegenwartssprache[2] and Hornby's An Advanced Learner's Dictionary[3]. Hornby lists for each verb the complement structures with which it may occur and the meanings it has in each environment. Thus, *observe* may mean *to take notice of (to watch)* or *to say as comment* in the environment "*that* S", e.g. *He observed that his wife had arrived*. However, in the environment "NP", *observe* can only have the first interpretation, e.g. *He observed the arrival of his wife*[4].

In view of the possibility of specifying the meaning of a lexical item or selecting a proper translation equivalent for it by taking its environment into account, it may seem surprising to the uninitiated that earlier MT systems had attempted to make such selections based on different criteria: considerations of the type of text to be translated or of probability of occurrences of lexical items. The difficulties confronting attempts to access the selection restrictions of a lexical item during the surface analysis of a sentence by means of a context-free grammar have been described in various monographs. These difficulties are multiplied when attempting the translation of languages, such as German, where various agreement and government relations hold between constituents, where lexical items and phrasal expressions often occur as discontinuous elements, and where sentence constituents can occur in various orders. The attempt to incorporate selection restrictions of lexical items into non-terminal symbols of context-free grammars would have increased the number of such rules to unmanageable proportions. For this reason, the incorporation of such selection restrictions was consequently suppressed. The loss was two-fold:

a) The number of syntactic interpretations for a sentence often increased ("forced readings").

b) The selection of proper translation equivalents had to be based on different criteria.

II Background of the Experiment

In summer 1966 I began investigating the possibilities of improving various parts of the Linguistics Research System[5] in order to cope with the increasing difficulties encountered in the attempts to analyze and translate sentences in natural language: the prohibitively large number of syntactic and translation rules necessary for the description and translation of surface structures into surface structures and the inability to deal with discontinuous constituents.[6] The research was influenced by the following guidelines:

1) to improve translation by permitting access to selection restrictions;

2) to decrease the number of forced readings assigned to sentences without an unreasonable increase in the number of grammar and translation rules;

3) to preserve as many as possible of the various algorithms used for surface analysis, translation mapping and surface production.

The results were reported in December 1966 in an unpublished paper which stated:

a) that vastly improved translations were possible by performing translation not from surface structures into surface structures but from standardized surface structures (standard strings) into standardized surface structures;

b) that these standard strings could be derived from the syntactic reading of a sentence by means of an additional straightforward algorithm;

c) that these translations could be obtained with an overall decrease of grammar rules;

d) that the core of the LRS algorithms could be retained;[7]

e) that non-trivial paraphrases could be performed over standard strings which were not possible over surface strings.

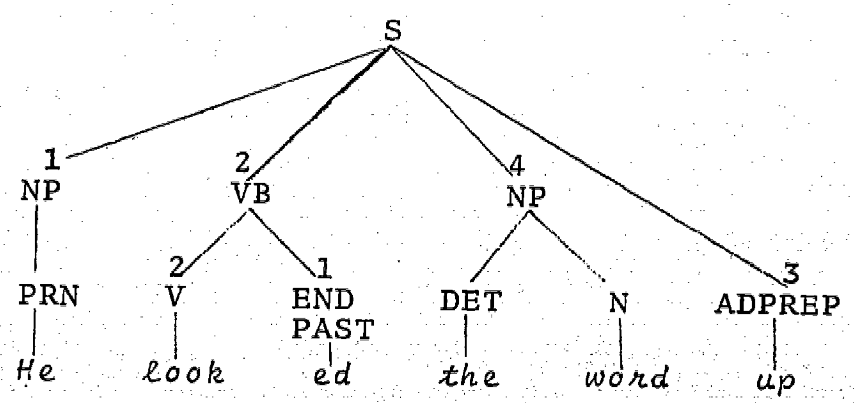
An experiment was subsequently performed to compare the proposed translation procedure with the established one. In order to facilitate this comparison, a text was selected for translation part of which had been translated in February 1966 using the Linguistics Research Center's first and second order translation system. Since the program which derived the standard strings from the corresponding sentence readings did not exist, the standard terminals were represented as surface terminals enclosed in asterisks. Only in cases where surface terminals occurred as homographs in the given text was a descriptor added in parentheses to reflect the disambiguating effect of the standardization procedure.

In order to reduce the time spent on this experiment, only one standard string of those sentences which had more than one surface reading was selected. (The number of readings for sentence 486 was 24, sentences 488, 489, and 492 had two readings each, all others had one.)

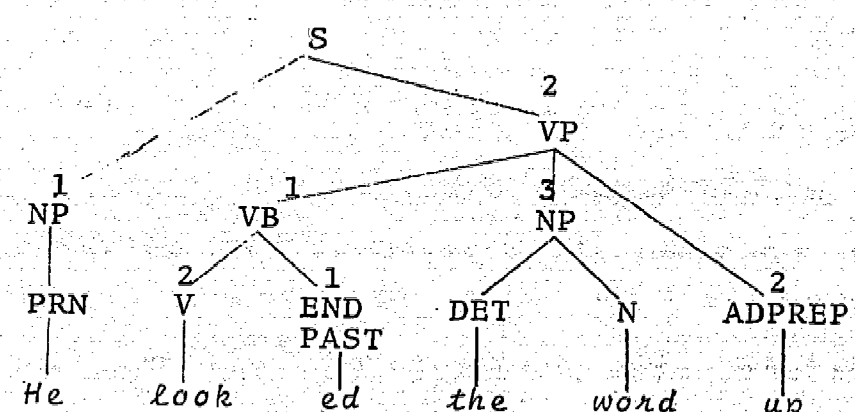
III Standard Strings

The standard representation of a sentence is a reordering of its terminal elements (with their part-of-speech interpretation) based on the surface interpretation of that sentence. The reordering could be performed by means of ordering instructions assigned to each constituent in the consequent of a rule which is part of the sentence reading.[8]

Assume the sentence *He looked the word up* is analyzed by the rules represented in the following tree diagram:



or, if you prefer, by



(The digits at the end of branches determine the mapping order of the sister nodes).

The standard string corresponding to this reading would then be:

he	*ed*	*look*	*up*	*the*	*word*
<PRN>	<END> <PAST>	<V>	<ADPREP>	<DET>	<N>

where the part-of-speech interpretation of each terminal is represented in angled brackets. (One can obtain a standard string by tracing down from each node, beginning with S, all branches in their indicated order and not tracing up a branch before all terminals below that branch have been reached).

The following standard order was defined for German surface constituents:

For clause level elements:

Subject (of an active sentence), agent adverbial (of a passive sentence), predicate, prefix, direct object, subject (of a passive sentence), predicative complement, indirect object, adverbials.

For phrase level elements:

Verbals: Finite verb, non-finite verb, prefix.

Noun phrases: Head, post-modifier, pre-modifier, determiner.

Prepositional phrases: Preposition, object.

For word level: Affixes, stem.

Conjoined elements "A, B and C": and , A B C .

The standard order defined for English differed from that for German only in that the elements of noun phrases occurred in the sequence: Determiner, pre-modifier, post-modifier, head of noun phrase. No significance is to be attributed to this difference; the distinction was made primarily to facilitate the reading of the output, the English standard strings. The distinction, however, shows the independence of the standard orders of the two languages.

The greater ease with which strings given in standard order could be analyzed may be evident when comparing the syntactic description of the following five sentences with the corresponding standard descriptions.

1) *Das Buch hat er seiner Frau gegeben.*
2) *Seiner Frau hat er das Buch gegeben.*
3) *Der Frau ist er gefolgt.*
4) *Seiner Frau hat er gehorcht.*
5) *Das Buch hat er gelesen.*

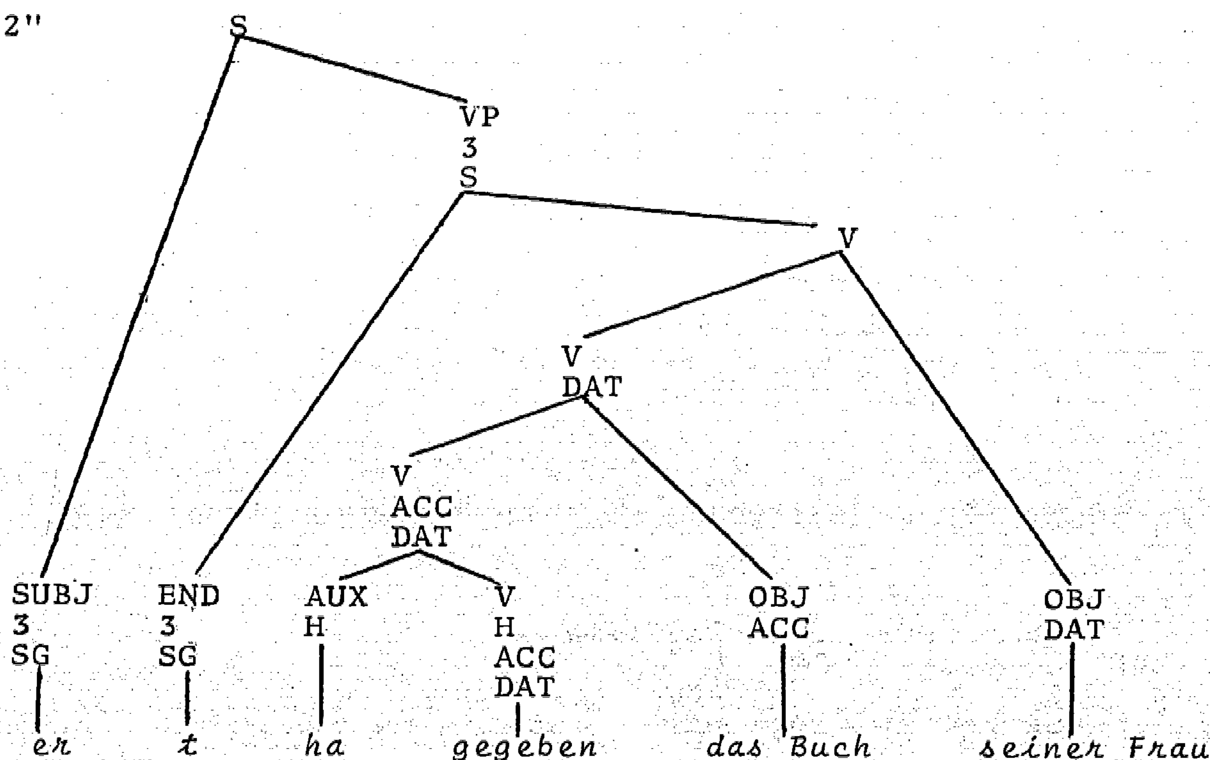
(Clause level constituents consisting of more than one word are underlined). These sentences were analyzed by the following rules:

1')	S →	OBJ ACC	AUX H 3 SG	SUBJ 3 SG	OBJ DAT	PASTPART[9] H ACC DAT
2')	S →	OBJ DAT	AUX H 3 SG	SUBJ 3 SG	OBJ ACC	PASTPART H ACC DAT
3')	S →	OBJ DAT	AUX S 3 SG	SUBJ 3 SG	PASTPART S DAT	

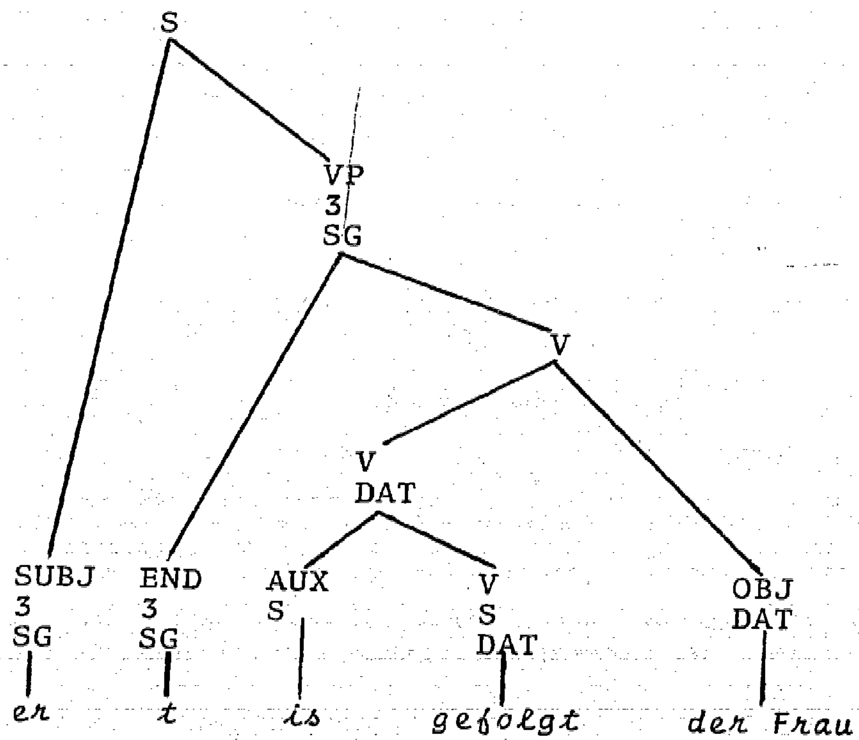
4')	S →	OBJ DAT	AUX H 3 SG	SUBJ 3 SG	PASTPART H DAT
5')	S →	OBJ ACC	AUX H 3 SG	SUBJ 3 SG	PASTPART H ACC

As we can observe, each change in word order (sentences 1 and 2), syntactic agreement (sentences 3 and 4) or government (sentences 4 and 5) had to be analyzed by a new sentence rule.[10] The corresponding standard representations, however, permitted a far more economic analysis.

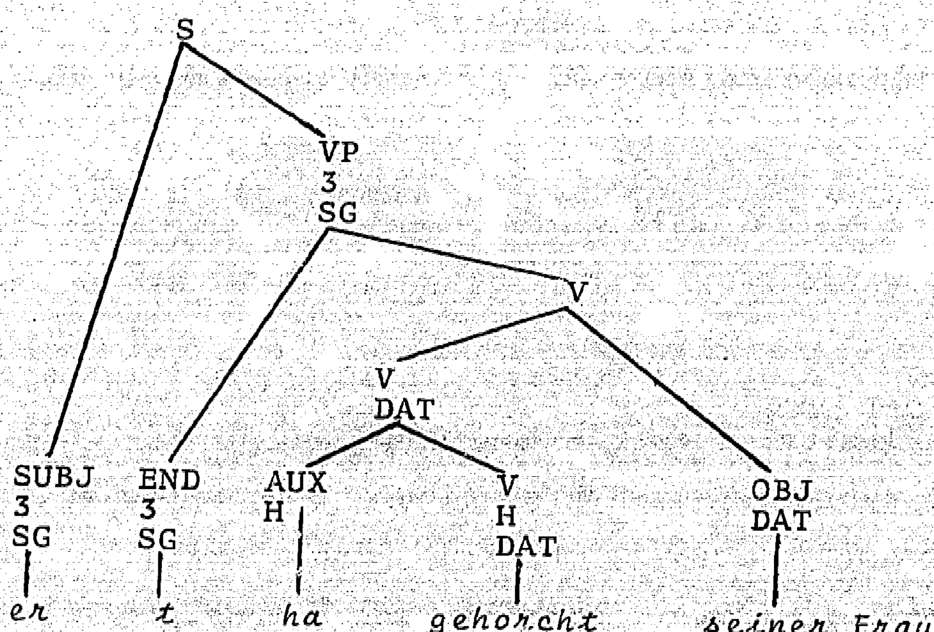
1'',2''



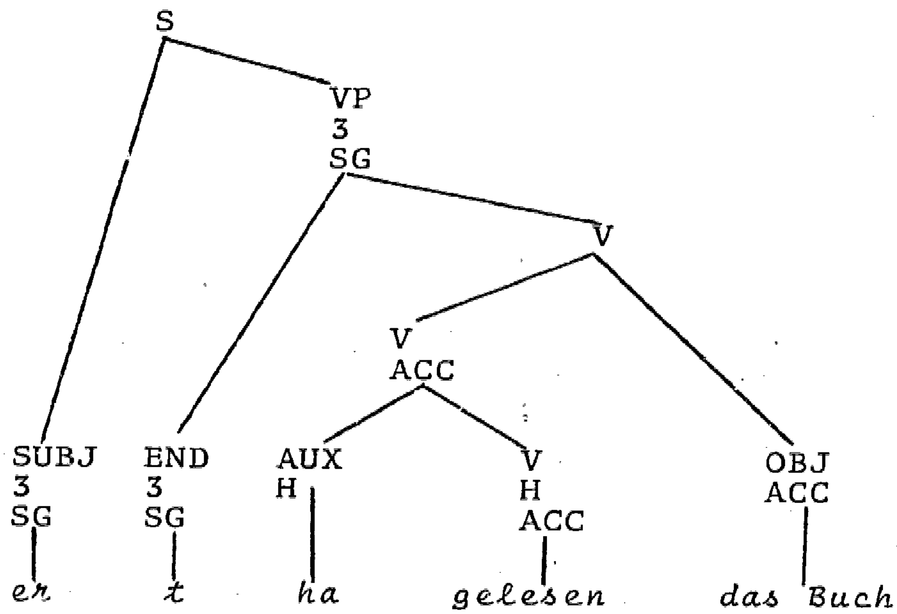
3''


S
VP
3
SG
V
V
DAT
SUBJ
3
SG
END
3
SG
AUX
S
V
S
DAT
OBJ
DAT
er
t
is
gefolgt
der Frau

4''


S
VP
3
SG
V
V
DAT
SUBJ
3
SG
END
3
SG
AUX
H
V
H
DAT
OBJ
DAT
er
t
ha
gehorcht
seiner Frau

5''



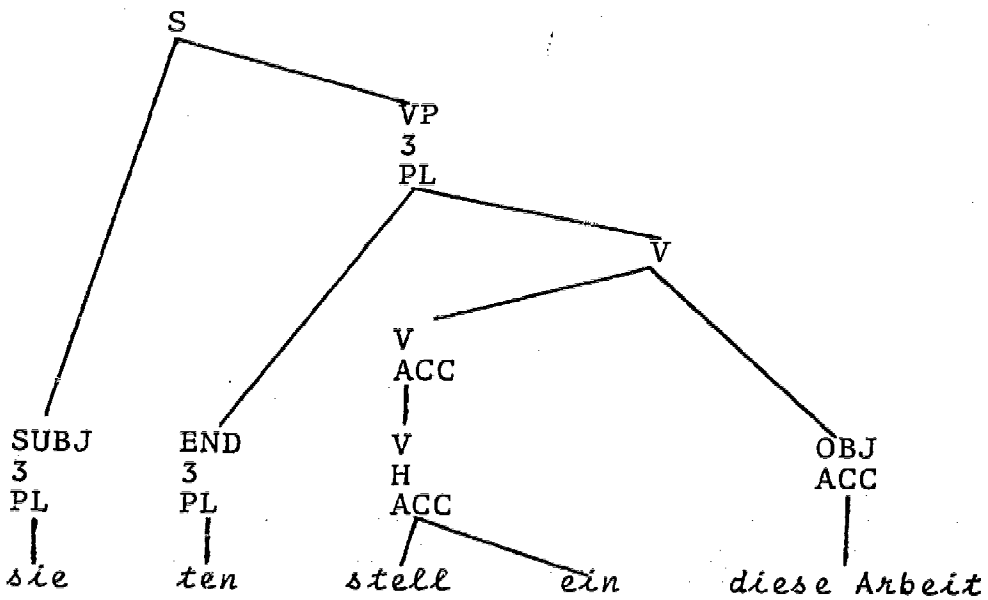
Firstly, it will be noticed that permutations as in sentences 1) and 2) were reduced to the same representation. Secondly, it was possible to concatenate the verb with its immediately contiguous elements, dropping with each concatenation the information that was necessary for the concatenation. This resulted in a considerably smaller number of grammar rules. Note that all four readings have in common the rules $\text{S} \rightarrow \begin{smallmatrix}\text{SUBJ}\\ 3\\ \text{SG}\end{smallmatrix}\ \begin{smallmatrix}\text{VP}\\ 3\\ \text{SG}\end{smallmatrix}$ and $\begin{smallmatrix}\text{VP}\\ 3\\ \text{SG}\end{smallmatrix} \rightarrow \begin{smallmatrix}\text{END}\\ 3\\ \text{SG}\end{smallmatrix}\ \text{V}$. Sentences 1), 3) and 4) also have in common the rule $\text{V} \rightarrow \begin{smallmatrix}\text{V}\\ \text{DAT}\end{smallmatrix}\ \begin{smallmatrix}\text{OBJ}\\ \text{DAT}\end{smallmatrix}$.[11] It was, finally, possible to treat discontinuous lexical items as one piece and assign them a new, their correct, syntactic interpretation.[12] Thus the rule $\text{S} \rightarrow \text{OBJ(4)}\ \begin{smallmatrix}\text{PRED(2)}\\ \text{ACTIVE}\end{smallmatrix}\ \text{SUBJ(1)}\ \text{PRFX(3)}$ - the desired order of the constituents is given in parentheses - interpreting sentences such as

6) *Diese Arbeit stellten sie ein = They discontinued this work.*

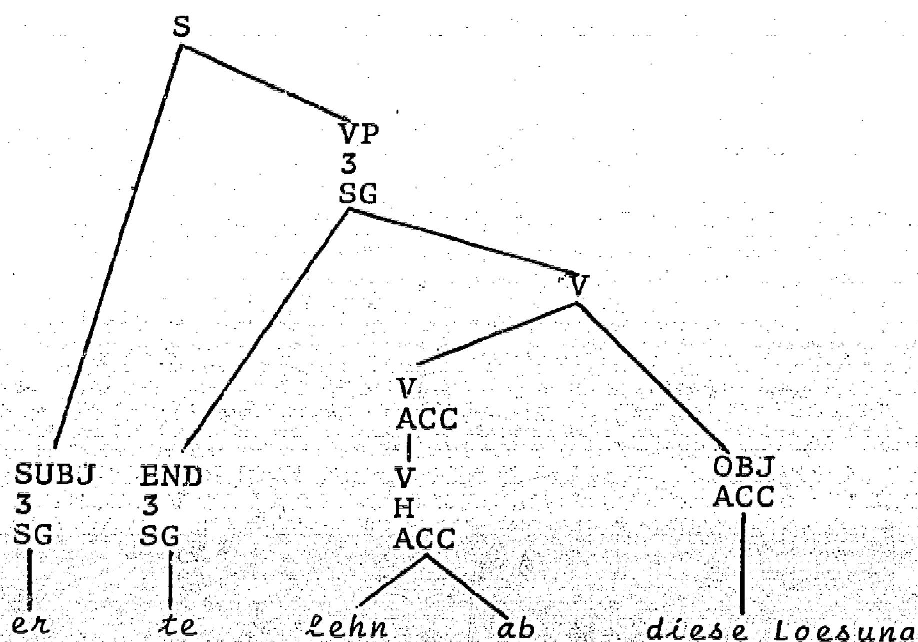
7) *Diese Loesung lehnte er ab = He rejected this solution.*

generated the standard strings given in the tree diagrams below.[13]

6''

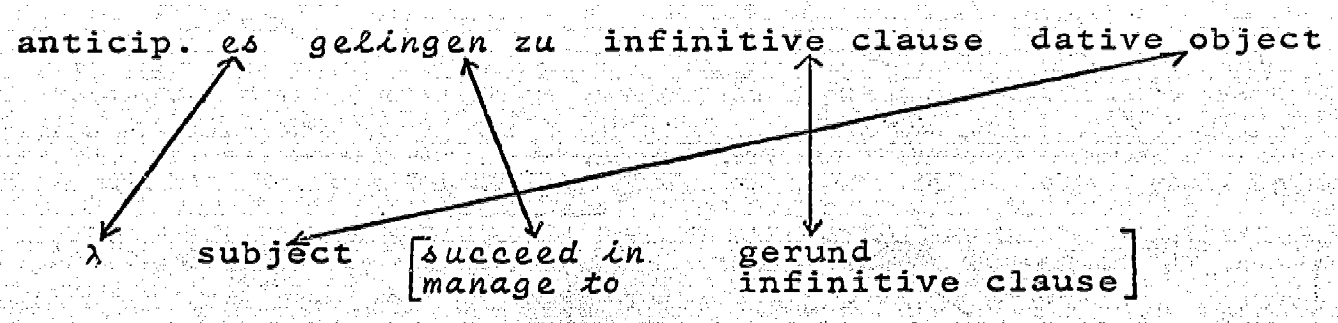


7''



IV The Selection of Translation Equivalents

The possibility of associating more comprehensive syntactic information with lexical pieces in standard strings as a consequence permitted an improved selection of translation equivalents. The list in Figures 7-1 through 7-6 contains a number of German items with their selection restrictions and the particular translations associated with each selection restriction. The lexical items are listed in the order in which they occur in the translated text. The selection restrictions which apply to the text are given a check mark. No semo-syntactic features, like HU, AN, AB (human, animate, abstract) were taken into account when performing the translation; for those features, cf. my appended paper "Requirements for Machine Translation: Problems, Solutions, Prospects."
The translation possibilities which resulted from the performed subclassification are indicated by light broken lines; the ones selected, by heavy underlines.[14] Of particular interest is one of the translations for *gelingen* (sentence 494, Figure 7), which permitted the mapping represented by the following diagram.



"*Breit* + unit of measure" could be mapped into "*wide* + unit of measure" or "unit of measure + *in width*", *Zuordnung zu* into *relation to* or *connection with*. The noun phrase *lange Zeit* could be recognized as an adverbial of extension in time instead of as an object due to the feature TIM.

V Paraphrases

In order to show the variety of translations or paraphrases possible over standard strings, a number of non-ad-hoc systematic synonymy relationships were defined for English resulting in the paraphrases given in Figures 3 and 4. Synonymy relationships were defined between lexical pieces and between syntactic structures. Examples of the latter are the active : passive transformation, the perfect tense : past tense transformation[15] and the noun-pre-modifier : noun-post-modifier transformation. Trivial examples of lexical paraphrases were simple synonymy substitutions like *get* : *obtain*, *prominence* : *protuberance*, or *circle* : *ring*; less trivial examples were *lunar* : *moon*, *solar* : *sun*, *luminous* : *light*, *bright* :*(to) shine*, *manage to* (+infinitive) : *succeed in* (+gerund). The effect of the syntactic classification of lexical items which had been defined as synonymous resulted in a selection of only those syntactic superstructures which interpreted them. Thus syntactic superstructures which were interpreted by the same normal form expression (translation term) but which could not form a well-formed tree with the selected lexical items were filtered out during the production phase.[16] The effect of this filtering function is shown for two examples in Figure 6; the sequence of normal form expressions S108, S100, S108, S104, L176, S104, L125 (to be read from top to bottom, left to right) simultaneously represents the four paraphrases *the solar disk, the disk of the sun, the sun's disk, the sun disk.*[17]

VI Translations

The simulated standard representation of the German original text (Figure 1) is given in Figure 2. The computer output, the mechanical translations, is shown in Figures 4-1 through 4-9. The translations in Figures 5-1 through 5-3 show an approximation to English normal word order. A more precise rendering would have required a separate processing stage, a rearrangement part. This stage seemed unnecessary for the purpose of the experiment since it is a simple reversal of the generation of standard strings from surface strings. A surface representation of the English translations of the German corpus is given in Figures 3-1 through 3-2.

The translation was performed using some of the then existing LRC analysis and translation algorithms. These, in order to speed up the actual processing time, stored in core all readings found. Whenever the number of readings exceeded the space allotted for them, certain readings were irretrievably dropped. If those readings were needed during the production phase, the corresponding German lexical or syntactic structures were used instead. This effect is noticeable in the occurrence of asterisked items in the English translations (also items, given in script in Figure 3), in the occurrence of the German standard order in noun phrases,[18] which is different from the defined English standard order, or simply in the ungrammaticality of the generated sentence.

VII Conclusion

In spite of the improved translation capabilities through translation over standard structures, the number of rules necessary, using context-free grammars with simple vocabulary symbols, was felt to be unnecessarily high. The changes made to remedy this deficiency are described in Lehmann/Stachowitz 1970, Vol. II.

FOOTNOTES

1 Thus the meaning of the noun *man* is different from that of the verb *man*, the meaning of the 'non-human' noun *conductor* different from that of the 'human' noun.

2 Woerterbuch der deutschen Gegenwartssprache, herausgegeben von Ruth Klappenbach und Wolfgang Steinitz, Akademieverlag Berlin, 1968 ff.

3 An Advanced Learner's Dictionary by Hornby, Gatenby and Wakefield, London, Oxford University Press, 1948.

4 This nominalization of the *that*-clause can be interpreted as a counterexample to various claims:
1) The combined claim that transformations are meaning-preserving and nominalizations are derived transformationally from sentences;
2) that semantic interpretations apply to deep structures before non-lexical transformations have applied.
Other verbs which behave like *observe* are *remark* and *notice*. Note that *watch* cannot occur in the environment "*that* S".

5 A comprehensive statement on the algorithms of the Linguistics Research System as used until May 1968 is given in Chapter VIII of Final Report, Linguistic Information Processing Study, DA 36-039 AMC-2162(E), 1 May 1965 - 30 April 1966; and Dynamic Adaptive Data Base Management Study, DA 28-043 AMC-02276(E), 16 May 1966 - 15 May 1967, The University of Texas, Linguistics Research Center, Austin, Texas, November 1968.

6 A comprehensive description of the problems encountered can be found in Lehmann/Stachowitz: Research in German-English Machine Translation on Syntactic Level, Vol. II, The University of Texas at Austin, August 1970.

7 Research performed during Spring of 1968 has led to the design of completely new analysis and translation algorithms which process context-free grammars with complex terminal and non-terminal symbols. Cf. Lehmann/Stachowitz 1970 and the appended paper "Requirements for Machine Translation: Problems, Solutions, Prospects."

8 Constituents in a rule consequent were assigned a predetermined order to permit the translation of sentences whose constituents could occur in different surface orders, e.g. *Mark bewunderten sie* = *Sie bewunderten Mark* = *They admired Mark*.

9 The LRC verb dictionaries only contained descriptors per-

taining to paradigmatic information. The verb constituents in those rules thus did not contain the descriptors pertaining to case government or auxiliary agreement information.

10 A trivial improvement for rules 1' and 2', resulting from the concatenation of the participle with the contiguous object before concatenating the new constituent with the other sentence constituents, was not possible in the earlier LRC system due to the ordering instructions attached to each constituent. Cf. Lehmann/Stachowitz, 1970, pp. T1 - T59.

11 The affixes are actually represented by "dummy" terminals; these are again replaced by the proper affixes during the output phase. Cf. Lehmann/Stachowitz 1970.

12 The translation of verb-prefix combinations, which occur discontinuously in German main clauses, would have required sentence rules in which the actual prefix would have had to be mentioned as a feature of the constituents involved. For example, *Diese Loesung schlug er vor (He proposed this solution)* would have had to be analyzed by a rule containing as constituents:

OBJ	PRED	SUBJ	PRFX.
ACC.	VOR	3	VOR
	ACC	SG	
	3		
	SG		

Each change of prefix would have required a new sentence rule, e.g. *Diese Loesung nahm er an (He accepted this solution)*:

OBJ	PRED	SUBJ	PRFX.
ACC	AN	3	AN
	ACC	SG	
	3		
	SG		

Such rules, of course, were never written.

13 Compare the translation equivalents *einstellen = suspend, ablehnen = refuse* in contrast to the translation of the corresponding simple verbs *stellen = put, lehnen = lean.*

14 In cases where the actually performed subclassification did not suffice to distinguish between different meanings of an item (e.g. *erhalten* with the readings *preserve, maintain* vs. *receive, obtain*), the translation given in the February 1966 translation was accepted. Cf. also footnote 21.

15 This paraphrase was defined to permit the translation of

the German perfect tense as in sentence 492 into both English present perfect and past tense.

16 One can interpret a sequence of normal form expressions as instructions to generate a tree by attaching the top node of a substructure to a non-terminal node of another structure, provided the respective labels are identical. The sequence of normal form expressions interpreting a tree thus imposes a well-formedness condition on the construction of all sentence trees with that normal form reading. Cf. also McCawley 1968.

17 The letter S stands for "non-lexical (syntactic) tree", the letter L for "lexical tree". The numbers were assigned in ascending order beginning with 100. These expressions can, of course, be replaced by meaningful expressions which can be interpreted as the vocabulary symbols of an interlingua or universal grammar.

18 The English subject *B. Edlen* in sentence 494 corresponding to the German dative object appeared in the position for "indirect object" whenever a necessary structure was dropped.

19 Figure 3: Only the paraphrases given in Figures 4-1 through 4-7 are given here. The items in script do not occur in any translation; the items in parentheses were provided as optional translations. The repeated "optionality" of *the* is due to the fact that it was not provided as a lexical equivalent of German /der/ but supplied by means of a syntactic normal form expression which should have been based on the non-encoded information that some nouns may optionally occur without *the*, like *earth, the earth*. The equivalents *completely, wholly, entirely, very* were not subclassified for adjective vs. participle modification (sentence 486). *Luminous corona* (sentence 492) results from an incorrect rule.

20 Figure 7: This translation, not given in any dictionaries, was provided in the February 66 translation.

21 The selection of the correct translation equivalent for this pattern depends on the understanding of the sentence.

Bibliography

Bech, Gunnar, Studien ueber das deutsche Verbum Infinitum, Det Kongelige Danske Videnskabernes Selskab. Dan. Hist. Filol. Medd. 35, no.2. Copenhagen, 1955; 36, no.6, 1957.

Bierwisch, Manfred, Grammatik des deutschen Verbs, Studia Grammatica II, Akademie Verlag, Berlin, 1963.

Chomsky, Noam, Aspects of the Theory of Syntax, M.I.T. Press, Cambridge, 1965.

Chomsky, Noam, Syntactic Structures, Mouton, The Hague, 1957.

Gruber, Jeffrey S., Studies in Lexical Relations, M.I.T., Cambridge, September 1965.

Harris, Zellig S., String Analysis of Sentence Structure, Mouton & Co., The Hague, 1962.

Harris, Zellig S., "Transformational Theory", Language, 41, No.3, 1965.

Hornby, A.S., A Guide to Patterns and Usage in English, Oxford University Press, London, 1960.

McCawley, James D., "Concerning the Base Component of a Transformational Grammar", Foundations of Language, Volume 3, No.3, August 1968.

Messinger, Heinz, Langenscheidts Handwoerterbuch Deutsch-Englisch, Langenscheidt KG, Berlin-Schoeneberg, 1960.

Postal, P., Constituent Structure - A Study of Contemporary Models of Syntactic Structure, Publications of the Research Center in Anthropology, Folklore, and Linguistics, Indiana University, Bloomington, 1964.

Tesnière, Lucien, Éléments de Syntaxe Structurale, Librairie C. Klincksieck, Paris, 1966 (deuxième édition revue et corrigée).

Wildhagen, Karl and Will Héraucourt, English-German German-English Dictionary, Vol.II German-English, Brandstetter Verlag, Wiesbaden, 1953.

GERMAN CORPUS

999,487
DIE LINIEN DES WASSERSTOFFS, DES HELIUMS UND VIELER METALLE TRETEN HIER AUF.

999,488
WENN DIE MONDSCHEIBE DIE SONNE GANZ VERDECKT, ERSCHEINT EIN ROTER 10 -- 15 BOGENSEKUNDEN BREITER RING UM DIE SONNE.

999,489
DAS IST DIE CHROMOSPHAERE MIT DEN PROTUBERANZEN.

999,490
WEITER AUSSEN SCHLIESST ALS SILBERWEISSER LICHTSCHWACHER SAUM DIE SONNENKORONA AN.

999,491
IN DER CHROMOSPHAERE FINDET MAN HAUPTSAECHLICH WASSERSTOFF-, HELIUM- UND KALZIUMLINIEN, ABER AUCH SPEKTRALLINIEN ANDERER METALLE.

999,492
IM LICHTE DER KORONA SIND MEHRERE HELLE SPEKTRALLINIEN AUFGEFUNDEN WORDEN, DEREN ZUORDNUNG ZU BEKANNTEN ELEMENTEN LANGE ZEIT UNBEKANNT BLIEB.

999,494
ERST IM JAHRE 1941 GELANG ES B. EDLEN IN UPSALA DIESE SPEKTRALLINIEN IN GEEIGNETEN IRDISCHEN LICHTQUELLEN ZU ERHALTEN.

999,486
DIE HELLEN LINIEN DER DAMPFFOERMIGEN SONNENATMOSPHAERE KANN MAN IN DER SOGENANNTEN UMKEHRENDEN SCHICHT, EINER SCHMALEN DAMPFHUELLE OBERHALB DER AEUSSEREN SONNENBEGRENZUNG, DER PHOTOSPHAERE, FUER EINIGE WENIGE AUGENBLICKE BEOBACHTEN, WENN BEI EINER SONNENFINSTERNIS DER FORTSCHREITENDE MOND GERADE EBEN NOCH EINEN GANZ SCHMALEN RAND DER SONNENOBERFLAECHE AUF DER EINEN SEITE FREI LAESST (SOG. FLASHSPEKTRUM).

Fig. 1

German Standard Strings

```
  JOB PROOF G-TXT RETRIEVAL OF 30.JANUARY.'67
PAGE 1

0001 999,486,RST,011867
0002     *MAN* ** *KANN* *EN* *BEOBACHT* *N* *LINIE* *ATMOSPHAERE* *N*
0003 *SONNE* ** *EN* *DAMPFFOERMIG* ** *DER* ** *EN* *HELL* ** *DIE*
0004 *IN* *SCHICHT* *,* *HUELLE* *DAMPF* ** *OBERHALB* *BEGRENZUNG*
0005 *N* *SONNE* ** *EN* *AEUSSER* ** *DER* ** *,* *PHOTOSPHAERE* *DER*
0006 *,* *EN* *SCHMAL* ** *EINER* *EN* *END* *KEHR* *UM*(PFX) *EN*
0007 *SOGENANNT* ** *DER* ** *FUER* *E* *BLICK* *N* *AUGE* ** *E*
0008 *WENIG* *E* *EINIG* ** ** *WENN* *MOND* *E* *END* *SCHREIT* *FORT*
0009 ** *DER* ** *T* *LAESS* *FREI* *RAND* *OBERFLAECHE* *N* *SONNE* **
0010 *DER* ** *EN* *SCHMAL* *GANZ* ** *EINEN* *AUF*(PP) *SEITE* *EN*
0011 *EIN* ** *DER* ** *NOCH* *EBEN* *GERADE* ** *BEI* *FINSTERNIS* *N*
0012 *SONNE* ** *EINER* ** *(* *UM*(FLX) *SPEKTR* *FLASH* ** *SOG.* **
0013 *)* ** *,* ** *.*

0001 999,487,RST,011867
0002     *N* *LINIE* *UND* *,* *S* *WASSERSTOFF* *DES* *S* *HELIUM*
0003 *DES* *E* *METALL* *ER* *VIEL* ** ** *DIE* ** *EN* *TRET* *AUF*
0004 *HIER* ** ** *.*

0001 999,488,RST,011867
0002     *RING* *,* *ER* *BREIT* *N* *SEKUNDE* *BOGEN* ** *--* *10*
0003 *15* ** *ER* *ROT* ** *EIN* ** *T* *ERSCHEIN* *UM* *SONNE* *DIE*
0004 ** *WENN* *SCHEIBE* *MOND* ** *DIE* ** *T* *VERDECK* *SONNE*
0005 *DIE* *GANZ* ** ** *,* ** *.*

0001 999,489,RST,011867
0002     *DAS*(D) ** *IST* *CHROMOSPHAERE* *DIE* *MIT* *EN* *PROTUBERANZ*
0003 *DEN* ** ** *.*

0001 999,490,RST,011867
0002     *A* *KORON* *N* *SONNE* ** *DIE* ** *T* *SCHLIESS* *AN* *ALS*
0003 *SAUM* *ER* *LICHTSCHWACH* *ER* *SILBERWEISS* ** *AUSSEN* *ER*
0004 *WEIT* ** ** *.*

0001 999,491,RST,011867
0002     *MAN* ** *ET* *FIND* *, ABER AUCH* *N* *LINIE* *UND* *,* *-*
0003 *WASSERSTOFF* *-* *HELIUM* *KALZIUM* ** *HAUPTSAECHLICH* *N* *LINIE*
0004 *AL* *SPEKTR* *E* *METALL* *ER* *ANDER* ** ** *IN* *CHROMOSPHAERE*
0005 *DER* ** ** *.*

0001 999,492,RST,011867
0002     ** *SIND* *WORDEN* *GE* *EN* *FUND* *AUF* *N* *LINIE* *AL*
0003 *SPEKTR* *E* *HELL* *E* *MEHRER* ** *I* *E* *LICHT* *A* *KORON*
0004 *DER* ** *M* ** ** *ZUORDNUNG* *ZU* *EN* *ELEMENT* *EN* *BEKANNT*
0005 ** ** *DEREN* ** *BLIEB* *BEKANNT* *UN* ** *ZEIT* *E* *LANG* **
0006 ** *,* *.*

0001 999,494,RST,011867
0002       *ES* ** *GELANG* *ZU*I *EN* *ERHALT* *N* *LINIE* *AL* *SPEKTR*
0003 *DIESE* *IN* *N* *QUELLE* *LICHT* ** *EN* *ISCH* *IRD* *EN* *GE* *ET*
0004 *EIGN* ** ** ** *B. EDLEN* *IN* *UPSALA* ** *I* *E* *JAHR* *1941*
0005 *M* *ERST* ** ** *.*
```

Fig. 2

Figure 3-1

English Paraphrases of German Corpus in Surface Representation[19]

487 Lines of (the) hydrogen, (the) helium and many
metals occur here.
appear

488 When (the) lunar disk hides (the)
disk of moon covers

completely
sun wholly , a red ring 10 to
entirely circle

15 arc seconds in width appears around
seconds of arc wide

(the) sun.

489 This is (the) chromosphere with (the) prominences .
protuberances

490 (The) corona of (the) sun follows a silvery white
solar corona

dim border farther out.
boundary

Above all
491 Mainly hydrogen's, helium's and calcium's
Chiefly hydrogen, helium and calcium

lines, but also other metals' spectrum lines
spectral
spectral lines of other metals

are found in (the) chromosphere.

One finds ... in (the) chromosphere.

492 Several bright spectral lines were found
shining spectrum discovered
have been found

corona lights
in (the) light of (the) corona of which the
luminous corona whose

relationship to known elements remained unknown
connection with

(for) a long time.

Figure 3-2

494 Only in / Not before / Not until — 1941 did B. Edlen in Upsala

succeed in getting / manage to get / manage to obtain — these spectral lines in

suitable terrestrial / earth — luminous / light — sources.

B. Edlen in Upsala managed to obtain / managed to get / succeeded in obtaining / succeeded in getting

these .... only in / not before / not until — 1941.

486 *One* can observe *the* bright / shining — lines *of the* vaporous

sun / solar — atmosphere in *the* so-called reversing layer,

a completely / wholly / entirely / very — narrow / thin — vaporous coat / veil / envelope

above / beyond — the outer solar border / boundary — , the

photosphere, for a few moments when the advancing moon just

barely leaves visible a very thin / narrow

solar surface edge / edge *of* the solar surface — on one side during

a darkness of the sun / an eclipse of the sun / a solar darkness / a sun eclipse — , the so-called

flash spectrum / spectrum of flash.

```
99487001      OF HYDROGEN, HELIUM, AND MANY S METAL ES LIN OCCUR HERE .
99487001      OF HYDROGEN, HELIUM, AND MANY S METAL ES LIN APPEAR HERE .
99487001      OF THE HYDROGEN, HELIUM, AND MANY S METAL ES LIN OCCUR HERE .
```

```
99487001      OF THE HYDROGEN, HELIUM, AND MANY S METAL ES LIN APPEAR HERE .
99487001      OF THE HYDROGEN, THE HELIUM, AND MANY S METAL ES LIN APPEAR
99487002    HERE .
99487001      OF HYDROGEN, HELIUM, AND MANY S METAL ES LIN  OCCUR HERE ** *.*
99487001      OF THE HYDROGEN, HELIUM, AND MANY S METAL ES LIN OCCUR HERE *.*
99487001      OF HYDROGEN, HELIUM, AND MANY S METAL ES LIN *DIE*  OCCUR  HERE
99487002    ** ** *.*
99487001      OF THE HYDROGEN, HELIUM, AND MANY S METAL ES LIN  APPEAR  HERE
99487002    ** ** *.*
```

Fig. 4-1

```
99488001    A RED 10 TO 15 ARC S SECOND IN WIDTH E CIRCL S APPEAR AROUND
99488002   SUN WHEN AR LUN DISK S COVER SUN LY WHOL , .

99488001    A RED 10 TO 15 ARC S SECOND IN WIDTH E CIRCL S APPEAR AROUND
99488002   SUN WHEN AR LUN DISK ES HID SUN LY WHOL , .

99488001    A RED 10 TO 15 ARC S SECOND IN WIDTH E CIRCL S APPEAR AROUND
99488002   THE SUN WHEN AR LUN DISK S COVER SUN LY WHOL , .

99488001    A RED 10 TO 15 ARC S SECOND IN WIDTH E CIRCL S APPEAR AROUND
99488002   THE SUN WHEN AR LUN DISK ES HID SUN LY WHOL , .

99488001    A RED 10 TO 15 ARC S SECOND IN WIDTH RING  S APPEAR AROUND THE
99488002   SUN WHEN AR LUN DISK S COVER SUN LY WHOL , ** *.*

99488001    RED 10 TO 15 ARC S SECOND IN WIDTH S RING *EIN*  S APPEAR
99488002   AROUND THE SUN *WENN*  THE AR LUN DISK S COVER SUN LY WHOL *,* **
99488003   *.*

99488001    RED 10 TO 15 ARC S SECOND IN WIDTH S CIRCLE *EIN*  S APPEAR
99488002   AROUND SUN WHEN AR LUN DISK S COVER SUN ELY COMPLET , ** *.*

99488001    *RING*  RED 10 TO 15 OF ARC S SECOND IN WIDTH ** *EIN*  S APPEAR
99488002   AROUND THE SUN WHEN THE AR LUN DISK S COVER SUN LY WHOL , ** *.*

99488001    *RING* *,*  10 TO 15 OF ARC S SECOND E WID *ER* *ROT* **  A **
99488002   *T*  APPEAR *UM* *SONNE* *DIE* ** *WENN*  THE AR LUN DISK S COVER
99488003   THE SUN LY WHOL *,* ** *.*

99488001    *RING* *,*  10 TO 15 OF ARC S SECOND E WID  RED ** *EIN*  S
99488002   APPEAR AROUND THE SUN WHEN AR LUN DISK S COVER SUN ELY ENTIR , **
99488003   *.*

99488001    *RING* *,* *ER*  WID  ARC S SECOND  TO  10  15 ** *ER* *ROT* **
99488002   A ** *T* *ERSCHEIN* *UM* *SONNE* *DIE* ** *WENN*  THE AR LUN DISK
99488003   S COVER SUN ELY COMPLET *,* ** *.*

99488001    *RING* *,* *ER* *BREIT*  OF ARC SECOND *--*  10 *15*
99488002   ** *ER* *ROT* ** *EIN*  S APPEAR AROUND SUN WHEN AR LUN DISK ES
99488003   HID SUN ELY COMPLET , ** *.*

99488001    *RING* *,* *ER* *BREIT*  OF ARC SECOND *--*  10 *15*
99488002   ** *ER* *ROT* ** *EIN*  S APPEAR AROUND SUN WHEN AR LUN DISK ES
99488003   HID SUN ELY ENTIR , ** *.*

99488001    *RING* [illegible] *ER* *BREIT*  OF ARC SECOND *--* *10* *15* ** *ER*
99488002   *ROT* ** *EIN*  S APPEAR AROUND SUN *WENN*  OF MOON DISK ES HID
99488003   SUN ELY COMPLET *,* ** *.*

99488001    *RING* [illegible] *ER* *BREIT* *N* *SEKUNDE* *BOGEN* ** *--* *10* *15*
99488002   ** *ER* *ROT* ** *EIN*  S APPEAR AROUND SUN WHEN OF MOON DISK ES
99488003   HID SUN VERY , ** *.*
```

Fig. 4-2

99489001 THIS IS CHROMOSPHERE WITH S PROMINENCE .
99489001 THIS IS CHROMOSPHERE WITH S PROTUBERANCE .
99489001 THIS IS CHROMOSPHERE WITH S PROTUBERANCE .
99489001 THIS IS CHROMOSPHERE WITH S PROTUBERANCE .
99489001 THIS IS CHROMOSPHERE WITH S PROTUBERANCE .
99489001 THIS IS CHROMOSPHERE WITH S PROTUBERANCE .
99489001 THIS IS CHROMOSPHERE WITH S PROTUBERANCE .
99489001 THIS IS CHROMOSPHERE WITH S PROTUBERANCE .
99489001 THIS IS THE CHROMOSPHERE WITH S PROMINENCE .
99489001 THIS IS THE CHROMOSPHERE WITH S PROTUBERANCE .
99489001 THIS IS THE CHROMOSPHERE WITH S PROTUBERANCE .
99489001 THIS IS THE CHROMOSPHERE WITH S PROTUBERANCE .
99489001 THIS IS THE CHROMOSPHERE WITH S PROTUBERANCE .
99489001 THIS IS THE CHROMOSPHERE WITH S PROTUBERANCE .
99489001 THIS IS CHROMOSPHERE WITH S PROMINENCE *.*
99489001 THIS IS CHROMOSPHERE WITH S PROTUBERANCE *.*
99489001 THIS IS CHROMOSPHERE WITH THE S PROMINENCE .
99489001 THIS IS CHROMOSPHERE WITH THE S PROTUBERANCE .
99489001 THIS IS CHROMOSPHERE WITH THE S PROTUBERANCE .
99489001 THIS IS CHROMOSPHERE WITH THE S PROTUBERANCE .
99489001 THIS IS CHROMOSPHERE WITH THE S PROTUBERANCE .
99489001 THIS IS CHROMOSPHERE WITH THE S PROTUBERANCE .
99489001 THIS IS THE CHROMOSPHERE WITH S PROTUBERANCE .
99489001 THIS IS THE CHROMOSPHERE WITH S PROTUBERANCE .
99489001 ~~THIS IS THE CHROMOSPHERE WITH THE S PROMINENCE~~
99489001 ~~THIS IS THE CHROMOSPHERE WITH THE S PROTUBERANCE~~
99489001 THIS IS THE CHROMOSPHERE WITH THE S PROTUBERANCE .

Fig. 4-3

```
99490001     AR SOL CORONA S FOLLOW  AS A SILVERY E WHIT DIM BOUNDARY
99490002    FARTHER OUT .

99490001     THE AR SOL CORONA S FOLLOW  AS A SILVERY E WHIT DIM BOUNDARY
99490002    FARTHER OUT .

99490001     AR SOL CORONA S FOLLOW  AS A SILVERY E WHIT DIM BOUNDARY
99490002    FARTHER OUT *.*

99490001     THE AR SOL CORONA S FOLLOW  AS A SILVERY E WHIT DIM BOUNDARY
99490002    FARTHER OUT *.*

99490001     AR SOL CORONA  S FOLLOW  AS A SILVERY E WHIT DIM BORDER FARTHER
99490002    OUT ** *.*

99490001     OF THE SUN CORONA  S FOLLOW   AS A SILVERY E WHIT DIM BORDER
99490002    FARTHER OUT ** *.*

99490001     OF SUN CORONA S FOLLOW  AS A SILVERY E WHIT DIM BOUNDARY
99490002    FARTHER OUT *.*
```

Fig. 4-4

```
99491001     ARE FOUND LY CHIEF HYDROGEN, HELIUM, AND CALCIUM ES LIN, BUT
99491002   ALSO OTHER S' METAL AL SPECTR ES LIN IN CHROMOSPHERE .

99491001     ARE FOUND LY MAIN HYDROGEN, HELIUM, AND CALCIUM ES LIN, BUT
99491002   ALSO OTHER S' METAL AL SPECTR ES LIN IN CHROMOSPHERE .

99491001     ARE FOUND LY CHIEF HYDROGEN, HELIUM, AND CALCIUM ES LIN, BUT
99491002   ALSO OTHER S' METAL AL SPECTR ES LIN IN THE CHROMOSPHERE .

99491001     ARE FOUND LY MAIN HYDROGEN, HELIUM, AND CALCIUM ES LIN, BUT
99491002   ALSO OTHER S' METAL AL SPECTR ES LIN IN THE CHROMOSPHERE .

99491001     ARE FOUND LY CHIEF HYDROGEN, HELIUM, AND CALCIUM ES LIN, BUT
99491002   ALSO OTHER S' METAL AL SPECTR ES LIN IN CHROMOSPHERE *.*

99491001     ARE FOUND LY MAIN HYDROGEN, HELIUM, AND CALCIUM ES LIN, BUT
99491002   ALSO OTHER S' METAL AL SPECTR ES LIN IN CHROMOSPHERE *.*

99491001    *MAN*  S FIND LY MAIN HYDROGEN, HELIUM, AND CALCIUM ES LIN, BUT
99491002   ALSO OTHER S' METAL AL SPECTR ES LIN IN CHROMOSPHERE ** *.*

99491001    *MAN*  S FIND LY CHIEF HYDROGEN, HELIUM, AND CALCIUM ES LIN, BUT
99491002   ALSO OTHER S' METAL AL SPECTR ES LIN  IN CHROMOSPHERE ** ** *.*

99491001    *MAN*  S FIND  BUT ALSO  ES LIN  AND  HYDROGEN, HELIUM  S
99491002   CALCIUM  LY MAIN  OF OTHER S METAL AL SPECTR ES LIN  IN
99491003   CHROMOSPHERE ** *.*

99491001    *MAN*  S FIND  BUT ALSO  ES LIN  AND *,*  S' HYDROGEN  S'
99491002   HELIUM  S' CALCIUM  LY MAIN  OTHER S' METAL AL SPECTR ES LIN   IN
99491003   CHROMOSPHERE ** *.*

99491001    *MAN* ** *ET* *FIND*  BUT ALSO  ABOVE ALL HYDROGEN, HELIUM, AND
99491002   CALCIUM ES LIN  AL SPECTR E LIN  OTHER S METAL **  IN
99491003   CHROMOSPHERE ** ** *.*

99491001    *MAN* ** *ET* *FIND*  BUT ALSO *N* *LINIE*  AND  ( *-*
99491002   *WASSERSTOFF*  S HELIUM  S CALCIUM *HAUPTSAECHLICH*  OTHER S'
99491003   METAL UM SPECTR ES LIN   IN THE CHROMOSPHERE ** *.*
```

Fig. 4-5

```
99492001     WERE ED DISCOVER SEVERAL ING SHIN AL SPECTR ES LIN *I*  CORONA
99492002   S LIGHT *M* ** **  WHOSE RELATIONSHIP TO N KNOW S ELEMENT ED
99492003   REMAIN UN N KNOW FOR A LONG E TIM , *.*

99492001     WERE ED DISCOVER SEVERAL ING SHIN AL SPECTR ES LIN *I*  CORONA
99492002   S LIGHT *M* ** **  WHOSE CONNECTION WITH N KNOW S ELEMENT ED
99492003   REMAIN UN N KNOW FOR A LONG E TIM , *.*

99492001     WERE ED DISCOVER SEVERAL ING SHIN AL SPECTR ES LIN *I*  CORONA
99492002   S LIGHT *M* ** **  OF WHICH THE RELATIONSHIP TO N KNOW S ELEMENT
99492003   ED REMAIN UN N KNOW FOR A LONG E TIM , *.*

99492001     WERE ED DISCOVER SEVERAL ING SHIN AL SPECTR ES LIN *I*  CORONA
99492002   S LIGHT *M* ** **  OF WHICH THE CONNECTION WITH N KNOW S ELEMENT
99492003   ED REMAIN UN N KNOW FOR A LONG E TIM , *.*

99492001     WERE FOUND SEVERAL ING SHIN AL SPECTR ES LIN  IN  OUS LUMIN
99492002   CORONA ** *M* ** **  WHOSE RELATIONSHIP TO N KNOW S ELEMENT  ED
99492003   REMAIN UN N KNOW FOR A LONG E TIM ** *,* *.*

99492001     HAVE BEEN FOUND SEVERAL ING SHIN AL SPECTR ES LIN  IN  CORONA S
99492002   LIGHT *M* ** **  WHOSE CONNECTION WITH N KNOW S ELEMENT  ED
99492003   REMAIN UN N KNOW FOR A LONG E TIM ** *,* *.*

99492001     HAVE BEEN FOUND SEVERAL ING SHIN AL SPECTR ES LIN  IN  CORONA S
99492002   LIGHT *M* ** **  OF WHICH THE RELATIONSHIP TO N KNOW S ELEMENT
99492003   ED REMAIN UN N KNOW FOR A LONG E TIM ** *,* *.*

99492001     HAVE BEEN FOUND SEVERAL ING SHIN AL SPECTR ES LIN  IN  CORONA S
99492002   LIGHT *M* ** **  OF WHICH THE CONNECTION WITH N KNOW S ELEMENT
99492003   ED REMAIN UN N KNOW FOR A LONG E TIM ** *,* *.*

99492001     WERE ED DISCOVER SEVERAL BRIGHT AL SPECTR ES LIN IN OF CORONA
99492002   LIGHT **  CONNECTION WITH N KNOW S ELEMENT  WHOSE  ED REMAIN  N
99492003   KNOW *UN* **  A LONG E TIM ** *,* *.*

99492001     WERE ED DISCOVER SEVERAL BRIGHT AL SPECTR ES LIN IN OF CORONA
99492002   LIGHT **  RELATIONSHIP TO N KNOW S ELEMENT  WHOSE  ED REMAIN  N
99492003   KNOW [illegible] **  A LONG E TIM ** [illegible].*

99492001     [illegible]E ED DISCOVER  ES LIN  UM SPECTR  BRIGHT  SEVERAL  IN  OF
99492002   CORONA LIGHT ** **  RELATIONSHIP TO  ELEMENT  N KNOW **  WHOSE
99492003   ED REMAIN  N KNOW *UN* **  TIM  LONG ** *,* *.*

99492001     WERE ED DISCOVER SEVERAL BRIGHT AL SPECTR ES LIN IN THE OF
99492002   CORONA LIGHT **  RELATIONSHIP TO  N KNOW S ELEMENT **  WHOSE **
99492003   *BLIEB*  N KNOW *UN* **  FOR A LONG E TIM ** *,* *.*

99492001     WERE ED DISCOVER SEVERAL BRIGHT AL SPECTR ES LIN IN THE OF
99492002   CORONA LIGHT **  CONNECTION WITH  N KNOW S ELEMENT **  WHOSE **
99492003   *BLIEB*  N KNOW *UN* **  FOR A LONG E TIM ** *,* *.*

99492001     WERE ED DISCOVER SEVERAL BRIGHT AL SPECTR ES LIN IN OF THE
99492002   CORONA LIGHT **  CONNECTION WITH  N KNOW S ELEMENT **  WHOSE **
99492003   *BLIEB*  N KNOW *UN* **  FOR A LONG E TIM ** [illegible]*.*
```

Fig. 4-6

```
99494001    B. EDLEN DID MANAG TO GET THESE AL SPECTR ES LIN IN SUITABLE
99494002   IAL TERRESTR OUS LUMIN S SOURCE IN UPSALA NOT UNTIL 1941 .

99494001    B. EDLEN DID MANAG TO GET THESE AL SPECTR ES LIN IN SUITABLE
99494002   IAL TERRESTR OUS LUMIN S SOURCE IN UPSALA NOT BEFORE 1941 .

99494001    B. EDLEN DID MANAG TO OBTAIN THESE AL SPECTR ES LIN IN SUITABLE
99494002   IAL TERRESTR OUS LUMIN S SOURCE IN UPSALA NOT UNTIL 1941 .

99494001    B. EDLEN DID MANAG TO OBTAIN THESE AL SPECTR ES LIN IN SUITABLE
99494002   IAL TERRESTR OUS LUMIN S SOURCE IN UPSALA NOT BEFORE 1941 .

99494001    B. EDLEN DID MANAG TO GET THESE AL SPECTR ES LIN IN SUITABLE
99494002   IAL TERRESTR OUS LUMIN S SOURCE IN UPSALA ONLY IN 1941 .

99494001    B. EDLEN DID MANAG TO GET THESE AL SPECTR ES LIN IN SUITABLE
99494002   IAL TERRESTR OUS LUMIN S SOURCE IN UPSALA ONLY IN 1941 .

99494001    B. EDLEN DID SUCCEED IN TING GET THESE AL SPECTR ES LIN IN
99494002   SUITABLE IAL TERRESTR OUS LUMIN S SOURCE IN UPSALA NOT UNTIL 1941
99494003

99494001    B. EDLEN DID SUCCEED IN TING GET THESE AL SPECTR ES LIN IN
99494002   SUITABLE IAL TERRESTR OUS LUMIN S SOURCE IN UPSALA NOT BEFORE
99494003   1941 .


99494001    ED SUCCEED IN  ING OBTAIN  ES LIN  UM SPECTR  OBSERV  IN
99494002   SOURCE  LIGHT  EARTH  SUITABLE ** **  B. EDLEN  IN  UPSALA **
99494003   ONLY IN 1941 ** *.*


99494001    ED MANAG TO  ING OBTAIN  AL SPECTR ES LIN  OBSERV  IN SUITABLE
99494002   IAL TERRESTR LIGHT S SOURCE **  B. EDLEN  IN UPSALA  ONLY IN
99494003   1941 ** *.*
```

Fig. 4-7

```
PATHS 1,1..1,2..2,1..2,2..3,1..3,2......12,1..12,2

99486001    *MAN* **  CAN  E OBSERV  ES LIN  AR SOL S ATMOSPHERE *EN*  OUS
99486002    VAPOR ** *DER* **  ING SHIN *DIE*  IN  LAYER *,*  OUS VAPOR S
99486003    ENVELOPE  BEYOND THE OUTER AR SOL BOUNDARY **  , PHOTOSPHERE,
99486004    THIN  A  ING REVERS  SO-CALLED *DER* **   FOR A FEW S MOMENT
99486005    WHEN ING ADVANC MOON ES LEAV E VISIBL A VERY THIN AR SOL SURFACE
99486006    EDGE ON ONE E SID JUST LY BARE DURING A AR SOL DARKNESS ,
99486007    SO-CALLED FLASH UM SPECTR, , ** *.*

99486001    *MAN* **  CAN  E OBSERV  ES LIN  AR SOL S ATMOSPHERE *EN*  OUS
99486002    VAPOR ** *DER* **  ING SHIN *DIE*  IN  S' LAYER *,*  OUS VAPOR S
99486003    ENVELOPE  ABOVE THE OUTER AR SOL BOUNDARY **  , PHOTOSPHERE
99486004    NARROW  AN  ING REVERS  SO-CALLED *DER* **   FOR A FEW S MOMENT
99486005    WHEN ING ADVANC MOON ES LEAV E VISIBL A VERY THIN AR SOL SURFACE
99486006    EDGE ON ONE E SID JUST LY BARE DURING A AR SOL DARKNESS ,
99486007    SO-CALLED FLASH UM SPECTR , ** *.*

99486001    *MAN* **  CAN  E OBSERV  ES LIN  AR SOL S ATMOSPHERE [illegible]  OUS
99486002    VAPOR ** *DER* **  BRIGHT *DIE*  IN  LAYER *,*  OUS VAPOR S
99486003    ENVELOPE  BEYOND THE OUTER AR SOL BOUNDARY **  , PHOTOSPHERE,
99486004    THIN  A  ING REVERS  SO-CALLED *DER* **   FOR A FEW S MOMENT
99486005    *WENN*  ING ADVANC MOON  ES LEAV E VISIBL A VERY THIN OF AR SOL
99486006    SURFACE EDGE  ON* SID  ONE  JUST LY BARE  DURING A AR SOL
99486007    DARKNESS  , SO-CALLED OF FLASH UM SPECTR, ** *,* ** *.*

99486001    *MAN* **  IS *EN*  OBSERV *N* *LINIE*  ATMOSPHERE *N*  SOL **
99486002    *[illegible]  OUS VAPOR ** *DER* ** *EN*  SHIN ** *DIE* *IN*  S LAYER  ,
99486003    ENVELOPE  VAPOR **  BEYOND  OUTER AR SOL BOUNDARY *DER* **  ,
99486004    PHOTOSPHERE *,*  THIN *EINER* *EN* *END*  REVERS *EN*  SO-CALLED
99486005    *** *DER* **   FOR A FEW S MOMENT ** *WENN*  ING ADVANC MOON
99486006    *DER*  ES LEAV E VISIBL  VERY THIN OF AR SOL SURFACE EDGE
99486007    *EINEN*  ON *SEITE*  ONE  JUST LY BARE **  DURING A AR SOL
99486008    DARKNESS **  , THE SO-CALLED OF FLASH UM SPECTR, ** *,* ** *.*

99486001    *MAN* ** *KANN* *EN* *BEOBACHT* *N* *LINIE*  SUN S ATMOSPHERE
99486002    *EN* *DAMPFFOERMIG* ** *DER* ** *EN* *HELL* **  *DIE* *IN*
99486003    *SCHICHT* *,*  OUS VAPOR VEIL *OBERHALB*  OUTER AR SOL BORDER
99486004    *DER* ** [redacted] *EN* *SCHMAL* ** *EINER*   *EN*
99486005    SO-CALLED ** *DER* **   FOR A FEW S MOMENT ** *WENN*  LUN *E*
99486006    *END*  ADVANC ** *DER*  ES LEAV E VISIBL A ELY ENTIR NARROW OF AR
99486007    [redacted] *AUF*(PP) *SEITE* *EN*  ONE ** *DER* **   JUST
99486008    LY BARE  DURING AN OF SUN ECLIPSE *(*  SO-CALLED OF FLASH UM
99486009    SPECTR *)* ** *,* ** *.*

99486001    *MAN* ** *KANN* *EN* *BEOBACHT* *N* *LINIE* *ATMOSPHAERE* *N* *S-
99486002    ONNE* ** [illegible] *DAMPFFOERMIG* ** *DER* ** *EN* *HELL* **
99486003    *DIE* *IN* *SCHICHT* *,* *HUELLE* *DAMPF* ** *OBERHALB*  AR SOL
99486004    BOUNDARY *EN* *AEUSSER* ** *DER* ** *,* *PHOTOSPHAERE* *DER*
99486005    *,* *EN* *SCHMAL* ** *EINER* *EN* *END*   *EN* *SOGENANNT* **
99486006    *DER* ** *FUER* *E*  MOMENT  A FEW ** *WENN* *MOND* *E*
99486007    *END* *SCHREIT* *FORT* ** *DER* ** *T* *LAESS* *FREI*  OF AR SOL
99486008    SURFACE EDGE  THIN  WHOL ** *EINEN* *AUF*(PP) *SEITE* *EN* *EIN*
99486009    *BEI*  AR SOL ECLIPSE *EINER* **  ,  OF FLASH UM SPECTR
99486010    SO-CALLED ** *)* ** *,* ** *.*
```

Fig. 4-8

```
PATHS 1,0,0,1-2,C,0,1---12,0,0,1-1,0,0,2-2,0,0,2---12,0,0,2

99486001   *MAN* **  IS  E OBSERV  ES LIN  ATMOSPHERE  SUN [illegible] OUS VAPOR
99486002   ** *DER* **  BRIGHT *DIE*  IN  LAYER  ,  ENVELOPE  OUS VAPOR
99486003   BEYOND  BORDER  SUN  OUTER *DER* **  , PHOTOSPHERE,  THIN  A  ING
99486004   REVERS  SO-CALLED *DER* **   FOR A FEW S MOMENT *WENN*  THE ING
99486005   ADVANC MOON  ES LEAV E VISIBL  EDGE  SURFACE  SUN *DER* **  THIN
99486006   COMPLET **  A  ON' SID  ONE  JUST LY BARE **  DURING  ECLIPSE
99486007   SUN  A **  , THE SO-CALLED OF FLASH UM SPECTR, ** *,* ** *.*

99486001   *MAN* ** *KANN* *EN*  OBSERV *N*  LIN  ATMOSPHERE *N*  SUN **
99486002   VAPOR ** *DER* **  BRIGHT *DIE*  IN  LAYER  (  COAT  OUS VAPOR
99486003   ABOVE  BOUNDARY  THE SUN  OUTER *DER* **  , THE PHOTOSPHERE,
99486004   NARROW  AN  ING REVERS  SO-CALLED *DER* **   FOR A FEW S MOMENT
99486005   *WENN*  ING ADVANC MOON  ES LEAV E VISIBL  EDGE  SURFACE  THE SUN
99486006   *DER* **  NARROW  ENTIR **  AN  ON  SID  ONE  JUST LY BARE **
99486007   DURING  DARKNESS  THE SUN  AN **  , THE SO-CALLED OF THE FLASH UM
99486008   SPECTR, ** *,* ** *.*

99486001   *MAN* ** *KANN* *EN* *BEOBACHT* *N* *LINIE* *ATMOSPHAERE*
99486002   *N* *SONNE* ** *EN* *DAMPFFOERMIG* ** *DER* ** *EN* *HELL* **
99486003   *DIE* *IN* *SCHICHT* *,*  OUS VAPOR COAT *UBERHALB*  AR SOL S
99486004   BORDER *EN* *AEUSSER* ** *DER* ** *,*  PHOTOSPHERE *,*
99486005   *EN* *SCHMAL* ** *EINER*   *EN* *SOGENANNT* ** *DER* ** *FUER*
99486006   *E* *BLICK* *N* *AUGE* ** *E* *WENIG* *E* *EINIG* ** **
99486007   *WENN*  THE ING ADVANC MOON ES LEAV E VISIBL A VERY THIN AR SOL
99486008   SURFACE EDGE ON ONE E SID JUST LY BARE DURING A AR SOL DARKNESS ,
99486009   THE SO-CALLED FLASH UM SPECTR, *,* ** *.*
```

Fig. 4-9

```
99490001    AR SCL CORONA S FOLLOW  AS A SILVERY E WHIT DIM BOUNDARY
99490002   FARTHER OUT .

99490001    AR SOL CORONA S FOLLOW  AS A SILVERY WHITE DIM BOUNDARY FARTHER
99490002   OUT .

99490001    AR SOL CORONA S FOLLOW AS A SILVERY E WHIT DIM BOUNDARY FARTHER
99490002   OUT .

99490001    AR SOL CORONA S FOLLOW AS A SILVERY WHITE DIM BOUNDARY FARTHER
99490002   OUT .

99490001    SOLAR CORONA S FOLLOW  AS A SILVERY E WHIT DIM BOUNDARY FARTHER
99490002   OUT .

99490001    SOLAR CORONA S FOLLOW  AS A SILVERY WHITE DIM BOUNDARY FARTHER
99490002   OUT .

99490001    AR SOL CORONA S FOLLOW  AS A SILVERY E WHIT DIM BOUNDARY
99490002   FARTHER OUT *.*

99490001    AR SCL CORONA S FOLLOW  AS A SILVERY WHITE DIM BOUNDARY FARTHER
99490002   OUT *.*

99490001    AR SOL CORONA S FOLLOW AS A SILVERY E WHIT DIM BOUNDARY FARTHER
99490002   OUT *.*

99490001    AR SCL CORONA S FOLLOW AS A SILVERY WHITE DIM BOUNDARY FARTHER
99490002   OUT *.*

99490001    SOLAR CORONA S FOLLOW  AS A SILVERY E WHIT DIM BOUNDARY FARTHER
99490002   OUT *.*

99490001    SOLAR CORONA S FOLLOW  AS A SILVERY WHITE DIM BOUNDARY FARTHER
```

Fig. 5-1

```
99492002   LIGHTS *M* ** **  WHOSE RELATIONSHIP TO N KNOW S ELEMENT ED
99492003   REMAIN UN N KNOW FOR A LONG E TIM , *.*

99492001     SEVERAL ING SHIN AL SPECTR LINES WERE ED DISCOVER *I*  CORONA
99492002   LIGHTS *M* ** **  WHOSE RELATIONSHIP TO N KNOW S ELEMENT ED
99492003   REMAIN UN N KNOW FOR A LCNG TIME , *.*

99492001     SEVERAL ING SHIN AL SPECTR LINES WERE ED DISCOVER *I*  CORONA
99492002   LIGHTS *M* ** **  WHOSE CONNECTION WITH N KNOW ELEMENTS ED REMAIN
99492003   UN N KNOW FOR A LCNG E TIM , *.*

99492001     WERE ED DISCOVER SEVERAL ING SHIN SPECTRAL ES LIN *I*  CORONA
99492002   LIGHTS *M* ** **  WHOSE CONNECTION WITH N KNOW ELEMENTS ED REMAIN
99492003   UN N KNOW FOR A LCNG TIME , *.*

99492001     SEVERAL ING SHIN SPECTRAL ES LIN WERE FOUND  IN  OUS LUMIN
99492002   CORONA ** *M* ** **  WHOSE RELATIONSHIP TO N KNOW ELEMENTS  ED
99492003   REMAIN UN N KNOW FOR A LONG E TIM ** *,* *.*

99492001     SEVERAL ING SHIN SPECTRAL ES LIN HAVE BEEN FOUND  IN  CORONA S
99492002   LIGHT *M* ** **  WHOSE RELATIONSHIP TO N KNOW S ELEMENT  ED
99492003   REMAIN UN N KNOW FOR A LONG TIME ** *,* *.*

99492001     SEVERAL ING SHIN SPECTRAL LINES WERE FOUND  IN  CORONA LIGHTS
99492002   *M* ** **  WHOSE CONNECTION WITH N KNOW ELEMENTS  ED REMAIN UN N
99492003   KNOW FOR A LONG TIME ** *,* *.*

99492001     SEVERAL ING SHIN SPECTRAL LINES HAVE BEEN FOUND  IN  CORONA
99492002   LIGHTS *M* ** **  WHOSE CONNECTION WITH N KNOW S ELEMENT  ED
99492003   REMAIN UN N KNOW FOR A LONG TIME ** *,* *.*

99492001     SEVERAL ING SHIN AL SPECTR ES LIN WERE FOUND  IN  CORONA LIGHTS
99492002   *M* ** **  OF WHICH THE RELATIONSHIP TO N KNOW ELEMENTS  ED
99492003   REMAIN UN N KNOW FOR A LONG TIME ** *,* *.*

99492001     SEVERAL ING SHIN AL SPECTR ES LIN HAVE BEEN FOUND  IN  CORONA
99492002   LIGHTS *M* ** **  OF WHICH THE RELATIONSHIP TO N KNOW S ELEMENT
99492003   ED REMAIN UN N KNOW FOR A LONG TIME ** *,* *.*

99492001     WERE ED DISCOVER  ING SHIN SPECTRAL ES LIN *E*  SEVERAL ** *I*
99492002   S' LIGHT  CORONA *DER* ** *M* ** **  RELATIONSHIP TO N KNOW
99492003   ELEMENTS  WHOSE  ED REMAIN UN N KNOW  A LONG E TIM ** *,* *.*

99492001     WERE FOUND SEVERAL ING SHIN SPECTRAL ES LIN *I* *E* *LICHT*
99492002   CORONA ** *M* ** **  RELATIONSHIP TO N KNOW S ELEMENT *DEREN*  ED
99492003   REMAIN UN N KNOW  A LONG TIME ** *,* *.*

99492001     WERE FOUND SEVERAL ING SHIN SPECTRAL LINES *I* *E* *LICHT*
99492002   CORONA ** *M* ** **  CONNECTION WITH N KNOW ELEMENTS *DEREN*  ED
99492003   REMAIN UN N KNOW  A LONG TIME ** *,* *.*
```

Fig. 5-2

```
99494001     B. EDLEN DID MANAG TO GET THESE SPECTRAL ES LIN IN SUITABLE
99494002    TERRESTRIAL OUS LUMIN S SOURCE IN UPSALA NOT UNTIL 1941 .

99494001     B. EDLEN DID MANAG TO GET THESE SPECTRAL ES LIN IN SUITABLE
99494002    TERRESTRIAL OUS LUMIN S SOURCE IN UPSALA NOT BEFORE 1941 .

99494001     B. EDLEN DID MANAG TO GET THESE SPECTRAL ES LIN IN SUITABLE
```

Fig. 5-3

```
99494002    TERRESTRIAL OUS LUMIN SOURCES IN UPSALA NOT UNTIL 1941 .

99494001     B. EDLEN DID MANAG TO GET THESE SPECTRAL ES LIN IN SUITABLE
99494002    TERRESTRIAL OUS LUMIN SOURCES IN UPSALA NOT BEFORE 1941 .

99494001     B. EDLEN DID MANAG TO GET THESE SPECTRAL ES LIN IN SUITABLE IAL
99494002    TERRESTR OUS LUMIN S SOURCE IN UPSALA NOT UNTIL 1941 .

99494001     B. EDLEN DID MANAG TO GET THESE SPECTRAL ES LIN IN SUITABLE IAL
99494002    TERRESTR OUS LUMIN S SOURCE IN UPSALA NOT BEFORE 1941 .

99494001     B. EDLEN DID SUCCEED IN TING GET THESE SPECTRAL ES LIN IN
99494002    SUITABLE TERRESTRIAL OUS LUMIN S SOURCE IN UPSALA NOT UNTIL 1941
99494003    .

99494001     B. EDLEN DID SUCCEED IN TING GET THESE SPECTRAL ES LIN IN
99494002    SUITABLE TERRESTRIAL OUS LUMIN S SOURCE IN UPSALA NOT BEFORE 1941
99494003    .

99494001     B. EDLEN DID SUCCEED IN TING GET THESE SPECTRAL ES LIN IN
99494002    SUITABLE TERRESTRIAL OUS LUMIN SOURCES IN UPSALA NOT UNTIL 1941

99494001     B. EDLEN DID SUCCEED IN TING GET THESE SPECTRAL ES LIN IN
99494002    SUITABLE TERRESTRIAL OUS LUMIN SOURCES IN UPSALA NOT BEFORE 1941
99494003    .

99494001     B. EDLEN DID SUCCEED IN TING GET THESE SPECTRAL ES LIN IN
99494002    SUITABLE IAL TERRESTR OUS LUMIN S SOURCE IN UPSALA NOT UNTIL 1941
99494003    .

99494001     B. EDLEN DID SUCCEED IN TING GET THESE SPECTRAL ES LIN IN
99494002    SUITABLE IAL TERRESTR OUS LUMIN S SOURCE IN UPSALA NOT BEFORE
99494003    1941 .

99494001     B. EDLEN DID MANAG TO GET THESE SPECTRAL ES LIN IN SUITABLE
99494002    TERRESTRIAL OUS LUMIN S SOURCE IN UPSALA NOT UNTIL 1941 *.*

99494001     B. EDLEN DID MANAG TO GET THESE SPECTRAL ES LIN IN SUITABLE
99494002    TERRESTRIAL OUS LUMIN S SOURCE IN UPSALA NOT BEFORE 1941 *.*

99494001     B. EDLEN DID MANAG TO GET THESE SPECTRAL ES LIN IN SUITABLE
99494002    TERRESTRIAL OUS LUMIN SOURCES IN UPSALA NOT UNTIL 1941 *.*
```

Fig. 6

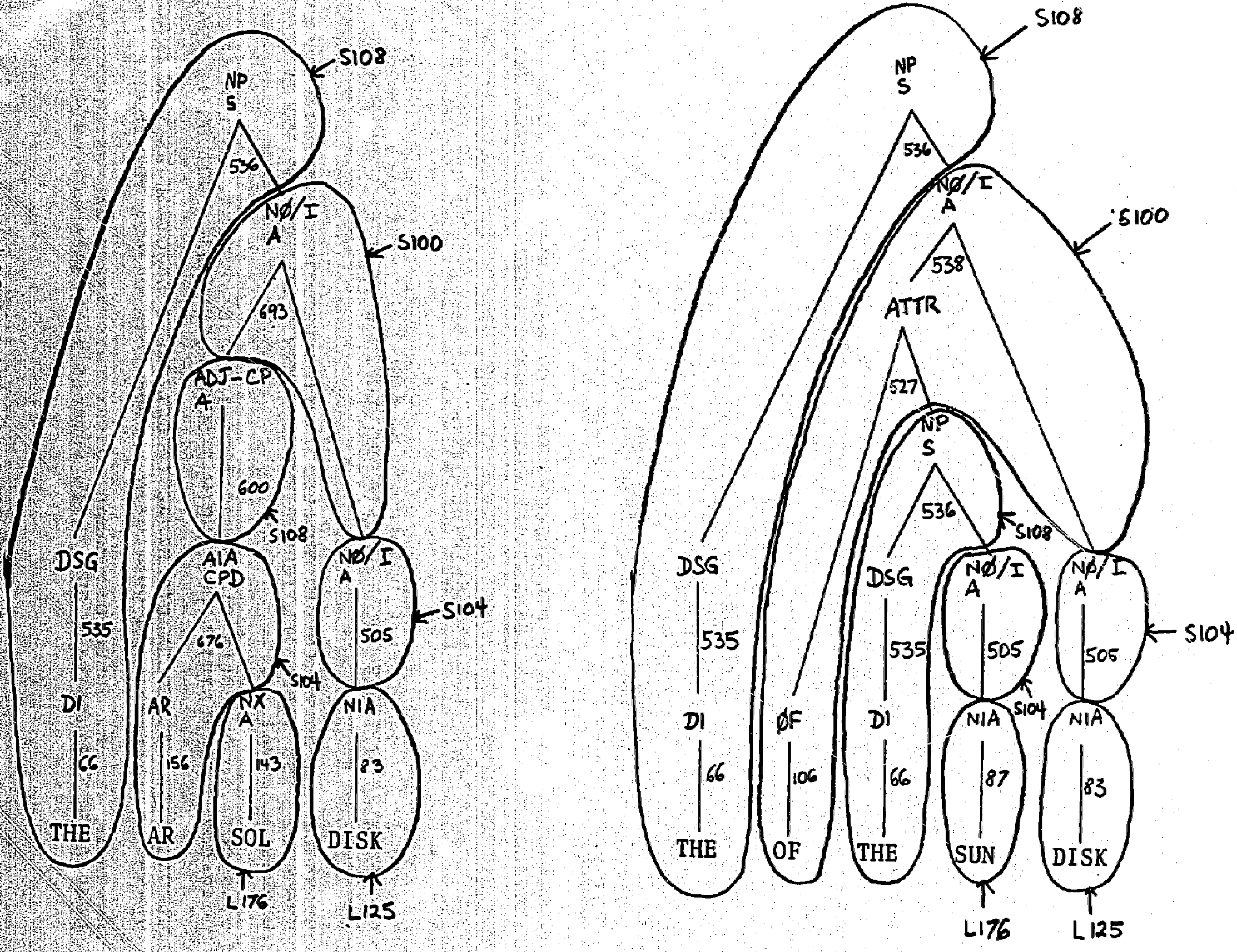

S108
NP
S
536
NØ/I
A
S100
693
ADJ-CP
A
600
S108
AIA
CPD
676
S104
NØ/I
A
505
S104
DSG
535
DI
66
THE
AR
156
AR
NX
A
143
SOL
L176
NIA
83
DISK
L125
S108
NP
S
536
NØ/I
A
S100
538
ATTR
527
NP
S
536
S108
DSG
535
DI
66
THE
ØF
106
OF
DSG
535
DI
66
THE
NØ/I
A
505
S104
NIA
87
SUN
L176
NØ/I
A
506
S104
NIA
83
DISK
L125

Figure 7-1

486 BEOBACHTEN:

1. ✓	SBJ +HU	_____	OBJ +ACC		<u>observe</u>, <u>watch</u>

Ex: Mark beobachtete Sylvia = Mark watched Sylvia.

2.	SBJ +HU	_____	OBJ +ACC +AB	an OBJ +DAT +HU	observe in sb., notice in sb.

Ex: Mark beobachtete Zeichen von Triumph an Sylvia = Mark noticed signs of triumph in Sylvia.

3.	SBJ +HU	_____	ADV +MAN		observe

Ex: Mark beobachtet gut = Mark observes well.

4. ✓	SBJ +HU	_____	OBJ +ACC +AB		follow, obey, observe, respect, comply with

Ex: Die Roemer beobachteten das Gesetz = The Romans observed the laws.

FREILASSEN:

1. ✓	SBJ +HU	_____	OBJ +ACC +HU		free, set free, liberate

Ex: Mark liess Sylvia frei = Mark set Sylvia free.

2. ✓	SBJ +HU	_____	OBJ +ACC -HU		<u>leave blank</u>, <u>leave open</u> <u>leave vacant</u>, <u>leave visible</u> 20

Ex: Mark liess eine Zeile frei = Mark left a line blank.

487 AUFTRETEN:

1.	SBJ +HU	_____	OBJ +ACC +PO -AN		kick open

Ex: Mark trat die Tuer auf = Mark kicked the door open.

2. ✓	SBJ +AN	_____			step, tread, walk

Figure 7-2

Ex: Mark trat leise auf = Mark trod softly.

3. SBJ +HU ____ gegen OBJ +ACC — come up against, rise against, oppose

Ex: Die Griechen traten gegen die Tuerken auf = The Greeks rose against the Turks.

4. SBJ +HU ____ fuer OBJ +ACC — stand up for

Ex: Mark trat fuer Sylvia auf = Mark stood up for Sylvia.

5. SBJ +HU ____ vor OBJ +DAT +HU — perform before

Ex: Mark trat vor dem Koenig auf = Mark performed before the king.

6. SBJ +HU ____ als CMPL +NOM — figure as, pose as

Ex: Mark trat als Koenig auf = Mark posed as a king.

7. SBJ +AN ____ wie CMPL +NOM — behave like, act like

Ex: Mark trat auf wie ein Fuerst = Mark behaved like a duke.

8.✓ SBJ +AB ____ — occur, happen, arise, result, ensue, appear

Ex: Ein Fall von Cholera war aufgetreten = A case of cholera had occurred.

9.✓ SBJ +HU ____ — appear, perform, enter

Ex: Mark trat in einem Stueck auf = Mark appeared in a play.

488 ERSCHEINEN:

1.✓ SBJ ____ — appear, emerge

Ex: Ein Wagen erschien = A car appeared.

Figure 7-3

2. SBJ +HU ____ OBJ +DAT +HU — appear to sb.

Ex: Der Geist war Mark erschienen = The ghost had appeared to Mark.

3. SBJ ____ OBJ +DAT +HU ADJ — seem, appear, look

Ex: Die Loesung erschien Mark gut = The solution looked good to Mark.

BREIT:

1.✓ ____ ADV +MEAS — wide, in width

Ex: drei Meter breit = three meters wide

2. N +PO ____ — broad, wide, spacious, large, vast

Ex: ein breites Gesicht = a broad face

3. N +AB ____ — extensive

Ex: eine breite Darstellung = an extensive descr'ption

490 ANSCHLIESSEN:

1. SBJ ____ OBJ +ACC -AB — chain, connect, fasten with a lock

Ex: Mark schloss das Fahrrad an = Mark fastened the bike with a lock.

2. SBJ +HU ____ OBJ +ACC +AB — add

Ex: Mark schloss eine Bemerkung an = Mark added a remark.

3. SBJ ____ OBJ +ACC -AB an OBJ +ACC -AB — chain to, connect to, join to, link up with

Figure 7-4

Ex: Mark schloss das Fahrrad an den Zaun an = Mark chained the bike to the fence.

4.	SBJ	____	OBJ +ACC +AB	an	OBJ +ACC +AB	add to

Ex: Mark schloss die folgende Bemerkung an seine Rede an = Mark added the following remark to his speech.

5.	SBJ +AN	____	OBJ +REFL +ACC		OBJ +DAT +HU	accompany, join

Ex: Mark schloss sich Sylvia an = Mark joined Sylvia.

6.	SBJ +AN	____	OBJ +REFL +ACC	an	OBJ +ACC +HU	accompany, join

Ex: Mark schloss sich an Sylvia an = Mark joined Sylvia.

7.	SBJ -AN	____	OBJ +REFL +ACC	an	OBJ +ACC -HU	be adjacent to, border on

Ex: An Texas schliesst sich Oklahoma an = Oklahoma borders on Texas.

8.✓	SBJ	____	follow

Ex: Weiter aussen schliesst die Sonnenkorona an = The corona of the sun follows further out.[20]

491 FINDEN:

1.✓	SBJ +HU	____	OBJ +ACC			discover, find, come across

Ex: Mark fand einen Diamanten = Mark found a diamond.

2.	SBJ +HU	____	OBJ +REFL +ACC	in	OBJ +ACC +AB	be reconciled with, resign oneself to, put up with

Ex: Mark fand sich in die Lage = Mark resigned himself to the situation.

3.	SBJ +HU	____	OBJ +ACC	ADJ	find, think, consider

Ex: Mark fand Sylvia huebsch = Mark considered Sylvia pretty.

Figure 7-5

492 ZUORDNUNG:

1. ____ zu OBJ +HU∨+AB +DAT — assignment to, relationship to, connection with

2. N +AB — coordination

ZEIT:

1. N +TIM — time

494 GELINGEN:

1. SBJ +AB ____ OBJ +DAT +HU — succeed in

Ex: Das Experiment gelang Mark = Mark succeeded in the experiment.

2. SBJ +AB ____ — be successful, succeed, work

Ex: Das Experiment gelang = The experiment was successful.

3.✓ es ____ zu INF OBJ +DAT +AN — succeed in + Gerund, manage to + Inf.

Ex: Es gelang Mark, das Experiment durchzufuehren = Mark succeeded in performing the experiment.

ERHALTEN:

1.✓ SBJ ____ OBJ +ACC — get, obtain, receive; keep, preserve 21

Ex: Mark erhielt ein Buch = Mark got a book.
Die Italiener versuchten, Venedig zu erhalten = The Italians tried to preserve Venice.

2.✓ SBJ +HU ____ OBJ +ACC +HU — support

Ex: Mark erhielt seine Eltern = Mark supported his parents.

Figure 7-6

3.	SBJ +HU	_____	OBJ +ACC +HU	von	OBJ +DAT +AB	maintain sb. on, support sb. on

Ex: Mark erhielt seine Eltern von seinem mageren Gehalt = Mark supported his parents on his small salary.

4.	SBJ +HU	_____	OBJ +ACC +REFL	von	OBJ +DAT +AB	subsist on, support onself on

Ex: Mark erhielt sich von Almosen = Mark subsisted on alms.

REQUIREMENTS FOR MACHINE TRANSLATION: PROBLEMS, SOLUTIONS, PROSPECTS

by

Rolf Stachowitz

Linguistics Research Center
The University of Texas at Austin

TABLE OF CONTENTS

1. Introduction

Today it is generally accepted that the expression "science" no longer refers to a discipline which deals with a particular subject area but in general to any discipline which uses a particular method of research: the so-called "scientific method". We classify various disciplines according to whether they make use of the scientific method or not. Thus, we exclude disciplines like history or literary analysis from the sciences.[1]

We shall only deal with two of the criteria which constitute the scientific method: intersubjectivity and verifiability. Intersubjectivity means that the result obtained by one person starting from certain assumptions and working according to a particular method should be obtainable by other persons operating with the same assumptions and the same method. By verifiability we mean that the statements on certain phenomena in a particular research area have to be empirically verifiable. The "principle of tolerance" (Toleranzprinzip), formulated originally but later abandoned by Carnap, no longer holds in the sciences. Introspective, phenomenological, and transnatural verifiability may only be used if they are reduceable eventually to verifiability through the senses.[2]

The development of linguistic theory and advances in computer hardware and software have put linguistic science into the forunate position of being able to verify by computer the various hypotheses and theories made about linguistic phenomena because of a correspondence between formal languages and programming languages: everything that can be formalized can be programmed and vice versa. A number of computational linguists have consequently written programs which process transformational grammars, so-called grammar testers, and have made them available to the linguistic community. The linguistic community has as yet made little use of such programs. The few linguists who have had their grammars processed by such an algorithm soon found out that their hypotheses were falsified.

The reluctance of linguists to use a computer is, of course, based on the fact that there is no comprehensive theory of grammar that works. Estimates on the length of time required to construct such a grammar vary considerably. We have heard opinions indicating a time of about 500 years. Though we are inclined to regard this figure as an exaggeration, a number of renowned linguists have seriously stated they feel that it may take about 150 years of grammatical research to come up with a comprehensive grammar for a language.

What are the avenues open to the linguist who is not patient enough to wait that long in order to test his hypotheses or theories? He can resign himself to the view that language is a phenomenon which cannot be treated algorithmically, at least not from a recognition point of view, which is true for formulas of the predicate calculus. We personally are disinclined to accept such a resignation since we know that everybody can speak but not everybody can prove logical theorems.

The second possibility is to assume that grammars are indeed highly complicated and that we must work patiently, hoping that future generations will be able to make use of our preliminary work.

The third possible course of action, the one we are going to follow, is to investigate whether all the scientific and methodological premises of current grammar theory, especially its descriptive and explanatory apparatuses, are really necessary, or whether they can be replaced by a simpler system of apparatuses under preservation of the observational, descriptive, and explanatory adequacy. We shall thus treat current linguistic theory as the object of research of another science, its meta-science. We shall investigate linguistics from a meta-linguistic point of view according to which the components of a grammatical model are subject to scientific investigation based on the scientific method. Which empirically observable, experienceable phenomena correspond to a competence model, as grammar models are normally called, and to its various components, the deep phrase-structure component, the transformational component, and the semantic component? (For present purposes, we shall ignore the phonological component.) Which are the phenomena explained by such a model, which remain unexplained?

To accept the stipulation of transformational grammarians that competence models not be regarded as performance models imposes a heavy burden on our research, but instead of discussing whether such a request is legitimate, we decide that we can still investigate such models and their components as part of a hypothesized performance model.

It is very difficult to believe the claim that a grammar of a language with a finite set of terminal symbols is an adequate representation of a phenomenon that occurs almost any day: the introduction of new words in a language, which either name new objects or which are introduced by means of definitions. A grammar model as it is normally defined is basically static, something that, I believe, Humboldt would have called not an energeia but an ergon, incapable of representing the changes that occur in any living language. (Cf. the interesting footnote in Hans Hermes: "The schematical execution of a given general procedure (i.e. algorithm, our addition) evidently offers (after some attempts) no particular interest to a mathematician. We can thus state the remarkable fact that a creative mathematician - through the specific mathematical achievenent of the development of a general procedure - renders valueless, as it were, the area covered by this procedure."[3])

Which possibilities for verification do we have for a competence model?

a) We could check its output. Apart from the fact that this output does not exist yet, this criterion, if used alone, could also be used to represent as a model for the human capability to divide and multiply a computer program which performs division and multiplication by iterative subtraction and iterative addition.

b) We could consider the structural description which is assigned to surface sentences. We grant that the structural description which a competence model assigns to a surface sentence corresponds to our linguistic intuition. However, we see no means to decide that such surface structures are derived transformationally from deep structures; they might equally well be derived from a surface phrase-structure component. Recent development in standard transformational grammar which makes the deep structure representation correspond more and more to the surface representation actually argues in favor of the latter assumption.

Which empirical verifiability exists for a deep phrase-structure component? The claim that the deep structure representation permits a formal definition of semantic categories, as subject of a sentence or predicate of a sentence, has already been shown by various transformational grammarians not to be applicable for such semantic categories as objects or adverbials in the case of verbs with multiple objects and multiple adverbials. This claim, I believe, was shaken by Charles Fillmore[4], who pointed out that the deep representation is not really a representation of semantic relations between constituents. This has been admitted by Chomsky if I understand his comments in "Deep Structure, Surface Structure and Semantic Interpretation" correctly. Others pointed out that important linguistic concepts as "head" of a phrase cannot be expressed by means of the deep phrase-structure component.[5]

Which reality corresponds to the transformational component? We do not doubt that transformational relations exist between surface structures. But, as far as I know, there is no empirical verification for the existence of ordered transformations. The few examples, all based on reflexivization, can be explained in a different way.

Which observable phenomenon corresponds to an intermediate phrase marker? No real investigation has been performed on this aspect.[6] The reality of intermediate phrase markers can be easily tested by confronting a naive speaker with such sentences as "By John give Harry the book"; they normally find it unintelligible; occasionally they interpret as "Give Harry the book written by John". We know that the string, by means of preposition deletion, eventually results in "John gives Harry the book".

Which experienceable reality corresponds to a semantic component, which cannot explain the process of introducing a new word by definition, the modification of meaning by explication, which cannot represent in a sentence reading the synonymity or the occasional intersection of the semantic readings of two words expressed by the "explicative or" (corresponding to the stylistic term "hendyadyoin") when no individual term in a language represents that semantic reading?[7]

The rigor which had been introduced into linguistics by means of the notion of rules and transformation rules in the earlier version of transformational grammar has gradually disappeared. We are not able to relate the surface phenomena that we can observe to the semantic representation or the deep structure since the increased complexity of the transformational apparatus makes the establishment of such relations and their verification extremely difficult if not impossible. The "remedies" which have been proposed: to make the deep structure more and more similar to the surface structure or more and more abstract to arrive at the semantic representation, we regard as futile in view of the results obtained by

Peters and Ritchie.[8]

In a science we set out to describe the facts that we observe and to try to relate them, to find an explanation for them, a system, a structure. The principles that in general are used in setting up the observational and explanatory apparatus are that they should be adequate and appropriate. These principles are also influenced by certain esthetic considerations: that the apparatus should be as simple as possible. From our point of view, this means: We now know a lot more about linguistic theory than we did twenty years ago. We know that language is the language of man, whose capabilities we should not exclude when dealing with language. We should begin research again by relating surface sentences to surface sentences by means of transformations, but by means of transformations which are kept as simple as possible, which relate surface structures to surface structures and which, if possible, need not apply in a particular order. Only if the facts force us to make changes in our assumptions, should we make the necessary changes; we should not start out by carrying over into our own discipline certain apparatuses useful and also necessary in others, at least not without weighing the pro's and con's carefully. We should start by constructing a model which reflects such considerations. It is not even necessary to find or develop such a model since the person who started it all, Zellig S. Harris, has been describing such a model for some time.[9] Our own model, which we are going to describe in Chapter 5 of this paper, is based on the notion of Harris' substitution transformations. It has been constructed with the aim to explain certain human capabilities, among them the acquisition of new words and their definition by means of the context. Our grammar is based on the assumption that sentences can be represented as connections of elementary predications. Thus the sentence "A young girl sang a song" is representable as the sequence of connected predications:

$$\mathrm{girl}(x_1) \wedge \mathrm{young}(x_1) \wedge \mathrm{song}(x_2) \wedge \mathrm{sing}(x_1, x_2)$$

Sentences are not generated by rewriting the initial symbol S but by reducing them to symbol S both during recognition and production. The model is a representation of recognition in that is derives meanings from surface sentences; a model of production, in that it derives surface sentences from a representation of their meanings.

When I first proposed, after my experiment in paraphrasing and translation,[10] to reduce sentences of a natural language mechanically to connected elementary sentences, means were not available to extend my experiment. For some time, the project lay dormant. That it has been revived I owe to three persons, to whom I would like to express my gratitude:

Winfred P. Lehmann, Rowena Swanson, and Zbigniew L. Pankowicz. In order to prepare for the discussion of our model, we shall introduce in Chapter 2 a simplified model of human comprehension. In Chapter 3 we will discuss the requirements for a quality or high quality machine translation system. In Chapter 4 we discuss the capabilities of current competence models from the point of view of applicability to machine translation. Chapter 5 gives an outline of our model, the Linguistics Research System. Chapter 6 discusses primarily a development in computer storage whose impact on the scientific community, in particular on linguistics and linguistic studies, cannot be estimated yet.

2. Comprehension and Translation

In order to describe and clarify the extent to which translation of a text is dependent on the comprehension of that text, we shall construct a simplified, restricted model of human comprehension and determine the components of this model which will have to be part of a translation device. To facilitate the description of such a model we shall introduce the following terms by example: State of affairs, state-of-affairs-description, the image of a state of affairs, the image of a state-of-affairs-description.

Assume that a number of people observe an incident Q, a traffic accident, involving two objects: a car and a pedestrian. Two or three observers make the following statements about Q:

1) There was a car-pedestrian accident.
2) A car hit a man.
3) A Porsche hit a man.

We shall say that the statements 1 through 3 describe the same state of affairs Q (SA Q), though with different information content. We shall call each statement a description of SA Q or an SA-description of Q. Clearly, each of the statements is not only an SA-description of Q but of several SA's; thus statement 3 describes all car accidents similar to SA Q which involve a Porsche hitting a man. We shall thus call statement 3 an SA-description, independent of the particular SA it describes.

We shall further posit that every sentence, whether command, request, question or statement, is a description of some SA.[11] An SA need not have any physical reality. This follows from the fact that an SA-description may be false.

Let us now assume a device K - with several components - which can process SA-descriptions, store them and reproduce them; it can also assign to an SA-description p all the syntactic, structural descriptions of p; it can further associate one or more images with each SA-description. Thus, K associates the different images a, b, and c with the SA-descriptions 1, 2, and 3, respectively. However, K associates the same image d with the SA-descriptions 4 and 5; it associates image e with the SA-descriptions 6 and 7, image f with the SA-descriptions 8 and 9, and two different images g and h with the SA-description 10.

4) A Porsche hit a man. 5) A man was hit by a Porsche.	image d
6) The man scaled the fish. 7) The man desquamated the fish.	image e
8) A car, a Porsche, hit a man. 9) A Porsche, which is a car, hit a man.	image f
10) George observed a man with a telescope.	image g image h

We call an image associated with an SA-description a DSA-image. (As we can observe, the relations between a DSA-image and an SA-description are similar to those that hold between an SA and an SA-description. A DSA-image can be associated with more than one SA-description. An SA-description can be associated with more than one DSA-image; whenever this is the case, we call the SA-description ambiguous.)

Let us now clarify the term DSA-image by constructing the DSA-image for the following sentence:

11) A woman sold a car to a man for some money.

For the time being let us refer to the woman as A, to the car as B, to the man as C, to the price as D. Let us now describe what happens during a sale of some property. a) A, the owner of B, gives B to C. Let us represent this by the following graph (arrows represent relations and unary actions):

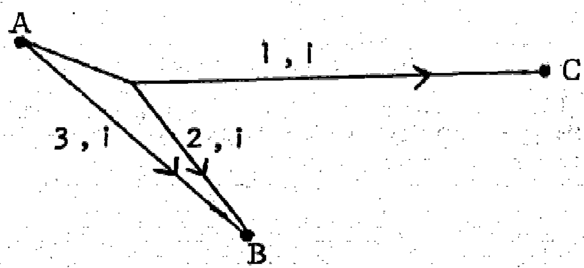


Figure 1

where "1,i" represents A gives to C at time i", "2,i" represents "A gives object B at time i", "3,i" represents "A owns B at time i". Note that the ternary relation "A gives B to C" is expressed by a "fork". b) After this act, C acquires property of B and A loses it, expressed as follows:

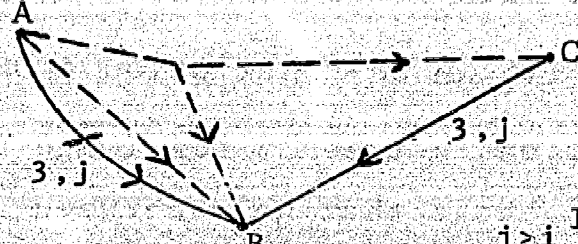


Figure 2

$j > i$[12]

Negation is expressed by a slash through the line representing the property or relation. c) Then C, who owns the money D, gives this money to A as a compensation for the acquisition of B. This results in the Fig.3 where k is

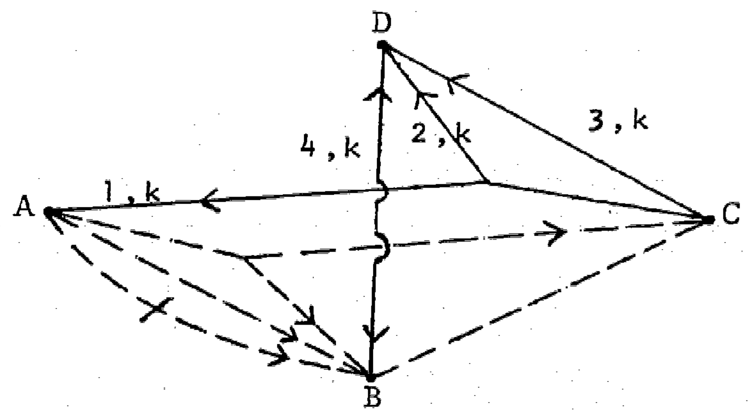


Figure 3

later than j. The double arrow, between D and B, represents a symmetrical relation. 4(B,D) stands for "B is a compensation for D". d) Finally, A acquires property of D and C loses property of D, resulting in the graph

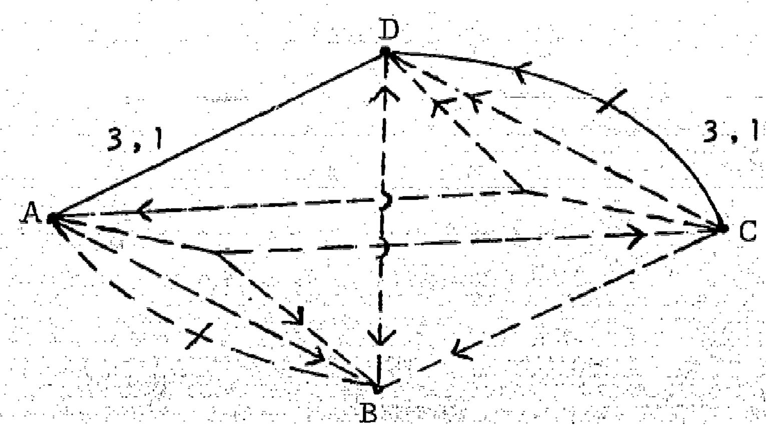


Figure 4

Sales transactions can only take place between human beings and/or legal entities. We thus add this information to nodes A and C.

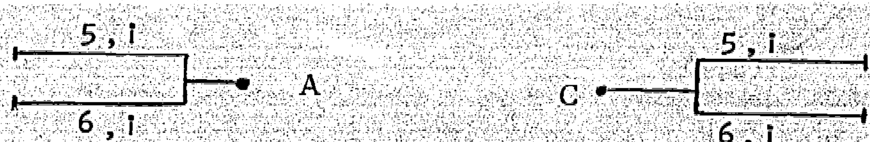


Figure 5

where the graph •——┤ represents a property of the node, and a line perpendicular to a property (or relation) a <u>logical or</u>.[13] 5 represents the property human; and 6, the property legal entity. The sold object B must finally be an object or a right to some object, D can be an object, or the right to some object, or money, which will be represented in the graphs

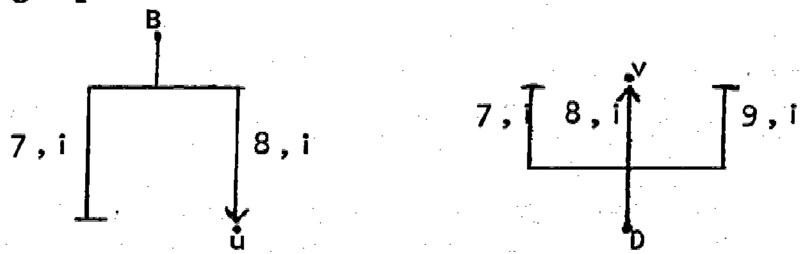


Figure 6

where 7 represents the property "physical object", 8 the relation "right to" and 9 the property "money".

Sentence 11': "A sells B to C for D" thus results in the following DSA-image:

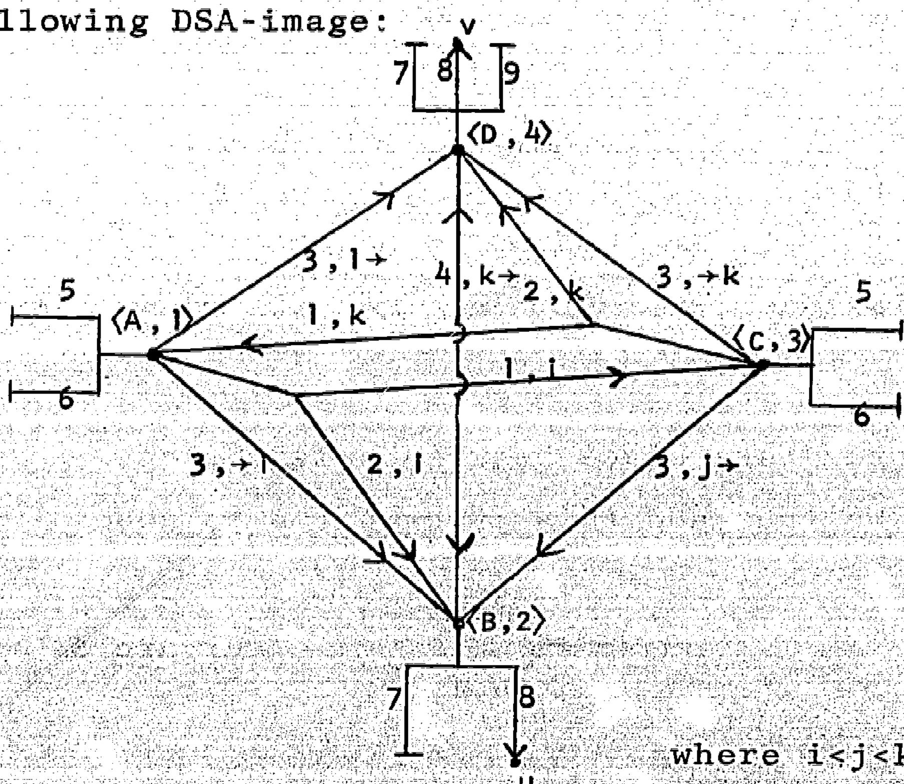
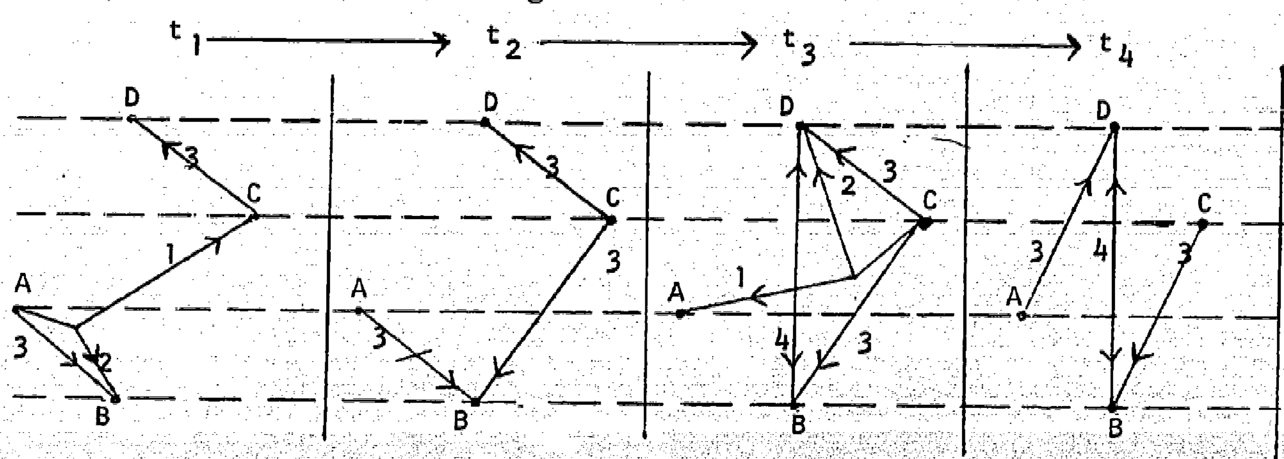


Figure 7

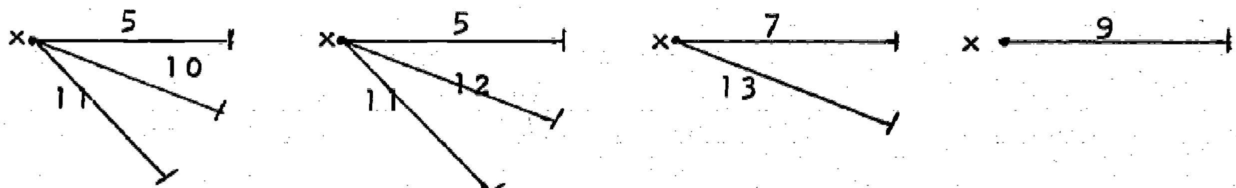
where i<j<k<l

The following conventions have been used in this figure: An expression of the form "number+,+→+letter" (e.g. 3,→k) is to be read as "The property or relation represented by the number ends at the point of time represented by the letter". An expression of the form "number+,+letter+→" is to be read "the property or relation represented by the number begins with the point of time represented by the letter". An expression of the form "number+,+letter" is to be read as "the property or relation expressed by the number begins at and terminates with the point of time represented by the letter". An expression of the form "number" (with no letter) expresses that the property or relation has no time boundaries. We prefer the representation in Figure 7 to the equivalent representation in Figure 8.

Figure 8



The graph in Figure 7 closely corresponds to the SA-image of the predication described by the verb "sell(<x,1>,<y,2>,<z,3>,<v,4>)".[14] This is obvious if we replace the node names A, B, C, and D by x, y, z, v, respectively. To obtain the SA-image of sentence 11, we still need to perform the predications upon the objects referred to by the expressions "a woman", "a man", "a car", and "money". These will be represented in that order by the following graphs:



10 = female
11 = adult

12 = male

13 = car

Figure 9

Sentence 11 will result in the following DSA-image:

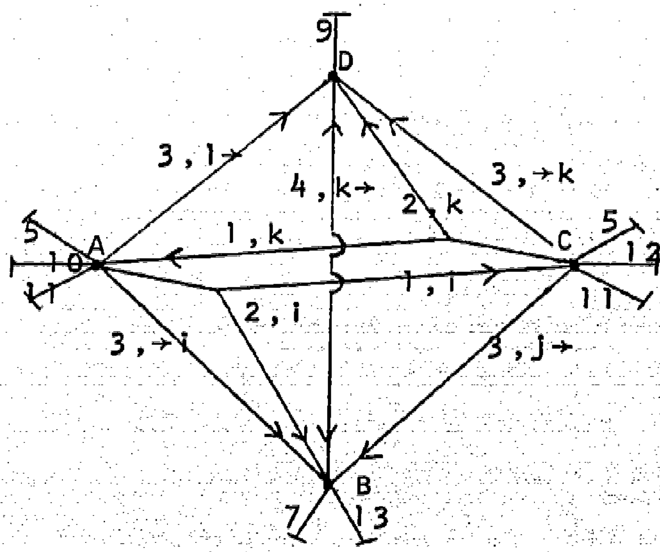


Figure 10

Note that in comparison with Figure 7 predications upon the objects A, B, C, and D have changed, i.e. we are dealing with human beings as seller and buyer, it is an object and not a right to something that has been sold, and the compensation for the object is money, not another object or right to something.[15]

We shall further assume that device K contains an additional component in which SA-images, images of the original state of affairs, are stored. Each SA-image is generated by means of the information provided by a DSA-image by replacing the object variables by constants. The SA-image constructed from sentence 11 would be identical with the DSA-image in Figure 10 if A, B, C, and D were replaced by x_1, x_2, x_3, x_4, respectively. Each SA-image t of SA Q is consequently a partial, i.e. imperfect, representation of the original SA Q.

A further component of K is able to superimpose two SA-images p and r of an SA Q and thus derive an SA-image v of SA Q by modifying - during the processing of a text - the current SA-image p of SA Q by means of the new SA-image r of SA Q; the result is a more precise representation of the original SA Q: SA-image v. Let us call such superimposed SA-images <u>connected</u> SA-images. This component also deletes all but the SA-image t of an ambiguous SA-description, as well as their DSA-images if t was connected with some SA-image q. This capability means that the device is able to connect SA-images, represented in different SA-descriptions, similar to the connection of SA-images represented in the change of the graph in Figure 7 to that of Figure 10. If two devices K_1 and K_2 with identical internal configurations, both beginning with an empty data storage, process

(4) A Porsche hit a man.

and

(2) A car hit a man. 12) It was a Porsche.

respectively, then, when each has processed its first sentence (4 and 2), the contents of the data storage of the two devices will be different in at least three respects: each will contain a different SA-description; each, a different DSA-image, and each, a different SA-image. When, however, K_2 has processed sentence 12, both devices will have an identical SA-image. That is, the sentence

(4) A Porsche hit a man.

and the sequence of sentences

(2) A car hit a man. (12) It was a Porsche.

result in the same SA-image.

When device K processes sentence 10:

(10) George observed a man with a telescope.

it will construct two DSA-images and two SA-images; this expresses the ambiguity of this sentence. If K subsequently processes sentence 13:

13) The man put the telescope down.

the DSA-image and SA-image which represent George as the user of the telescope will be deleted.

Of the states of K we shall call state T_q (such that there is no r>q) the <u>current state</u> of the device K. We will call the set of SA-images at state T_q the <u>current SA-image</u> of K;

the set of configurations of th- -image component from state T_0 through T_q, the m ory of K; the set of SA-image configurations of the states immediately preceding the current state of , the <u>short-span memory</u> of K.[16]

Further as ume that device K has a meaning rule component with ference rules, statements of definitions, and equi- lence rules. Examples of inference rules are

a) For all x: if x is a Porsche, then x is a car,

b) For all x: if x is a car, then x is a vehicle,

c) For all x: if x is human, then x is animate.

An example for a definition is:

the SA-image in Figure 11 $=_{Df}$ the SA-image in Figure 7',
(in Figure 7' A, B, C, D of Figure 7 have been replaced by x, y, z, v, respectively.)

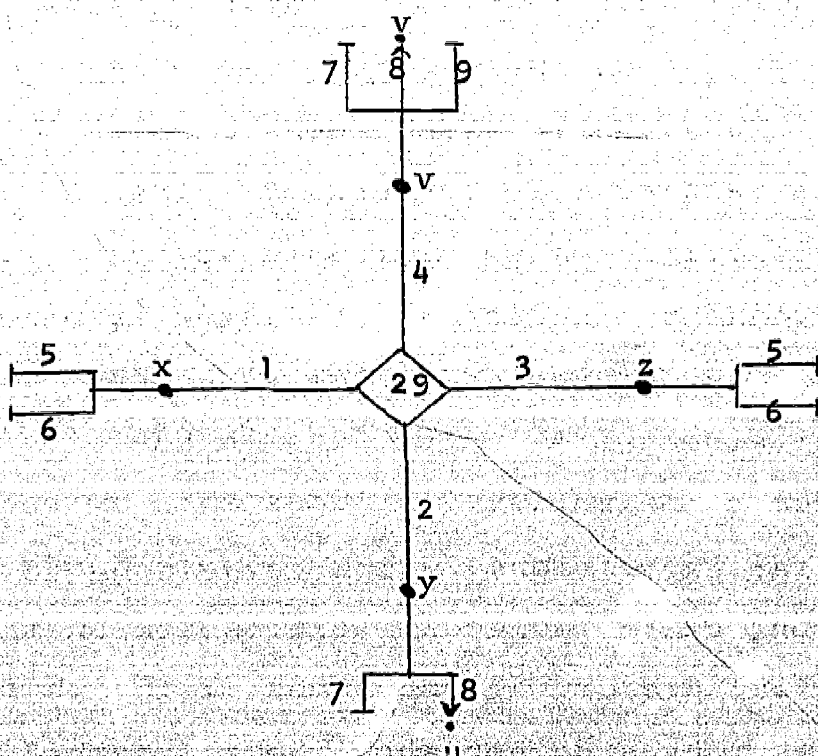
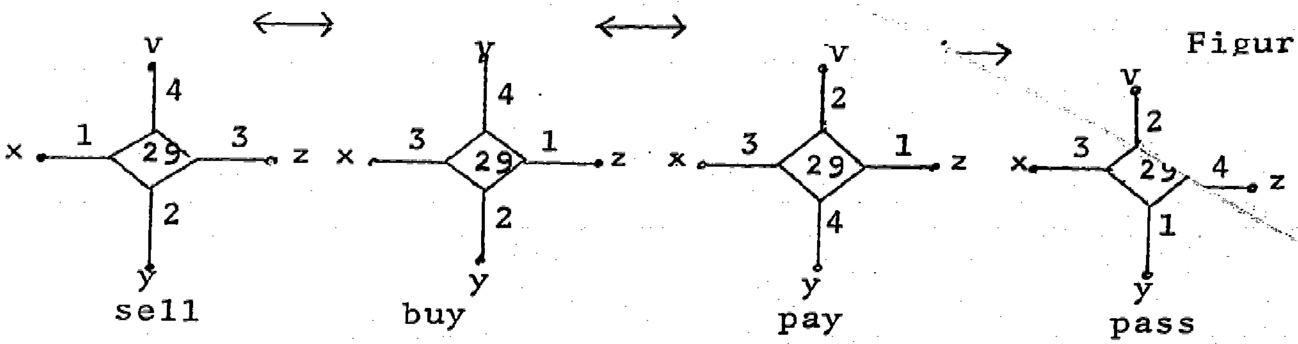


Figure 11

29 = sell
(Lines 5 through 9 represent atomic properties or relations, cf. Figure 7, page 15.)

Examples of equivalence rules a[illegible] given in Figure 12.

Figure 12



These graphs represent the meaning rules:

$\mathrm{sell}(\langle x,1\rangle,\langle y,2\rangle,\langle z,3\rangle,\langle v,4\rangle) \equiv_{Df} \mathrm{buy}(\langle x,3\rangle,\langle y,2\rangle,\langle z,1\rangle,\langle v,4\rangle) \equiv_{Df} \mathrm{pay}(\langle x,3\rangle,\langle y,4\rangle,\langle z,1\rangle,\langle v,2\rangle) \equiv_{Df} \mathrm{pass}(\langle x,3\rangle,\langle y,1\rangle,\langle z,4\rangle,\langle v,2\rangle)$.

The sentence "A woman (x) sold a car (y) to a man (z) for some money (v)" can thus also be represented as "A man bought a car from a woman for some money", "A man paid some money to a woman for a car", "A car passed for some money from a woman to a man".[17] Thus, device K can construct by means of the rules of the meaning rule component, in particular by means of the definitions, molecular SA-descriptions, molecular SA-images, and connected molecular SA-images from the SA-images, DSA-images, and connected SA-images which from now on we shall call atomic DSA-images, atomic SA-images, and connected atomic SA-images. It does this by replacing atomic and/or molecular expressions, which correspond to the right side of a definition, by the molecular expression on the left side of the definition, preserving the names of the object nodes involved. Molecular images do not show their internal structure. Thus, the graph in Figure 10, which represents sentence 11: "A woman sold a car to a man for some money", will result in the graph in Figure 13. (We represent molecular images by two-dimensional figures: quaternary relations by a diamond, properties of an object by a rectangle;[18] objects are represented by a dot, the names of relations and properties are represented by numbers in the geometrical figures. The names of objects occur besides the dots, the numbers on the lines between relations and objects represent the order of the arguments.)

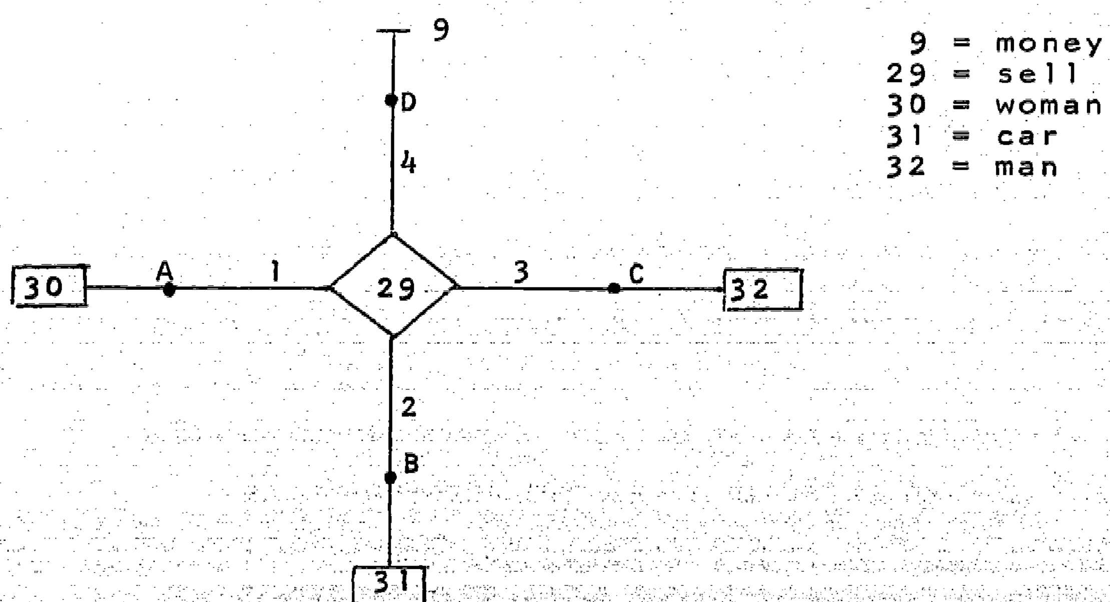


Figure 13

We obtain the molecular SA-image corresponding to the graph in Figure 13 by replacing the expressions A, B, C, and D by x_1, x_2, x_3, x_4, respectively.

We assume that device K will permanently store only molecular DSA-images and connected molecular SA-images, since it can construct the corresponding atomic DSA-images and SA-images by means of its meaning rule component with its definitions and inference rules, when required.

We suppose nobody will seriously doubt that, indeed, connected SA-images, atomic and/or molecular, or simulations of them are stored in comprehension devices, as e.g. in the human brain, or that SA-images are necessary besides DSA-images. Without this assumption, it would be fairly difficult to explain the inconsistencies in a number of SA-descriptions of some SA R when no two of them are inconsistent. Let us demonstrate this by the following three SA-descriptions of the same SA which may occur distributed over some text.

13) The final conference on the "Theoretical Study Effort of High Quality Translation" was held in Austin, Texas, from January 11 through January 15, 1971.

14) When the final conference on the "Theoretical Study Effort of High Quality Machine Translation" was held, it rained every day in Austin.

15) No rainfall occurred in Austin, Texas, during the period of January 11 through January 15, 1971.

As we can easily verify, each pair of the statements 13 through 15 is consistent. The three statements together, however, are inconsistent. Of course, the inconsistency of statements 11 through 13 does not simply follow from the connected SA-images representing the state of affairs described by statements 11 through 13. For this we need an additional component, a logical component.

That a process corresponding to the connection of SA-images actually occurs in the human brain is most obvious whenever a hearer encounters a sentence which - in isolation - is semantically anomalous or possibly even contradictory. Thus, sentences 16 and 17:

16) Haensel broke off a part of the roof and ate it.

17) This boy is a girl.

which are not semantically well-formed, i.e. whose DSA-images are not "well-formed", make sense in their proper context. Sentence 16 occurs in Grimm's fairytale Haensel und Gretel, sentence 17 in numerous stories in which a girl, in order to be near her lover, a soldier, disguises herself and joins the army. Her true identity is eventually discovered. In the case of sentence 16, the system has stored the fact that the witch's house consists of cake and candy, i.e. that the house and its parts are edible. Thus, the SA-image of sentence 16 is compatible with the established fact structure, the current connected SA-images, though the DSA-image of sentence 16 violates at least one of the rules of the system's meaning rule component. In the case of sentence 17, which is contradictory and thus logically false, the system establishes that one of the predications a and b with the argument x_j (the disguised girl) in the SA-image of sentence 17

a) x_j is a boy and b) x_j is a girl

is not consistent with the current SA-image pertaining to x_j. The system, depending on outside information, either rejects predication (a) as false, or predication (b), or both.

We shall now introduce the last necessary component of device K. So far, we have tacitly assumed that an SA-description describes an SA that occurs or exists outside of K. An SA-description may, of course, also describe SA's inside of K, as, for example, components, meaning rules, states, SA-descriptions, DSA-images, and SA-images. We shall classify two devices J and K, with the same properties mentioned so far and identical internal configurations, according to the way they process or react to the following statements:

17) Did Mary sell a car?

18) What did Mary sell?

19) Mary sold a car.

20) "Mary sold a car" is a sentence.

21) "Mary sold a car" is not a sentence.

Device J processes the sentences 17 through 21, storing for each sentence the SA-description, and the associated DSA-images and SA-images. Its only in-built reaction is that either sentence 20 or sentence 21 or both be deleted from the memory, since they are inconsistent. K reacts in the following way: (we shall use "SA:x" for "the SA described by the SA-description of SA x"): When K has established the DSA-image of SA:17, it searches through its memory. If an SA-image identical (except for the representation of negation) to the SA-image the device J would produce when processing sentence 19 has been stored or can be deduced from existing SA-images by means of meaning rules and logical rules, K prints out "no" if at least one negation occurs in the SA; "yes", if no negation occurs. If no such SA-image is found, K prints out the stereotype answer: "The question cannot be answered, insufficient information." For sentence 18: Again, recognizing that an answer is expected, searching through its memory and finding a representation of "Mary sold her house on the 20th of July, 1969. She got $25,000 from Henry for it.", K prints out: "Mary sold a house to Henry for $25,000 on July 20, 1969." K then continues processing statement 19 in the way J processes it. We shall call device J a somewhat sophisticated <u>language data processor</u>; we shall call K a model of comprehension or a device with <u>rudimentary artificial intelligence</u>.

A slightly more intelligent version of K, having generated the DSA-images of SA 20 (or SA 21), will analyze the DSA-images $x \bullet\!\!\underset{15}{\rule{2em}{0.4pt}}\!\!\dashv$ (and $x \bullet\!\!\underset{15}{\rule{1em}{0.4pt}\!\!/\!\rule{1em}{0.4pt}}\!\!\dashv$) by means of an operation rule; ($\bullet\!\!\underset{15}{\rule{2em}{0.4pt}}\!\!\dashv$ represents the predicator "sentence"). This operation rule, a subroutine called by $\bullet\!\!\underset{15}{\rule{2em}{0.4pt}}\!\!\dashv$ (or $\bullet\!\!\underset{15}{\rule{1em}{0.4pt}\!\!/\!\rule{1em}{0.4pt}}\!\!\dashv$), establishes that the SA-description is true (false) if x is generatable by the syntactic component; if x is not generatable, that the SA-description is false (true). The corresponding SA-images and DSA-images will be deleted.

This "awareness" component of K, if modified slightly in the way indicated below, would also make device K a restricted speech production device. The modifications necessary would be:

a) K may print out a sequence of SA-descriptions t_1, t_2, ... t_n;

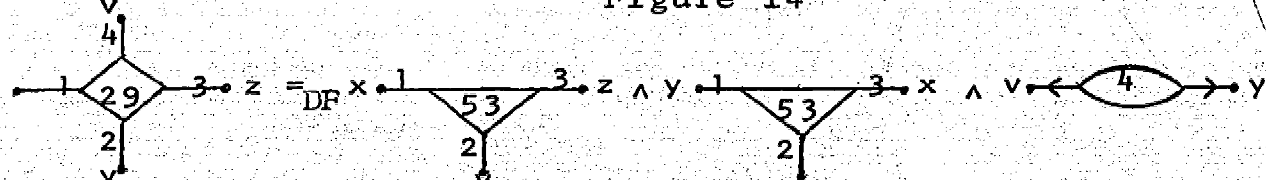
b) each t_i ($1 \leq i \leq n$) is a partial, incomplete representation of the underlying SA-image;

c) for each t_i, t_{i+1} ($1 \leq i < n$): the SA-image of t_i is connected with the SA-image of t_{i+1};

d) the conjunction of all SA-descriptions t_i ($1 \leq i \leq n$) is an exhaustive description of the underlying SA-image.[19]

By means of the semantic component and the definitions in the meaning rule component given in the following figure, K can produce the sequence of sentences below.

Figure 14



where 29 represents "sell"; 53, "give"; 4, "is a compensation for"; the caret stands for logical and.[20]

22) A woman sold a house. A man gave her money for it.

23) A house was sold. The owner, a woman, got some money for it. The present owner is a man.

24) A woman sold something. It was a house. Somebody, a man, gave her some money. The money is the compensation for the house. etc.

In addition to the necessary components already mentioned, the device may contain several others, as e.g. a component which associates a stylistic interpretation with an SA-description t, or a component which corrects printing errors.[21]

Let us recapitulate the major properties of the comprehension device. It is able to store and reproduce SA-descriptions. By means of a syntactic component, it can associate with each processed SA-description t all and only the syntactic descriptions of t. By means of a semantic component, it can associate with an SA-description t all and only the DSA-images of t. It can further associate all and only the SA-images of t with SA-description t by means of a discourse structure component. The association component of K performs the connection of SA-images pertaining to the same SA.

In addition, the device contains a meaning rule component, a logical component, and an "awareness" component. A more elaborate description of such a model of comprehension for purposes of Information Retrieval can be found in our report "Normalization of Natural Language for Information Retrieval".

Let us now represent the terms introduced above by their linguistic equivalences. An SA-description is a sentence in natural language. The syntactic description of an SA-description is the description of the surface structure of a sentence in natural language. An atomic DSA-image represents the meaning of a sentence in isolation. An atomic SA-image represents the meaning of a sentence in context. Molecular images may correspond to "semantic readings". We are not aware of an established linguistic term which corresponds to the set of connected SA-descriptions in the current state T_q of the device; it represents the current knowledge of facts of the device. The term "state of affairs" finally corresponds to the terms "referent", "significatum", "denotatum".[22]

We shall call a sentence t synonymous with a sentence u if t and u have the same SA-image or meaning.[23] In particular we shall call sentence t a paraphrase of sentence u if t is synonymous with u, and t and u are sentences of the same language. We shall call sentence t a translation of sentence u, if t and u are synonymous, and t and u do not belong to the same language.

The purpose of these explanations was to provide the basis for a discussion of the components of a translation device and, in particular, of the question which of the components of a comprehension device should be part of such a translation device.

3. Desirable Properties of a Translation Device

It is sometimes argued that in translation, at least in MT, it is not necessary to understand the meaning of a text as long as the target language equivalents for the words and syntactic structures of the source language can be correctly established or - in our formulation - as long as molecular or atomic expressions and syntactic structures of the source language can be mapped into the corresponding equimolecular or atomic expressions and structures of the target language.

We shall investigate, by means of the following German examples and their English translations, the extent to which this claim is justifiable by showing some of the problems that a mechanical translation device T will encounter and will have to solve. We shall try to indicate which of the components of device T will be involved in handling a particular problem, and, specifically, which components of device K must be part of T. (We do not restrict our attention to the translation of scientific texts. Statements on the greater ease with which such material may be mechanically translated seem to express to a greater extent opinions rather than careful investigations;[24] we also assume that MT device T will be able to translate scientific texts if it can translate "normal text", provided that the necessary vocabulary and their equivalences have been incorporated into T.)

The first requirement that an MT device should meet is to be able to derive the semantic reading R of t from a surface sentence t. In particular, an MT device should be able to handle syntactic problems represented by the following German examples. (In each of these examples, the correct English translation will be preceded by a literal translation.)

1. *Die Geschichte faengt mit einer Explosion an.*
   The history catches with an explosion at.
   History begins with an explosion.

2. *Er liess ihr Bescheid sagen, dass ...*
   He let her notice say that ...
   He sent word to her that ...

3. *Ich habe ihm aber Bescheid gesagt.*
   I have him but notice said.
   I gave him a piece of my mind.

4. *Die Sonne geht im Osten auf und im Westen unter.*
   The sun goes in the east up and in the west down.
   The sun rises in the east and sets in the west.

5. *Fritz ist nach Spanien, seine Frau nach Italien und ihre Tochter nach Griechenland gereist.*
   Fritz is to Spain, his wife to Italy, and their daughter to Greece traveled.
   Fritz traveled to Spain, his wife to Italy, and their daughter to Greece.

It may be obvious from these examples that the system will need the capability to deal with discontinuous elements as in sentence 1; it will have to be able to assign a syntactic description and semantic interpretation to such combinations of lexical items within a particular sentence, independent of the syntactic description and semantic interpretation of the individual items in the dictionary. The same capabilities are required for examples 2 and 3, which represent phrasal and idiomatic expressions. In particular, the system will need the capability of dealing with combinations of lexical items with internal variable slots. The items filling such slots may either not be translated at all, as in examples 6 and 7; or be translated, as in the idioms in examples 9 and 11. (Such items are underlined in the following examples.)

6. *Die Entwicklung nahm ihren Anfang mit ...*
   The development began with ...

7. *Der Aufstand nahm seinen Anfang mit ...*
   The revolution began with ...

8. *Er schoss einen Bock.*
   He shot a buck.
   He made a mistake.[25]

9. *Er schoss einen gewaltigen Bock.*
   He shot a tremendous buck.
   He made a tremendous mistake.

10. *Den Entschluss fassen, etwas zu tun.*
    To seize the decision to do something.
    To decide to do something.

11. *Den festen Entschluss fassen, etwas zu tun.*
    To seize the firm decision to do something.
    To decide definitely to do something.

(We observe in sentences 6 and 7 that the gender of the German possessive pronoun, which has no equivalent in the English translation, is dependent on the gender of the subject.)

The system must also be able to assign a semantic function to the constituents of sentences dependent on the meaning of those constituents and not necessarily on their syntactic function (cf. examples 12 through 16). Thus, the adverbs underlined in the German examples 12 through 14 have to be interpreted as semantic predicates or at least have to be mappable into predicates, given in broken underlines, of the output language; the German dative objects in sentences 15 and 16 appear as English possessives:

12. *Er studiert gern Physik.*
He likes to study physics.

13. *Er studierte lieber Physik.*
He preferred to study physics.

14. *Er sprach weiter.*
He continued to talk.

15. *Er kam ihr zu Hilfe.*
He came to her aid.

16. *Sie brachte es ihm zur Kenntnis.*
She called it to his attention.

(We may note in examples 12 through 14 that the tense of the original German predicate is associated with the English predicate which itself is a translation of the German adverb.)

With respect to the languages German and English, the system should also be able to translate the German article in cases of inalienable property as the English possessive:

17. *Er kreuzte die Arme.*
He crossed his arms.

18. *Er legte ihr die Hand auf die Schulter.*
He put his hand on her shoulder.

We further expect from a translation device that it not only associate a correct semantic reading with a sentence but rather that it provide the correct semantic reading. That is, it should be able to assign to a sentence t all its semantic readings in the case that t is ambiguous and should further be able to select from those readings the one which is correct in the textual environment.

19. *Die Maenner hatten die Frauen ermordet. Wir nahmen sie drei Tage spaeter gefangen.*
The men had murdered the women. We caught them three days later.

20. *Die Frauen waren von den Maennern ermordet worden. Wir nahmen sie drei Tage spaeter gefangen.*
The women had been murdered by the men. We caught them three days later.

21. *Die Maenner hatten die Frauen ermordet. Wir beerdigten sie drei Tage spaeter.*
The men had murdered the women. We buried them three days later.

22. *Die Frauen waren von den Maennern ermordet worden. Wir beerdigten sie drei Tage spaeter.*
The women had been murdered by the men. We buried them three days later.

The problem in examples 19 through 22 is the recognition of the proper referent of the pronoun "sie" in the second sentence of each example. We maintain that none of the four two-sentence combinations are ambiguous. "sie" in examples 19 and 20 uniquely refers to the men; in examples 21 and 22, it uniquely refers to the women. Since both men and women can be captured as well as buried, there is no clue in the semantic reading of the words "men" and "women" which permits the correct association of the proper referent for the subsequent pronoun. Thus, "wir nahmen sie drei Tage spaeter gefangen" in examples 19 and 20, and "wir beerdigten sie drei Tage spaeter" in examples 21 and 22 should be either ambiguous or vague. We can explain the non-ambiguity and non-vagueness of the sentences by the fact that a meaning rule "for all Y: if X kills Y, then Y is dead", is used when the SA-image of the first sentence of each sentence pair is constructed; i.e. that an SA-image is generated in which the argument "women" receives the predication "dead". Assuming that the verb "gefangen nehmen" requires for semantic wellformedness a human object that is alive and "beerdigen", an animate object which is not alive, we can easily explain the establishment of the proper referent. The reader should not be misled by the fact that the English translations of the problematic German sentences display the same ambiguity in isolation. That access to the established SA-image is necessary will be obvious when we translate the sentences into Italian, where the selection of the pronoun *le* or *li* referring to the women and the men, respectively, has to be made.

The problems that have to be dealt with in examples 19 through 22 are, however, not restricted to such apparently constructed examples, which are possibly rare in actual texts, in particular in scientific texts. It is necessary to point out that this problem, in a different appearance, comes up fairly frequently in possibly every text. In the sentences 23 and 24 the predicate *liess ... frei* is translated correctly as *set ... free* in the environment animate (physical) object, and as *left ... blank* in the environment inanimate object, respectively.

23. *Er liess Sylvia schliesslich frei.*
    He finally set Sylvia free.

24. *Er liess schliesslich die Zeile frei.*
    He finally left the line blank.

However, in German and many other languages semantic features of nouns are neutralized when the nouns are pronominalized. Thus, the German sentences 23 and 24 both become sentence 25 under object pronominalization, which, consequently, is ambiguous in isolation.

25. *Er liess sie schliesslich frei.*

The sequences 26 and 27, each of which contains sentence 25, correctly show different translations for 25.

26. *Mark konnte Sylvias Qualen nicht laenger ertragen. Er liess sie schliesslich frei.*
    Mark couldn't bear Sylvia's ordeal any longer. He finally set her free.

27. *Mark wusste nicht, wie er die letzte Zeile ausfuellen sollte. Er liess sie schliesslich frei. Mark didn't know how to fill in the last line. He finally left it blank.*

It follows that for the proper translation of such German sentences, we need to be able to recover the disambiguating semantic features from the contextual information which has been lost due to the pronominalization of the disambiguating German nouns.

It may be interesting to point out that of the 36 selection restrictions associated with the eight verbs in the appendix of my paper "Lexical Features in Translation and Paraphrasing: An Experiment", 13 entries cannot be translated properly if the stated semantic feature for subject or object is neutralized due to pronominalization. This surprisingly high percentage might become even larger if we take into account that the semantic features listed in that paper sometimes are not sufficient for correct interpretation or translation, and additional, more refined semantic features might be required. (Cf., for example, the entry *erhalten*.)[26]

Attempts to solve such problems by assigning to the various translation equivalents a probability, possibly based on criteria of frequency of occurrence, we regard as being unsatisfactory. Assume that an item with two different translations is translated as X in 60% of all the cases and as Y in 40% of the cases. To base translation on their assigned probability will mean that on an average in 100 occurrences of the item we will obtain 40 wrong interpretations and translations. This, moreover, is independent of whether we use the translation X and Y or the translation X alone. In the case that some MT system needs to select translations on considerations of probability, we would regard the restriction of the translation to just the item X as more practical since the user could be warned that X contains a certain margin of error: namely, that it may mean Y in 40% of the cases, whereas, if translations X and Y were used, the user would have to learn that X may mean Y in 40% of the cases and Y may mean X in 60% of the cases.

28. WIE GEHT ES IHNEN? *Mir geht es gut.*
How are you$_{sg}$? I am fine.

29. WIE GEHT ES IHNEN? *Uns geht es gut.*
How are you$_{pl}$? We are fine.

30. WIE GEHT ES IHNEN? *Ihnen geht es gut.*
How are they? They are fine.

Examples 28 through 30, moreover, show that translation of individual sentences based on the information contained in the immediately <u>preceding</u> context is not always possible. The disambiguating information may be provided in sentences which follow the ambiguous sentence. The argument that these examples could be translated correctly if they were not given in the frequent key punch representation which loses the distinction between majuscule and miniscule holds only for English.

28., 29. *Wie geht es Ihnen?*
How are you?

30. *Wie geht es ihnen?*
How are they?

For translation into other languages, as for example Spanish, we still need to be able to access the responses.

(28.) *Wie geht es Ihnen? Mir geht es gut.*
Cómo está Ud? Estoy bien.

(29.) *Wie geht es Ihnen? Uns geht es gut.*
Cómo están Uds? Estamos bien.

It may sometimes not be necessary for device T to have access to the environment in the cases where the ambiguities of the input sentences can be mapped into a corresponding output ambiguity, as examples 19 through 22, 28, and 29, or sentence 31 show:

31. *Johann beobachtete den Mann mit dem Teleskop.*
John watched the man with the telescope.

The capabilities of translation device T would certainly increase if it contained a component which mapped input ambiguity into corresponding output ambiguity, if possible.

Whereas this capability may only be desirable, the corresponding capability to carry over input uniqueness into corresponding output uniqueness is certainly necessary. That output non-ambiguity does not simply follow from input non-ambiguity may be shown by means of sequence 32, where brackets indicate that any, but only one, of the pronouns in the brackets may be used; the subscript of a pronoun indicates that it refers to the word with the same subscript occurring in the preceding text.

32. *Diese Maschine*$_1$ *hat einen Atommotor*$_2$. *Gestern ist eines* $\begin{bmatrix} \text{ihrer}_1 \\ \text{seiner}_2 \end{bmatrix}$ *Raeder*$_3$ *zerbrochen. Wir werden* $\begin{bmatrix} \text{sie}_1 \\ \text{ihn}_2 \\ \text{es}_3 \end{bmatrix}$ *zurueckschicken und Ersatz verlangen.*

A translation which preserves the pronominalization would result in the following sequence:

32a. *This machine*$_1$ *has a nuclear engine*$_2$. *Yesterday one of its*$_{1,2}$ *wheels*$_3$ *broke. We will send it*$_{1,2,3}$ *back and demand a replacement.*

As we can see, this translation introduces ambiguities which do not occur in the German counterpart. The correct translation should be:

32b. *This machine*$_1$ *has a nuclear engine*$_2$. *Yesterday one of the* $\begin{bmatrix} \textit{machine's}_1 \\ \textit{engine's}_2 \end{bmatrix}$ *wheels*$_3$ *broke.*
We will send the $\begin{bmatrix} \textit{machine}_1 \\ \textit{engine}_2 \\ \textit{wheel}_3 \end{bmatrix}$ *back and demand a replacement.*

We finally expect from a good translation device that the syntactic structure of translation u of some input sentence t be isomorphic with or similar to the syntactic structure of t; we also expect that the stylistic evaluation of subgraphs of the structure of t be identical with the stylistic evaluation of the corresponding graphs of the translation u of t. Both statements, of course, are to be understood with the proviso that such corresponding, similar structures or stylistic evaluations occur in both languages.

So far, none of the examples mentioned have provided us with counterevidence to the claim that translation is possible by mapping molecular lexical items into equivalent molecular items. How shall translation device T react if it meets a molecular expression in one language which has no corresponding equivalent equimolecular expression in the target language, as predicted by adherents of the Humboldt-Cassirer-hypothesis, also called Sapir-Whorf-hypothesis?

Two solutions are possible: T may contain a dictionary in which two or more molecular expressions of the target language are given as the equivalent of the molecular expressions in the source language or - to quote Professor Bar-Hillel - by permitting the system to "tell a story". The first way is normally selected in dictionary entries, though very often not very successfully, as translations like that of the German entry *jemandem etwas absehen* illustrate. Wildhagen gives the translation equivalent *learn something by looking at a person*, Langenscheidt, *learn something from a person*.[27] According to these translations, the German sentence *Er hat seiner Mutter das Schoenschreiben abgesehen* would be translated as *He learned calligraphy by looking at his mother* (Wildhagen) or *He learned calligraphy from his mother* (Langenscheidt), whereas the exact translation should be *He learned calligraphy by watching his mother do it.* The first dictionary translation does not express the fact that there is a causal relation between someone's learning some action or behavior and his watching someone do it. The second translation does not indicate the fact that this someone is performing the action or displaying the behavior. A better translation would consequently have been: *to learn doing* x *by watching someone do* x, and/or: *to learn to be* x *by watching somebody be* x. Assume now that a translation for term q cannot be provided because the dictionary - due to lack of any translation equivalent - does not contain a translation for q. (We do not know of such occurrences.) In this case, System T needs to be equipped with the capability for describing the SA-image representing term q. This, however, can be simulated by permitting System T to have access to its meaning rule component, where it can read off the definition for the term in question. This, again, means that the user of the MT system can update the bilingual dictionary by providing as a translation the equivalents of the terms used in the definiens of the definition of q.

Real problems will arise only if a state of affairs is described in the source language which simply cannot be described by any language-means in the target language. In this case, both human and mechanical translation would be impossible. We doubt that this will happen, in particular, in scientific texts.[28]

We finally investigate whether "self-awareness" is required for translation device T. This may be discussed by means of an example which was given by Roman Jakobson during one of the conferences pertaining to the Study. In Polish as in other Slavic languages the equivalent of "I" is normally omitted, but stated in emphasis. In one of them (Czech, if I recall properly), the opposite is the case. A translation of a Polish text: *Whenever he spoke of himself, he* *__used__* *the word 'I'.* into Czech should read: *Whenever he spoke of himself, he* *__omitted__* *the word 'I'.* (Note that the translation of Polish *__I__* *am speaking* into Czech *(I) am speaking* (where underlining indicates occurrence of the pronoun in the surface; enclosure in parentheses, absence in the surface) is not beyond the capabilities of the device; this could be handled by the semantic or, possibly, the stylistic component.) Clearly, the correct translation of such examples requires that the system contain the ability to interpret statements about itself or part of itself and associate those statements with the corresponding parts of that system. The system would thus have to be able to 'think' about itself or some of its parts. This capability, artificial intelligence, we do not regard as necessary for an MT system for some time to come.

The gravest argument against the possibility of mechanical translation has been the claim that knowledge of the world and even knowledge of the subject matter is required for the translation of a text. This argument, reformulated for our device T, reads: There are sentences whose ambiguity cannot be resolved by access to the immediate preceding or following textual environment. Sentences 33 and 34 may represent such ambiguities:

33. *Fred and John had beaten Mary and Jane so brutally that we had to take them to a penal camp.*

34. *Fred and John had beaten Mary and Jane so brutally that we had to take them to a hospital.*

It seems obvious that we understand these sentences correctly, i.e. that we can determine the proper referent for *them* (necessary, e.g., for their proper translation into Spanish *los* and *las*) because we have stored knowledge about certain typical "sequences of states of affairs". The fact that <u>we</u> understand these sentences in isolation does not mean, however, that MT device T must have the same capability. Very often the preceding and/or following context may contain - for us redundantly - information which permits the disambiguation of such sentences. Consider for example as a continuation of sentences 33 and 34, respectively:

33a. *After three weeks Fred and John were released from the camp.*

34a. *After three weeks Mary and Jane were released from the hospital.*

Consider even the "counterevidence" given in the following sentences 33b and 34b:

33b. *There they were safe from Fred and John.*

34b. *There they posed no more danger to Mary and Jane.*

As we can see, our knowledge about typical sequences of states of affairs permits us to draw conclusions with some, normally high, probability but not with absolute certainty. (This probability may be 100% when the relation between states of affairs is a cause-effect relationship.)

A difficulty of a different nature is represented by the fact that certain terms have a different translation dependent on the particular subject area they pertain to.

35. *John had always wanted to become a conductor. (bus, orchestra)*

But again, we might expect continuations like:

> 35a. *He attended every performance of the local orchestra and watched the conductor with admiration.*

or:

> 35b. *As often as he could, he rode in a bus and watched the conductor with admiration.*

We do not intend to belittle these difficulties confronting successful mechanical translation. On the other hand, we believe it is fair to point out that no research has been performed to find out the extent to which the preceding or following context provides the information necessary for the proper disambiguation for such sentences. We do, however, believe that sentences do not occur in isolation, at least not in material presented for translating, and that the required factual knowledge may be replaceable by access to the information contained in the contextual environment. If difficulties should arise because the device, instead of printing out all readings in such cases, prints out just one with a warning signal, we may still rely on the powers of the reader to interpret statements pertaining to a subject area he is well acquainted with. Let us now recapitulate the properties that we expect MT device T to have:

a) It must be able to assign to a source language sentence t all its syntactic descriptions and all its semantic readings. This might be done without a genuine semantic component provided that the "semantic function" of the arguments, i.e.

their location on the numbered lines in representations as in Figure 12, page 21, can be computed from the syntactic structure and the information associated with the lexical items occurring in that structure; this, we are inclined to believe, is possible. (Cf. also Fillmore's arguments in "The Case for Case".) T will, however, have to contain a transformational component which permits at least permutations and deletion recovery (for the source language), and permutations and deletions (for the target language).

b) It must be able to map the lexical items and the semantic relations expressed in t into the equivalent equimolecular lexical items and semantic relations of the target language sentence t'. This requires either a translation component: Source language → Target language, or an interpretation component: Natural language ↔ Interlingua, for each of the languages involved in the translation process.

c) It must be able to derive at least one sentence t' with its syntactic description from the semantic reading of t'. The syntactic structure of t' will have to correspond to the syntactic environment required by the lexical items in t'. This again - for each language involved - requires an extensive dictionary with sufficient syntactic and semo-syntactic information for every entry.

d) T must further be able to disambiguate sentence t based on the contextual information preceding and/or following t.

This definitely requires aa) the association component of K, bb) the capability not to be restricted to sentence-by-sentence translation, and cc) a lexicon in which terms with different meanings in particular areas of provenience - which are not disambiguable by means of semantic features - are equipped with area of provenience information (remember *conductor* in example 35, page 45). Device T, of course, needs the capability for exploiting such area of provenience features.

e) T must have access to the definitions of a meaning rule component. This requirement can be replaced and, for the time being, should be replaced by updating the source-target language dictionary by providing a combined translation in the target language of the terms of the definitions of the "difficult" item in the source language; this combination can be treated as <u>one</u> lexical item, possibly with internal variable slots (cf. examples 6 and 7 in this chapter).

In addition, the following properties are desirable:

f) It should be able to provide a translation t' for sentence t whose syntactic description is identical or similar to the syntactic description of t. This requirement means that the system must be able to associate with some semantic representation R all target language sentences (with meaning R') with their syntactic description. In uni-directional translation, this requirement may be limited to only those structures which are isomorphic or similar to structures occurring in the input language.

g) It should be able to provide a translation t' for sentence t whose stylistic evaluation is identical or similar to the stylistic evaluation of t. This means T should have a stylistic component which can possibly be simulated by stylistic features associated with lexical and syntactic structures.

h) T should be able to associate a translation t' with a sentence t in such a way that, if t is ambiguous in some specified fashion, t' is ambiguous in the same fashion. This desired property of MT system T, complementary to requirement d, is really a makeshift solution, proposed because of the current but, hopefully, passing inaccessibility of the information provided by the context to mechanical devices.

i) Finally, T should be able to produce a non-ambiguous translation t' for a non-ambiguous sentence t.

As we see, an MT device should incorporate a greater part of the components of a comprehension device and some additional components pertaining to the output in a foreign language to provide syntactically similar and stylistic translations. We are not able to say whether a translation device needs to have access to a long-term memory or an "encyclopedic knowledge" component. Examples which clearly show this necessity for a comprehension device or an information-retrieval system may not be relevant for an MT system.

We conclude that translation by mapping semantic relations between molecular or atomic expressions of the target language into equivalent equimolecular expressions (or combinations of expressions), under preservation of the semantic relations, is possible. Such translation can, in general, be performed on the level of semantic readings (DSA-images). Access to the short-span memory, the association component of K, to select the proper reading in cases of ambiguity, will be necessary. The extent to which access to the association component cannot be avoided, or to which this necessity can be replaced by relying on the intellectual capabilities of the reader of the translation has not been investigated, so far.

We shall discuss in the subsequent chapter which of the better known, current linguistic theories account for the requirements that we expect from such an MT device or, at least, the extent to which they account for them.

4. The Capabilities of Current Competence Models or The Properties of a Realizable Mechanical Translation Device

In the preceding chapter we gradually developed the properties of a hypothetical MT device T, based, in part, upon the linguistic problems occurring in translation which T must be capable of solving, and, in part, on certain esthetic expectations. These require that T carry across into the target language the message to be translated, in a way closely corresponding to the structure and the evaluation associated with the message in the source language. In this way, we increased the capabilities of MT device T until it approximated to some extent the capabilities of a human translator.

In this chapter we want to determine the extent to which these hypothesized capabilities are actually realizable within the framework of the current better known grammatical models. The models we have in mind are: a) the various realizations of transformational grammar, as the "standard" model; the "extended standard" transformational grammar; and the "universal base hypothesis", the transformational grammar with a generative semantic base component, b) the case grammar of Fillmore, and c) the dependency grammar, which in several variations is prevalent in European and Soviet approaches to MT.

All of these models have been defined, explicitly or implicitly, by their proponents or adherents as competence models, i.e. abstract devices which enumerate an infinite list of individual or non-coherent sentences. Competence models are regarded as components of performance models which account for such human capabilities as the production and understanding of sentences in actual speech situations or simulations of them, i.e. the production and understanding of coherent sentences.

These limitations of the capabilities of a competence model limit the capabilities of our MT device T. The main requirements which cannot be met in current competence models are:

requirement d (page 47), the disambiguation of sentences based on the information given in the context;
requirement c, the derivation of sentences from their semantic representations;
requirement e, the production of translations for source sentence - target sentence pairs whose semantic representations contain equivalent combinations of items with different internal molecularity, by means of a meaning rule component;
requirement g, the production of translations which have the same stylistic interpretation as the corresponding source language sentences.

It seems that the "universal base hypothesis" grammar is theoretically able to account for requirement c, the derivation of sentences from their semantic representation, provided there are efficient production (and recognition) algorithms. This is due to the fact that the deep structure they propose, common to all languages, represents the meaning of the sentences transformationally derivable from those deep structures by means of the rules of the transformational components of the individual languages. Since, however, this model has only been scarcely described - by means of a few examples restricted to English - we arrive at the conclusion that none of the requirements stated above can be met by the current competence models.

We are thus confronted with the choice either of attempting to simulate or construct a component of a performance model which permits us to meet requirement d and possibly c (we assume we can dispense with requirement g without considerable loss to the quality of the translation) or of lowering our requirements for MT device T to make it compatible with the current capabilities of the existing competence models. The latter possibility is the one normally taken by proponents of MT and automatic information retrieval. For MT it means that the original definition of translation as an association of source language sentence t, with the meaning R(t), into the corresponding target language sentence t', with the same meaning, i.e. as a mapping of meaning into meaning, is changed to a definition of translation as an association of the lexical items in t and the syntactic structures which interpret them with the corresponding lexical items in t' and the corresponding syntactic structures interpreting them in the same fashion.

Clearly, the powers of the hypothetical MT device T have been considerably reduced: not all paraphrases can be accounted for. In additior, the restricted device cannot account for verbal phrases and idiomatic expressions if they do not have a literal correspondence in the other language. (To my knowledge, Gruber's proposals have not been incorporated into transformational grammars or any other of the mentioned grammars.)[29] From a practical point of view, however, we can assume, based on experience in translating actual texts, that this restriction may still provide generally satisfactory translations, especially for languages whose syntactic structures are similar.

What are now the theoretical requirements for such a translation process? We need to be able to associate with a source language sentence t a syntactic representation, preferably the deep phrase marker; we need to map this representation into the corresponding deep phrase marker of the target language, and we need to derive from that deep phrase marker, by means of transformation rules, the corresponding surface sentence t'.

Though the algorithms which perform such recognition, mapping and production have been described and have been in existence for several years,[30] no machine translation system has been produced. This is due to two facts: the lack of comprehensive grammatical descriptions for any language and the lack of a component which is part of all four competence models: A lexical component in which for each lexical item two types of information are listed:

a) its own syntactic and semo-syntactic properties, and

b) the syntactic and semo-syntactic properties of the environment in which it may occur.

Confronted with this gap, we again have two choices: to lower the requirements for an MT system even further by allowing a lexical component which does not contain such features, or to construct such a lexical component, a difficult, tedious, and time-consuming task.

The first choice, in spite of the intermediate development of transformational recognizers would lead to systems which perform only slightly better, as experience has shown, than the ones criticized in the ALPAC report.[31]

Thus, really only the second choice is open for a designer of an MT device: he has to rely on a complete lexical description of the languages that he is dealing with; he has to construct his own featurized lexicon or hope that somebody else may have produced one from which he may be able to profit.

This decision is independent of whether he is happy with the capabilities of the current competence models or whether he wants to simulate additional capabilities of a performance model by permitting access to the contextual environment, i.e. whether he wants to perform research in discourse analysis.

What sort of approximations to the additional capabilities of device T can we expect from a restricted hypothetical MT device T' which performs mechanical translation based on a lexicon with features and a grammatical description of the languages involved in the translation process? (We do not share Petrick's opinion about the length required and the extent of difficulties involved in the construction of comprehensive grammars; we believe that his pessimism is based on the fact that he considers the difficulties primarily from the point of view of transformational grammars.)[32]
Those additional requirements are:

requirement f (page 48), syntactic similarity of source and target sentence structure or at least preservation of the relative order of the lexical equivalents; furthermore,

requirement h, the carrying across of lexical and/or syntactic source sentence ambiguity; and

requirement i, the carrying across of source sentence non-ambiguity.

The first requirement might be met by establishing additional correspondences between the relevant reverse (source language) transformations and the order in which they apply with the corresponding forward (target language) transformations (in opposite order). We have no opinion on how these correspondences should be established. The checking of the coincidence of the relative order of the corresponding lexical items may be easily incorporable into T' and may thus serve as a means to select one translation from a set of transformationally related translations.

The second requirement would mean that from the translations, i.e. the sets of surface sentences:

$$A_1 = \{t'_{1,1}, t'_{1,2} \ldots t'_{1,m}\}$$
$$A_2 = \{t'_{2,1}, t'_{2,2} \ldots t'_{2,k}\}$$
$$\vdots$$
$$A_n = \{t'_{n,1}, t'_{n,2} \ldots t'_{n,j}\}$$

the one occurring in each or in the greatest number of the sets A_1 through A_n would have to be selected (where source sentence t has the deep phrase-markers DM_1, DM_2, $\ldots DM_n$). Clearly, such a procedure would not be practical.

The third requirement would mean that T' would have to generate all sentences generatable from the mapped deep structure, analyze each of the generated surface sentences again by means of the input component of the target language and select one of those sentences which have only one deep phrase structure representation.

We thus also relinquish requirement h and the first part of requirement f. (The abandonment of the second part of requirement f, preceding page, would possibly impose too heavy a burden on the powers of the reader to interpret correctly.) Within the capabilities of the current competence models translation by means of MT device T' can thus be represented as a sequence of three processes:

1) recognition of the deep phrase-marker(s) of source language sentence t,
2) mapping of the deep phrase-marker(s) of t into the deep phrase-marker(s) of t',
3) production of some target language sentence t' from (each of) the phrase-marker(s) of t'.

We assume that such translations may be satisfactory, especially if performed between related languages. In view of the problems which will confront such a translation procedure (cf. Chapter 3), we regard MT device T' as an intermediary solution. We personally feel that the model which should be strived for is MT device T. In the following chapter we shall describe an approximation to such a device T, the Linguistics Research System.

5. The Linguistics Research System

"Everything in nature, in the unorganic world as well as in the organic world, happens according to rules, though we do not always know these rules ...

The use of our capabilities also occurs according to certain rules which we follow, at first unconsciously, until gradually, through attempts and continuous usage of our capabilities, we obtain a knowledge of them, even acquire such a fluent usage of them that it takes much effort to imagine them in the abstract. Thus, e.g. the general grammar is the form of a language as such. But one does speak without knowing the grammar, one has indeed a grammar,and speaks according to rules, but one is not conscious of them.

Like all our capabilities, our reasoning is subjected in its actions to rules which we can investigate." (Translated from the first through third paragraphs of Kant's Introduction to his *Logik*.[33]

The purpose of the Linguistics Research System (LRS), which is currently being constructed at the Linguistics Research Center of The University of Texas at Austin, is to provide a description and an explanation of human linguistic capabilities by performing recognition and production of sentences in natural language, mechanical translation, and information retrieval. LRS is a system of components which can be connected like building blocks to form larger configurations. Each component consists of a set of algorithms and instructions which are executed by the algorithms; they modify the general operations of the algorithms in a prescribed way. Such instructions are linguistic rules, dictionary rules, syntactic rules, interpretation rules; transformation rules, meaning rules, mapping rules, connection rules, and others.

In its basic configuration LRS is a grammatical model for the recognition and production of synonymous sentences in natural language with identical or different deep structures. By deep structures we mean the stage of a sentence derivation in standard transformational grammar when all base component rules, constituent and feature re-writing rules, have applied but before lexical insertions have been performed.

The purpose of this model is to associate with each sentence in a natural language all its canonical form (KF) representations. A sentence which has one semantic reading has one canonical form, a sentence which has n semantic readings has n canonical forms. Two sentences t and u which have one semantic reading in common have one canonical form in common. Two sentences t and u of the same language which have one canonical form k in common are called paraphrases in the reading k. Two sentences t and u of different languages which have the canonical form k in common are called translations of one another in the reading k.

LRS has the power of an interpretative semantic model in that it assigns the same KF reading to synonymous sentences with different deep structures. It has the power of a generative semantic model in that, given a particular KF reading k, it permits the generation of all sentences with different deep structures with that reading k.

A canonical form consists of a sequence of connected canonical form expressions (KF expressions). The language of canonical forms K has the following properties:

a) Each KF expression is a primitive element of K; (it has - for the user - one and only one (atomic) semantic interpretation); if a surface terminal k has n different senses or meanings, then n different KF expressions or connected KF expressions represent the different senses of k.

b) No two different (connected) KF expressions p and q are synonymous. If two surface terminals have one sense in common, then that reading is represented by the same (connected) KF expression.

Numerous statements have been made in history as to whether such a canonical language can be constructed. Counterarguments have mainly been given during the last few decades by proponents of the Humboldt-Cassirer hypothesis. Assuming that the "world views" of different natural languages are indeed different, a universal language can hardly be more than the logical sum of the different world views, which, however, should not be a reason to abandon this notion. However, compare Catford: A *Linguistic Theory of Translation. An Essay in Applied Linguistics*, and Hjelmslev: *Prolegomena to a Theory of Language*.[34]

Due to the lack of a theory of semantics applicable to the mechanical recognition and production of sentences in natural language and because of the immense difficulties involved in the construction of canonical forms, LRS represents the meaning of sentences by means of normal forms.

The normal forms of a language are distinct from canonical forms of a language in that the lexical primitives of normal forms may be both atomic and molecular with respect to the canonical forms, for example, *bachelor*$_1$, *unmarried man*, *unmarried human adult male*. When information retrieval or translation from any language into any language is attempted, the normal form representations will either have to be replaced by canonical form representations or, more economically, the meaning rule component will have to be expanded to permit the construction of the particular required canonical form when logical conclusions have to be found, or when different languages partition the "world" differently. Cf. Latin *patruus* (father's brother) and *avunculus* (mother's brother).

The process of associating with a surface sentence t all the normal forms of t is performed in three steps. To each step there corresponds a component:

The surface component, the standard component, and the normal form component.

One grammar, the surface grammar, the standard grammar, and the normal form grammar, is associated with each component. The non-terminal and terminal vocabulary symbols of each grammar are complex symbols (except for the terminal symbols of the surface grammar). Each complex symbol consists of a category symbol and zero or more subscript or feature symbols; each subscript may have zero or more values.

The grammar rules used during the recognition and production of sentences, both performed as a bottom-to-top direct substitution analysis, are generated by the processing algorithms by means of instructions represented as context-free rule schemata. A constituent in the consequent of a rule schema matches every analyzed (WS) complex symbol from which it is not distinct, i.e. it may match a whole complex WS symbol or a part of a complex WS symbol. A rule schema is successfully applied if each of the positive and negative conditions for each constituent in the rule schema is fulfilled by the matched complex WS symbol, and if all the required relations between two or more constituents stated in the rule schema hold between the corresponding complex WS symbols. If a rule schema is successfully applied, a new WS constituent is constructed according to the instructions stated in the antecedent of the rule schema.

The conditions that may be stated for individual constituents in a rule consequent are:

a) A particular category symbol may not or must contain a particular subscript or combinations of subscripts.

b) A particular category symbol may not or must contain a particular value or combinations of values.

c) Operations between subscripts of different constituents may not or must be successful. These operations, the set-theoretical operations Intersection, Sum and Difference, are performed with the values of the specified subscripts.

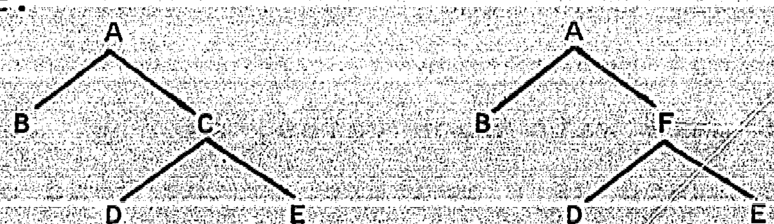
Each rule schema of each grammar consists of a syntactic part and an optional transformational part. For surface and standard grammar the syntactic part of each rule schema consists of context-free rewrite rules. The transformational part contains only transformations whose structural description is satisfied by a string of symbols interpreted by the constituents of the rule schema consequent. The transformations possible in surface and standard grammar are permutations, deletions, and insertions. The transformations are "feature-sensitive"; in particular, it is possible to lexicalize features of a particular constituent and to "featurize" terminal or non-terminal constituents. Thus, words like *up* which form a lexical unit with some verbs, as e.g. *look something up*, are assigned as a feature to the head of the verbal construction, resulting in *look something*.
+up

The rules of the normal form component differ from surface and standard rules in two respects:

a) They apply to connected graphs;

b) They are not rewrite rules.

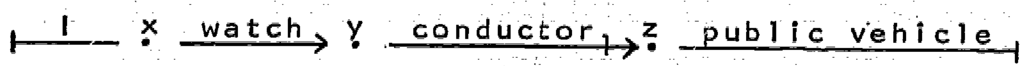
An NF rule applies to all graphs, terminal, non-terminal, or combinations of them, whose nodes, labeled by complex symbols, are non-distinct from the complex symbols in the consequent of the NF rule. The antecedent of the NF rule assigns to all graphs to which it applies a particular semantic reading, an NF expression, represented by that antecedent. Since NF expressions apply to graphs whose nodes are labeled by complex symbols, it is possible to assign a particular NF reading to a terminal k with a particular part of speech interpretation and with a particular selection restriction. At the same time, all graphs t_1, $t_2 \ldots t_n$ interpreted by the same NF expression k are substitutable for one another, regardless of whether the root and end nodes of t_i are identical or different from those of t_j $(1 \leq i,j \leq n; i \neq j)$. (It is theoretically possible that t_i and t_j have identical root and end nodes and still be different, cf.



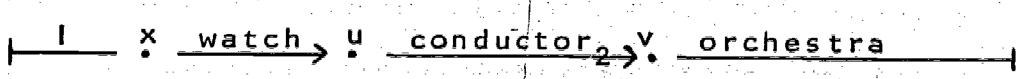
).

The normal forms of an ambiguous sentence t may be connected by means of "or" links, resulting in one connected normal form. Assume that the normal form of each of the following sentences is represented by the associated graph.

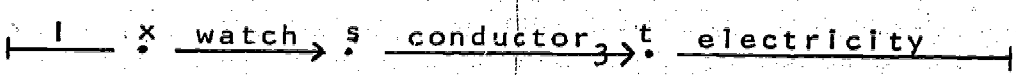
I watched a public vehicle conductor.



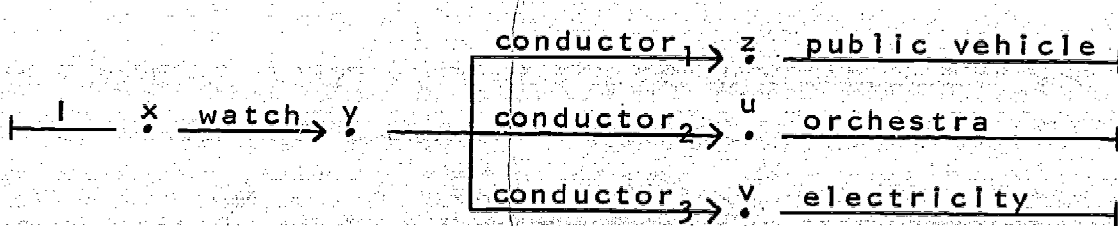
I watched an orchestra conductor.



I watched an electric conductor.



Then *I watched a conductor* can be represented as:



An "or" link is represented by a line which meets or intersects a labeled line at a right angle. (These graphs are simulated, simplified representations of the actual normal forms. For the actual normal form representations of such sentences cf. p. below.)

It is the function of the surface component to assign to each surface sentence t all its syntactic readings according to the surface grammar; ambiguous lexical items which have the same part-of-speech interpretation are represented as one "conflated" lexical item in a surface reading. After surface analysis all readings which are not dominated by the initial symbol S are deleted; then all transformation instructions contained in the remaining rules are executed. They associate with each of these readings a tentative standard string. Tentative standard strings consist of complex standard terminal symbols; these may be conflations of surface terminals and their (possibly disambiguated) dictionary interpretation, and dummy symbols which are introduced by the transformations of the surface rules which applied. Dummy symbols represent grammatical morphemes and elided lexical items. Elements which were discontinuous in the surface are contiguous in the tentative standard strings.

These strings are then analyzed by the standard grammar which assigns them a standard description and also filters out all those strings which are not well-formed according to the standard grammar.

The readings of the remaining standard strings are then analyzed by the NF grammar which assigns NF expressions to individual standard subtrees or combinations of them.

It is not necessary that the roots of the graphs interpreted by the same NF expression are labeled by the same category symbol. It is thus possible to define adjectives and nouns, e.g. *sun - solar, spectrum - spectral,* as synonymous in one reading by assigning each member of such pairs the same NF expression. The same holds for adjectives and verbs, e.g. *bright - to shine* or nouns and verbs, e.g. *destruction - destroy,* etc. It is also possible to define synonymy relations between lexical units and idiomatic expressions like *die - kick the bucket* or lexical units and phrasal expressions like *strike - give a blow - receive a blow* or *kill oneself - commit suicide*, etc. In the latter examples the actual synonymy relation is established between the verb *strike* and the noun *blow*, or between the verb *kill* with the feature reflexive and the noun *suicide*. The verbs *give, receive* and *commit* are introduced as empty verbal place holders; in addition, *receive* is defined as the logical converse of *give*, permitting such paraphrases as *Mary hit John, Mary gave John a blow, John received a blow from Mary.* It is also possible to define synonymy relationships between lexical pieces which have an internal variable slot without affecting their transformational possibilities.

To be paraphrases of one another, it is not necessary for two sentences t and u that each lexical piece in t be synonymous with some lexical piece in u and vice versa;

this may be realized from such generatable paraphrases as *All men are not virtuous - No man is virtuous*, etc. or *He overlooked this - He did not take this into account*, etc. or *A precedes B - B follows A*, etc. or *A is larger than B - B is smaller than A, B is not as large as A*, etc. How LRS assigns such paraphrases the same NF reading can be found in Lehmann - Stachowitz, 1970, Vol. II, pp. T217-268.

During production, the recognition process is reversed. Each NF expression k is replaced by all the standard rule schemata interpreted by k. The standard grammar rules thus obtained and only those are used for the generation of standard strings in a regular bottom-to-top recognition process. The combinations of all graphs which are connected with a root labeled by the symbol S represent the legitimate standard readings; all others are filtered out.

The terminal standard strings obtained from each well-formed standard reading are then analyzed by the rearrangement grammar of the language which

a) arranges the standard terminals in surface word order,

b) deletes the standard dummy symbols, and

c) re-introduces lexical pieces which are deleted after surface analysis.

In addition, the rearrangement grammar filters out all strings which are not well-formed according to its rules.

This basic component of LRS just described is based on the following linguistic assumptions:

1) that grammatical relations can be more easily and correctly stated for standard strings;

2) that surface information is necessary for correct semantic interpretation;

3) that synonymous sentences can be reduced to the same "universal" representation.

This component is part of the Linguistics Research System for Mechanical Translation and the Linguistics Research System for Information Retrieval.

In the remainder of this chapter, we will cursorily describe those components of LRS which are essential for performing mechanical translation of sentences in natural language. More detailed information can be found in our forthcoming report Lehmann - Stachowitz, 1971a, and in Lehmann - Stachowitz, 1970, Vol. II. The components of LRS pertaining to an information retrieval system are described in Lehmann - Stachowitz, 1971b.

Based on the problems represented in the examples of Chapter 3, we assume that high quality translation has to be based on the following kinds of information; these are:

textual information,

co-textual information, (and possibly also)

contextual information.

These terms correspond to the usage of Catford (op. cit.).

Contextual information is that type of information which can be derived from the speech situation, the belief systems and world knowledge of speaker and hearer. Terms also used to denote this type of information are: "Pragmatics", "pragmatic information", "socio-psychological information". Co-textual information refers to the speech acts that precede and follow an utterance. In case of a written utterance, co-textual information is represented by the written utterances which precede and follow the given utterance. Textual information is that information available from an utterance or a written utterance itself when contextual and co-textual information are ignored.

Translation based on all three types of information we regard at present as being beyond the requirements for an MT system; the situation may change, though, once intensive research in discourse structure shows the necessity for it.

The LRS translation system performs translation based on textual information derived from the basic input component and on co-textual information contained in the immediate environment of a sentence derived by means of an approximation of the short-span memory mentioned on page 20, Chapter 2. We have observed that LRS is capable of producing various paraphrases. This capability, though desirable for Information Retrieval purposes, may not always be desirable when performing translation, even if the connections of sentences with the preceding and following textual environment are properly preserved. Thus, we would prefer to translate *A geht B voraus* as *A precedes B* rather than as *B follows A*, or *Alle Menschen sind tugendhaft* as *All men are virtuous*, rather than *No man is not virtuous*, or *A verkauft B* as *A sells B* rather than *B is sold by A*.

This capability is obtained by means of the fact that NF-expressions are represented as complex symbols containing essential and accidental features. Essential features pertain to properties of an interlingua, accidental features to the properties of a particular language represented by lexical pieces and syntactic structures. Thus, the various graphs in Figure 12 which we repeat below as Figure 16, are all representations of the NF-expression 29.

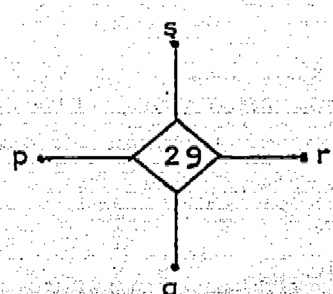


Figure 15

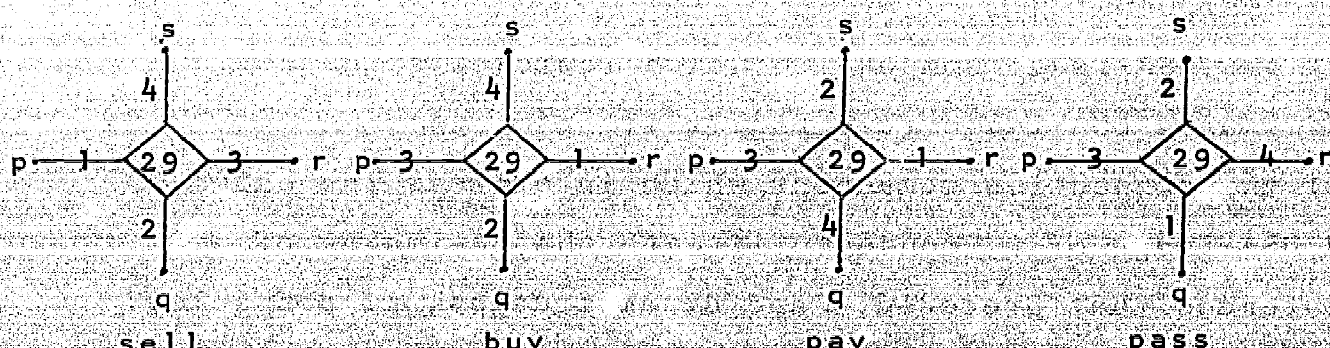


Figure 16

The numbers in Figure 16 represent accidental features. They permit a more precise translation as, for example, from or into the German counterparts represented as:

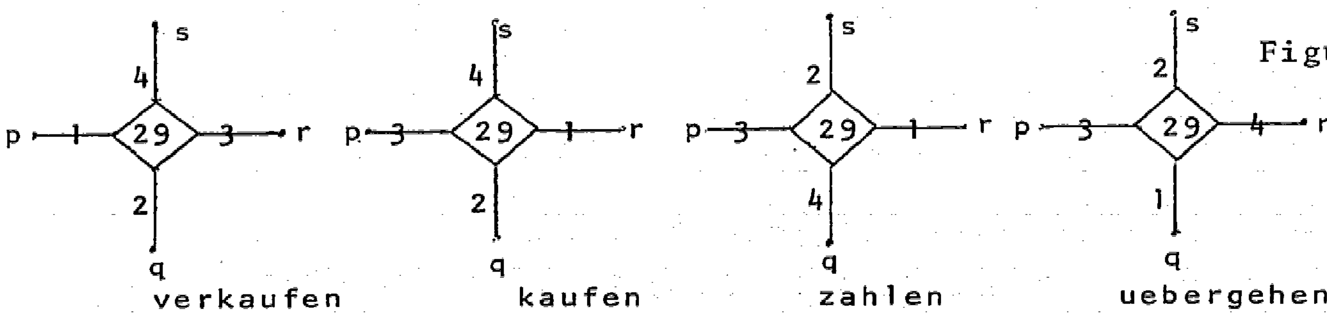


Figure 17

Similarly, syntactic information like active sentence, passive sentence, can be added by means of accidental features to the NF expressions which interpret these structures. (If an NF expression cannot be mapped into an identical (i.e. including accidental features) NF expression of the target language, all NF expressions in that normal form are mapped by means of only the essential features).[35]

Machine translation is performed by means of the following components:

1) the basic recognition component, which derives the normal form of surface sentence t, or the normal forms of t if t is ambiguous;

2) the DSA-image component, which represents the normal form of t as a DSA-image (cf. Figure 11, page 20);

3) the connection component, which interprets the established DSA-image of t, connects it with the SA-images of the sentences that preceded t, and disambiguates, if possible, the normal form of t;

4) the mapping component, which maps the normal form K of t into the normal form K' of the target language;

5) the production component, which produces, by means of the grammars of the target language, a translation t' of t.

Let us represent this translation process by means of the sequence of sentences 38 through 40:

38. *Im Museum sahen wir einen Leiter.*

39. *Den Leiter schaute sich eine alte Dame an.*

40. *Sie zerbrach ihn.*

The corresponding English translation of the individual sentences in the sequence is:

38a. *In the museum we saw a leader [or: conductor (animate), conductress, head, chief, executive, manager, manageress, president, director, directrice, superintendent, principal, conductor (inanimate)].*[36]

39a. *An old lady looked at the leader [or: conductor (animate), conductress, head, chief, executive, manager, manageress, president, director, directrice, superintendent, principal, conductor (inanimate)].*

40a. *She (it) broke him (it).*

Let us assume that the German sentence 38 of this sequence has already been analyzed and resulted in the following SA-image:

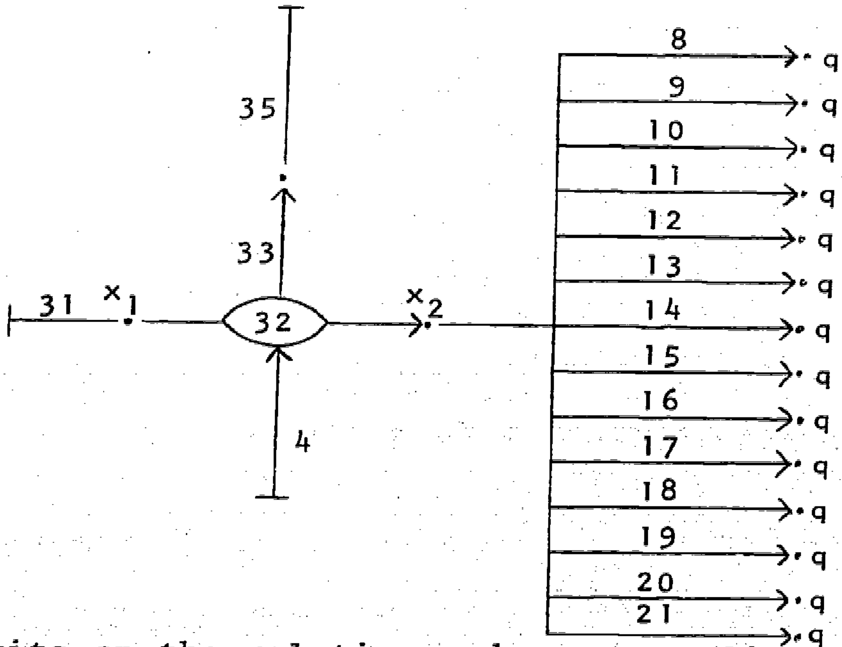


Figure 18

The digits on the relation and property lines represent molecular expressions. 31 may stand for "we"; 32 for "see"; 33 for "inside of"; 35 for "museum"; 4 for "past"; 8 for "leader"; 9 for "conductor (animate)"; 10 for "conductress"; 11 for "head"; 12 for "chief"; 13 for "executive"; 14 for "manager"; 15 for "manageress"; 16 for "president"; 17 for "director"; 18 for "directrice"; 19 for "superintendent"; 20 for "principal"; 21 for "conductor (inanimate)".

The input component, when processing sentence 39, assigns to this sentence a surface description which we represent, in a simplified manner, in the following graph (for a detailed description of a surface analysis cf. A. Stachowitz "Es liegt eine Anzahl von Elementen vor"):

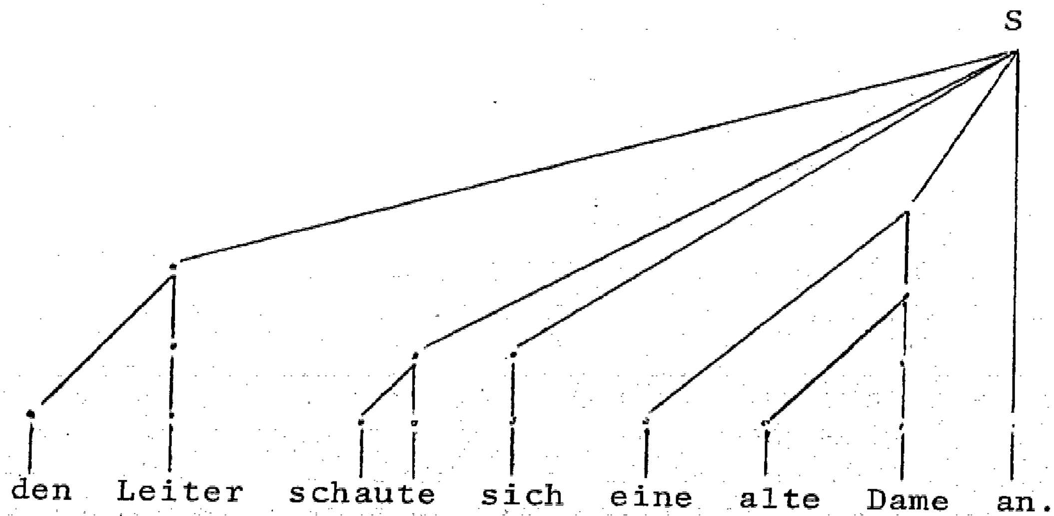


Figure 19

The semo-syntactic information associated with the rule which interprets the word *schauen* given in Figure 20 is exploited by the transformation instructions associated with the rule which rewrites the symbol S.

```
C V                        schau
+ PR(0'1'2'...)
+ TS(:':'AN'...)
+ SS(:':'2'...)
+ CG(:':'D.A'...)
+ TO(:':'R.PO,AB'...)
+ SO(:':'0.3'...)
```

Figure 20

This rule represents all (prefix-) verb combinations which contain the verb *schauen*. The symbol C identifies the category symbol VERB; subscripts are identified by a "+": PR stands for "prefix", TS for "type of subject required", SS for "deep order of subject", CG for "case government", TO for "type of object required", SO for "order assigned to objects". The expressions within parentheses are values; 2 stands for the prefix "an". AN for "animate", D for "dative", A for "accusative", R for "reflexive", PO for "physical object", AB for "abstract". The "." indicates that the verb takes two objects; a "," represents logical or;[37] the digits in SS and SO represent the order assigned to the subject and the objects. (The verb always has the order 1, the deep subject order 2, etc.) The value 0 expresses the fact that the reflexive object in the dative is to be deleted. (This deletion is only performed for genuine reflexive verbs.) The apostrophes represent columns in the "feature matrix" of the verb.

By means of this information and the transformation instructions associated with the nodes in the sentence, the following standard string is derived.

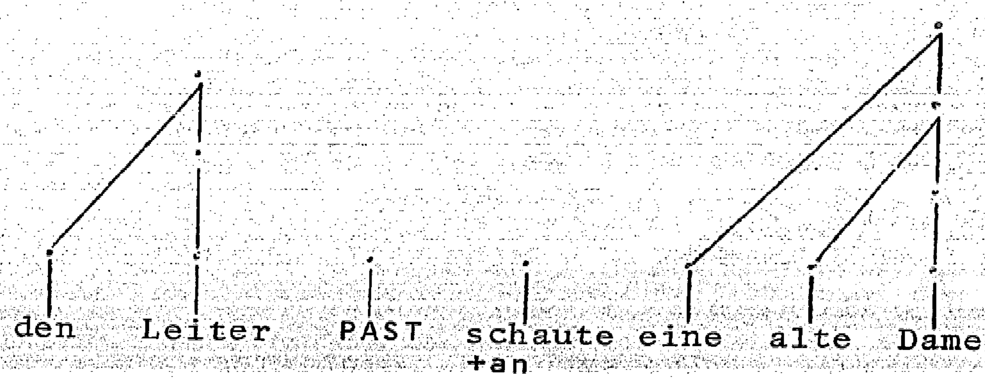


Figure 21

After applying the standard rules, the following structure is derived:

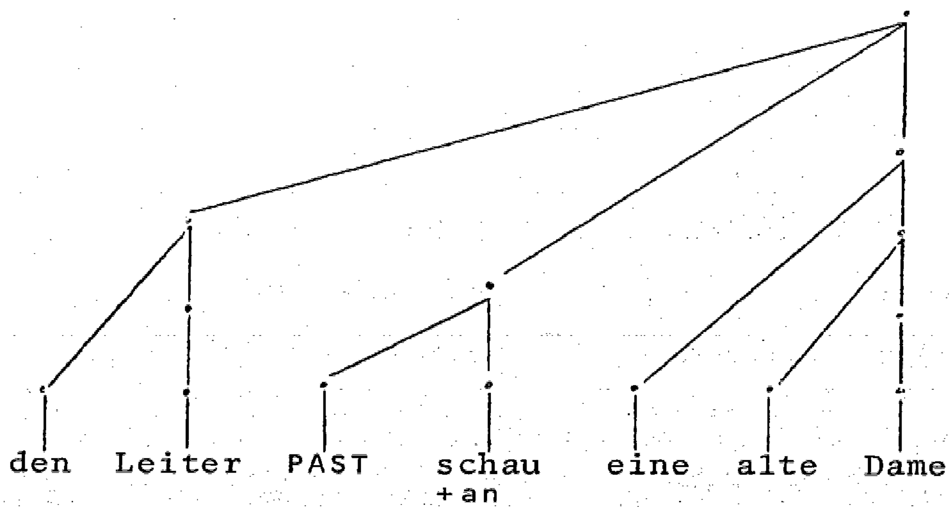


Figure 22

To this structure, the rules of the normal form grammar apply, which derive the following normal form:

$$\overset{1}{R(p,q)_3} \times \overset{2}{Time_2} \times \overset{3}{\underset{(p,q)}{Observe_0}} \times \overset{4}{Past_0} \times \overset{5}{\underset{Op(x)}{Argument_1}} \times \overset{6}{\underset{+}{AND_2}} \times$$

$$\overset{7}{Number(SG)_1} \times \overset{23}{Lady_0} \times \overset{22}{Old_0} \times \overset{24}{\underset{Op(a)}{Argument_1}} \times \overset{7}{Number(SG)_1} \times \{$$

$$\overset{8}{leader_0}, \overset{9}{music\text{-}conductor\text{-}male_0}, \overset{10}{music\text{-}conductor\text{-}female_0},$$

$$\overset{11}{head\text{-}person_0}, \overset{12}{chief\text{-}person_0}, \overset{13}{executive\text{-}person_0},$$

$$\overset{14}{manager\text{-}male_0}, \overset{15}{manager\text{-}female_0}, \overset{16}{president_0}, \overset{17}{director\text{-}male_0},$$

$$\overset{18}{director\text{-}female_0}, \overset{19}{superintendent_0}, \overset{20}{school\text{-}principal_0},$$

$$\overset{21}{electric\text{-}conductor_0}. \}$$

Figure 23

The items in script represent atomic or molecular NF expressions. The information given in light face print represents instructions for the DSA component; subscripts represent the degree of the normal form expression, which preserves information about the original standard constituency; the numbers above an NF expression refer to the connected sub-graph in the following figure, which has been interpreted by that NF expression.

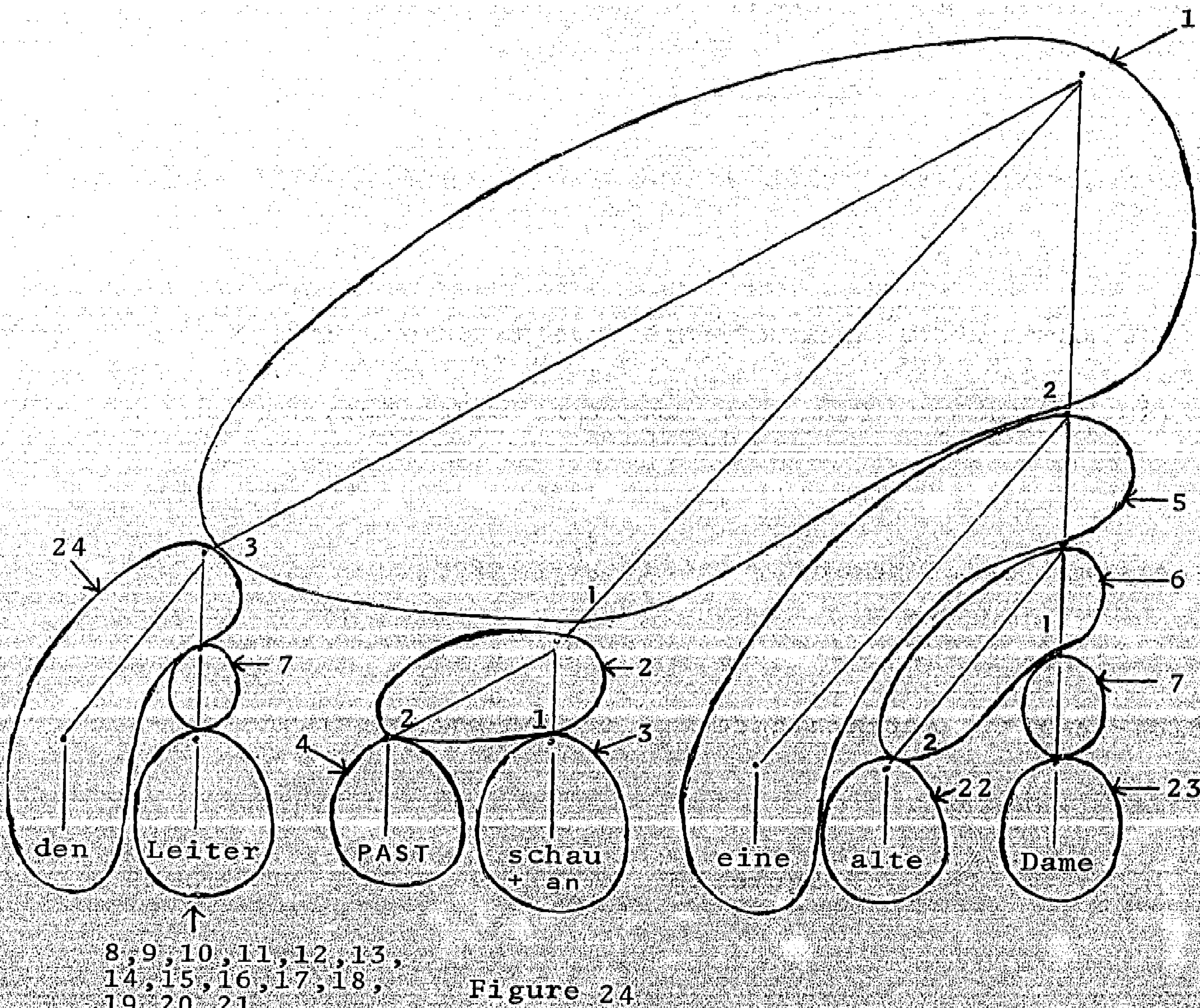
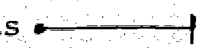


Figure 24

Note that the NF expressions represented by the digits 24 and 5 each interpret a sequence of connected standard trees.

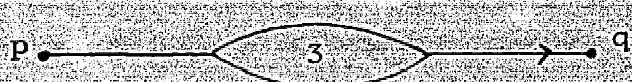
The normal form of a sentence is processed by the DSA-image construction component, which ignores all items which are not of degree 0, or which do not have an operator statement (indicated by light-face print), or which do not have an identifier (indicated by a "+"). For each non-ignored NF expression, the DSA-image component has an instruction:

a) every unary degree 0 symbol is represented as  .

b) n-ary degree 0 symbols are represented as lenses or arrows[38] (binary), triangles (ternary), diamonds (quaternary);

c) other normal form expressions have special instructions which have to be looked up in a set of operation statements. These representations are connected with objects represented by nodes according to wellformedness conditions computable from the degrees of the non-ignored NF expressions.

Let us now discuss the construction of the DSA-image of sentence 39 from its normal form representation. The first instruction, represented in NF expression 3, constructs a lens with the end nodes p and q and calls the lens by the name of the NF expression given in the DSA-image component. (We shall assume that this representation is the numerical representation above the NF expression in Figure 23.)
The first instruction results in the following graph:



Instruction 4 states: Assign the predication "past" to the last predication. NF expression 5 states: Replace one of the variable node names in the existing graph by name x_i (the order of replacement is dependent on the inherent order of the arguments of the predicator. It is reflected by the alphabetical sequence of the letters in the graph.) Expression 6 states: Attach two "and-branches" to the node with which it can be connected through the degree conditions[39]; NF expressions 23 and 22 call these branches LADY and OLD. We have so far obtained the following DSA-image:

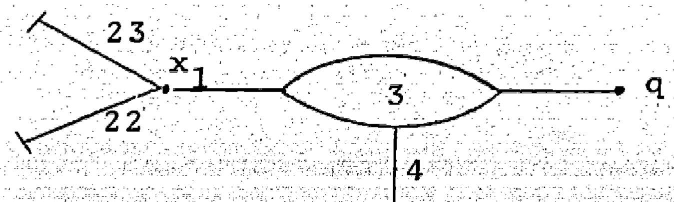


Figure 25

Expression 24 states "change the next variable name to a_i". Thus, q is changed to a_i. To this node, lines representing the NF expressions 8 through 21 are attached by "or" links, resulting in the graph:

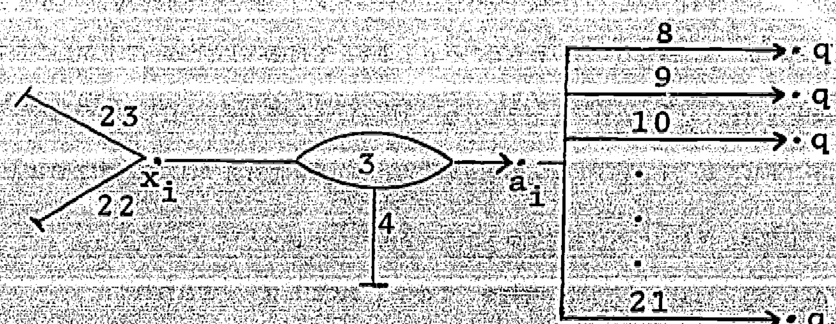


Figure 26

The output of the DSA component is processed by the connection component whose purpose is to replace, if possible, the names of the nodes in the DSA-image by the names of the nodes in the already established SA-image.

The connection component has the following instructions: For each node in the DSA-image which is named x_i it generates a numerical subscript which has not yet occurred in the SA-images, i.e. it assigns a numerical subscript which is larger by 1 than the last that was previously assigned. For each node named by a_i it performs a search through its short-span SA-image and tries to replace the name a_i of that node by the name of one of the nodes in its SA-image, based on the predication associated with the node a_i.[40] (We see that only node x_2 in Figure 18, page 75 , fulfills this condition.)[41] When all nodes in the DSA-image have been assigned their proper names represented in Figure 27, this image is connected with the established connected SA-image, resulting in Figure 28. Duplications of predications upon objects are not repeated.

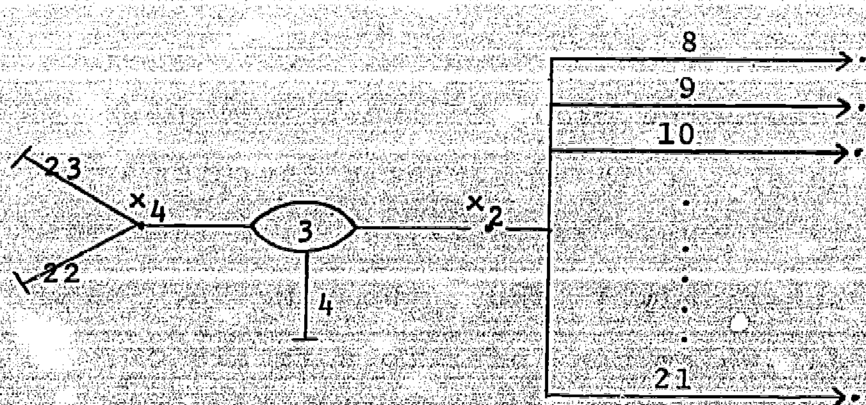


Figure 27

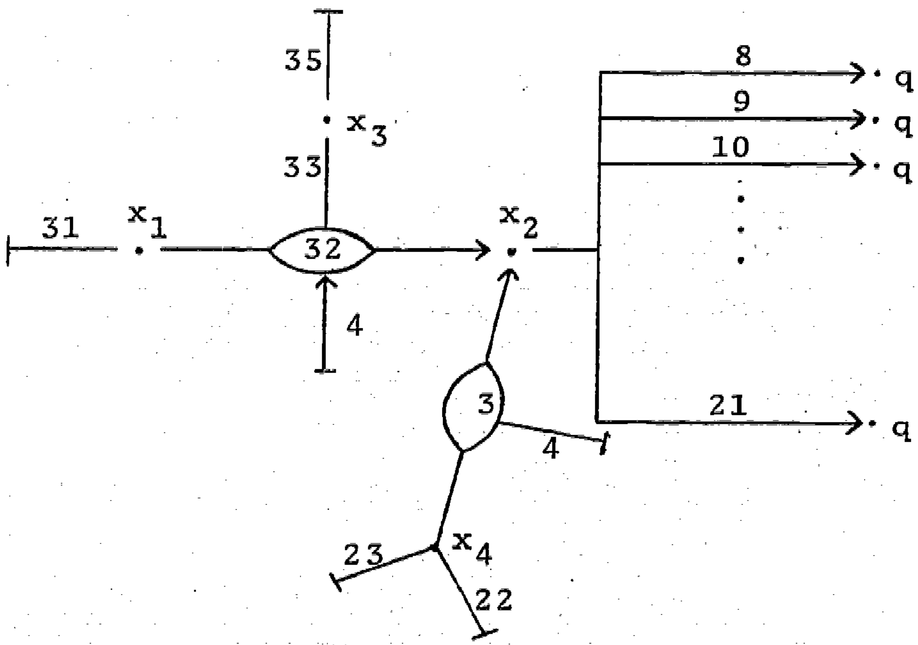


Figure 28

The processing of the sentence *Sie zerbrach ihn* results in the following DSA-image:

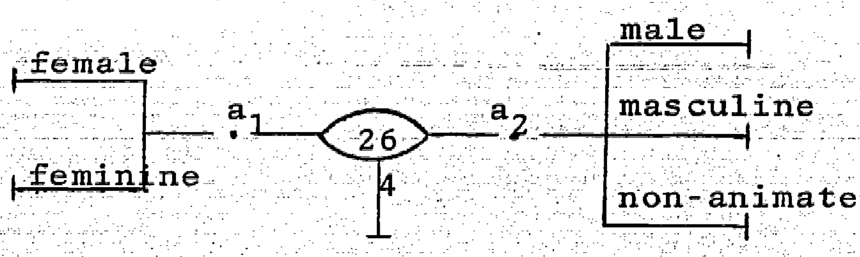


Figure 29

where the expressions above the property lines, connected by "or"-links, state the syntactic or semantic features of a_1 and a_2, respectively. (These are obtained from the pronouns *sie* and *ihn*, respectively.) Thus, the name "a_1" represents the pronoun *sie*, which can refer to a female or feminine object. "a_2" represents the pronoun *ihn*, which can refer to a male object, a non-animate object or an object of gender masculine.

The connection component tries to establish the referent for nodes a_1 and a_2, beginning with its most recent SA-image. a_1 could refer to x_4 or to x_2 since all of its predications meet at least one condition of a_1. a_2, however, can only refer to x_2. Consequently, the SA-image for this sentence results in Figure 30.

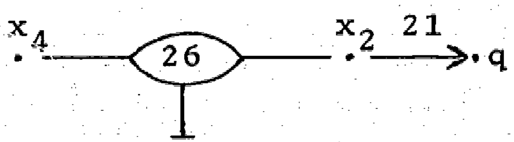


Figure 30

Its connection with the already established SA-image results in Figure 31 in which all the ambiguities represented by the "or" links associated with x_2 have disappeared.

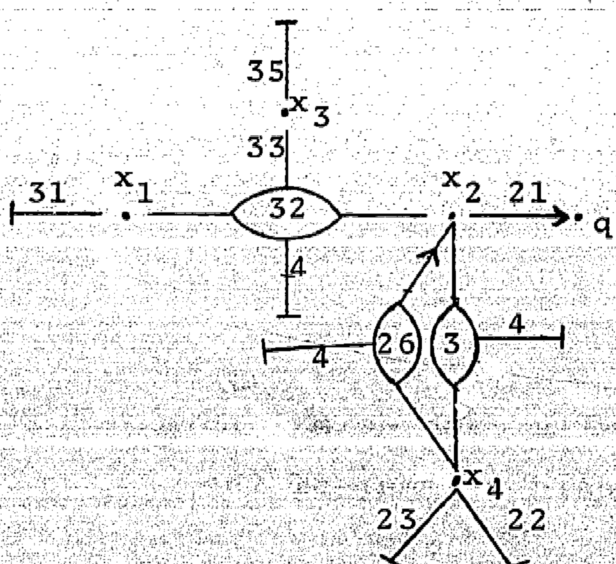


Figure 31

The first and second sentences are disambiguated by comparing their DSA-images against the established SA-image. The disambiguation of DSA-images results in a removal of the ambiguous NF interpretations for the term *Leiter*.

The resulting normal form of sentence 39 is then mapped into the identical normal form of the output language, and the graphs associated with each output NF expression are retrieved. One of these graphs is the NF rule

```
V LOOKAT    R 39
$ A(2)      + OB(at)
$ B(3)
```

where the values "2" and "3" represent accidental features carried over from German. This rule results in the retrieval of the standard rule (a subset of the surface rule):

```
R 39   V V           look
       + PX(0)
       + TY(AN)
       + SS(2)
       + OB(at)
       + TO(PO,AB)
       + SO(3)
```

The standard sub-graphs associated with each normal form expression are analyzed by the standard grammar of the output language, resulting in

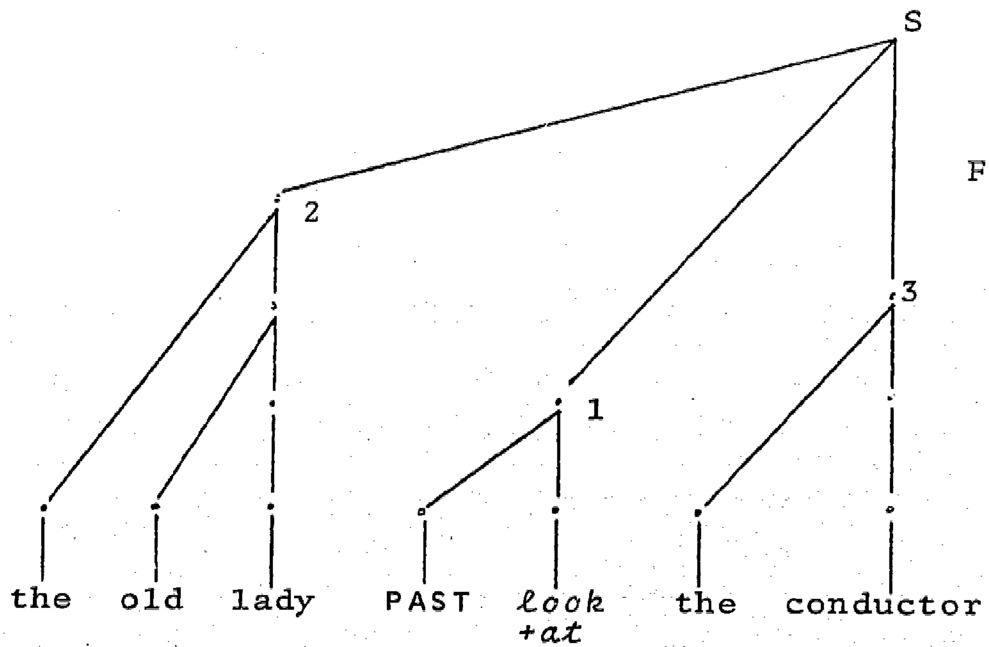


Figure 32

The rearrangement grammar featurizes the dummy symbol PAST and lexicalizes the feature *at*, resulting in the surface string

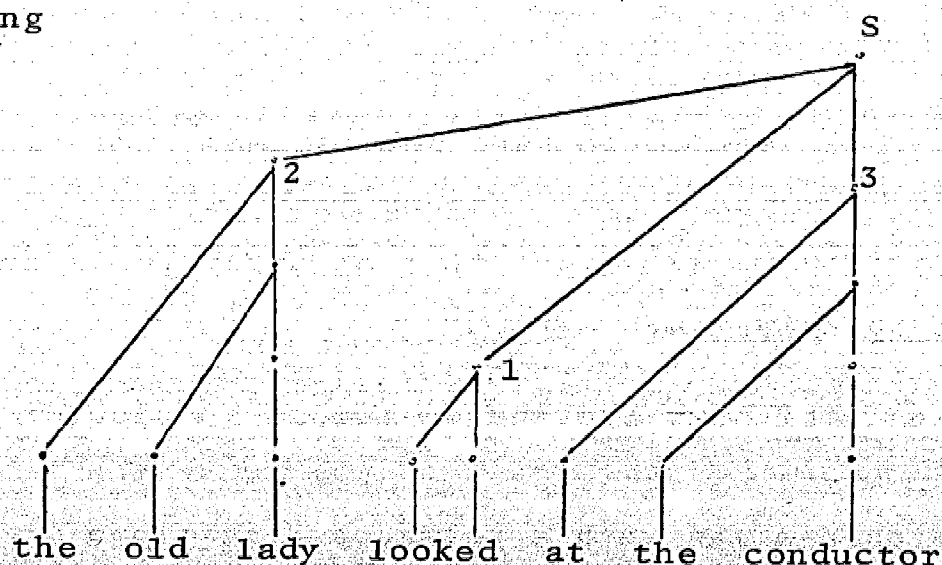


Figure 33

We performed this translation by actually using the memory of the designed LRS information retrieval system. We assume that in translation a short-term memory will be sufficient. DSA-images then only need to be constructed for the immediate environment of an ambiguous sentence. It may even be possible to restrict the construction of such DSA-images to the unary predications of the objects occurring in the environment. This decision, of course, is dependent on the results of research in discourse analysis. (In MT, it is not necessary to establish every referent of an expression as it is for information retrieval; it is only necessary to establish those referents which help to disambiguate a particular sentence.)

That the system has the power to carry input ambiguity across can be observed from the fact that the English terminal *conductor* will be retrieved twice as an equivalent item for *Leiter*, once through *conductor-music-male*, and once through *conductor-inanimate*. It is fairly simple to compute output terminals which have several meanings in common with an input terminal (cf. Lehmann-Stachowitz, 1970). However, this will only be necessary if the context does not provide any disambiguating information.

The construction of ambiguous syntactic structures also has to be performed by means of SA-images. Assume that the sentence "John watched a man with a telescope" is represented by the SA-image:

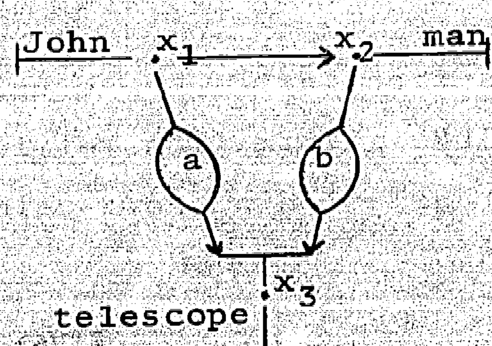


Figure 34[42]

where a stands for "use", b stands for "have". In order to map this ambiguity across, the system would have to be provided with the knowledge that the structure

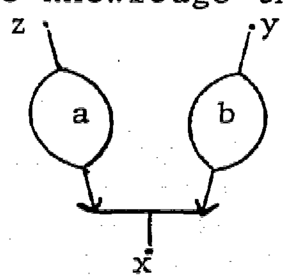


Figure 35

can be mapped as "with x" and the objects naming the nodes have to occur in the surface order z, y, x where x has to follow y directly. Such capabilities are those of a speech production device which are currently not regarded as being necessary for MT.

The capabilities of LRS are based on the following factors:

a) its subscript grammar with the feature-sensitive transformations;

b) its normal form component;

c) its DSA-image and connection component; and most important,

d) its lexicon.

The subscript grammar permits us to express in a rule relations like agreement and government, which correspond to the intuition of a human speaker. We can express grammatical categories as a) lexical categories: noun, verb; b) syntactic categories: NP, predicate; c) generic grammatical categories: number, tense, case; d) specific grammatical categories: singular

plural; present, past; nominative, accusative; etc.[43]
We can also express semantic categories like human, animate abstract, etc.; stylistic categories like colloquial, vulgar, learned; and lexical categories like morpheme and allomorph. The subscript grammar permits us to express in a natural manner such concepts as gender (with the values masculine, feminine, and neuter) instead of representing it as a bundle of unordered binary features as in

[+masculine] [-feminine]	[-masculine] [+feminine]	[-masculine] [-feminine]

where the combination

[+masculine]
[+feminine]

has to be excluded by means of an ad-hoc segment-structure rule.

By means of the subscript grammar rules we can formulate redundancy statements, conflate ambiguous trees into one tree; we can also update the lexicon by adding additional necessary semantic features to it without having to make corresponding changes in the syntactic rules interpreting them.

The transformational component permits the disambiguation of lexical items by means of "jump operations" within a disambiguating syntactic or semo-syntactic environment. It permits us to add stylistic interpretation to syntactic structures if certain conditions stated as features of the constituents are fulfilled.

The normal form component assigns an NF expression to (connected) syntactic subtrees, and to lexical subtrees with a specific set of semantic features within a specific semo-syntactic environment. It is also able to assign a semantic interpretation to verbal phrases and idiomatic expressions with or without internal variable slots and to map these NF expressions into the corresponding NF expressions of the output language without affecting their transformational properties (cf. the following graphs)

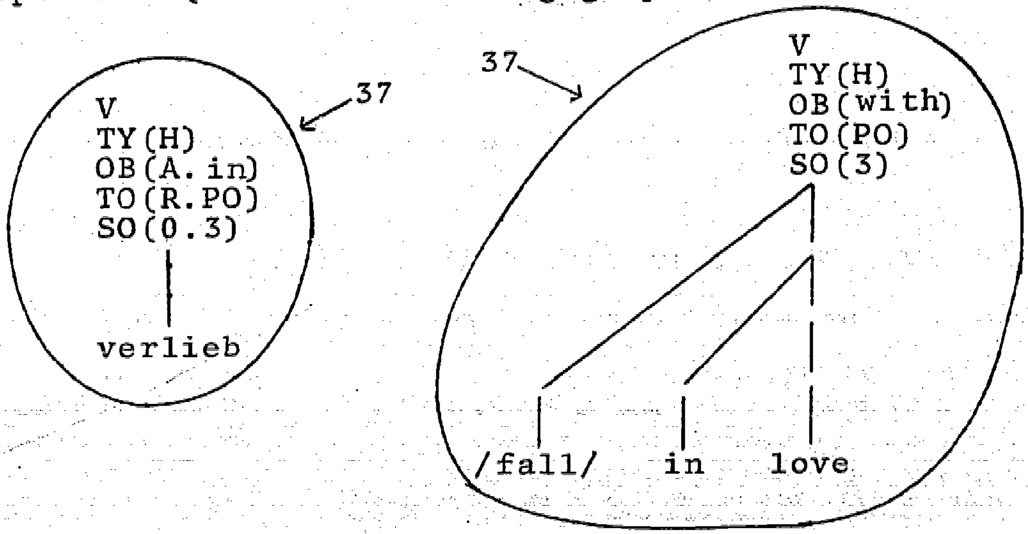


Figure 36

where /fall/ represents the morpheme *fall* (the actual allomorph is generated during the rearrangement stage).

NF rules are currently the only rules of the meaning rule component of LRS; we are planning to extend this component to include definition and inference rules, as for example

```
N KILL     →     N DEAD
$ P              $ Q
$ Q
```

which represents: "If P kills Q, then Q is dead."

The DSA-image component and connection component permit the disambiguation of an ambiguous sentence by means of its co-textual environment.

All the capabilities of the components mentioned would be ineffective if it were not for the lexicon which has to a large extent already been constructed at the Linguistics Research Center. The LRS dictionaries contain stems, inflectional affixes (and, for German, two types of derivational affixes: separable and inseparable prefixes) which are concatenated by means of the surface word grammar rules.

These dictionaries are currently being updated by establishing for each stem

a) its syntactic and semo-syntactic properties,

b) the syntactic and semo-syntactic properties of the environment in which the item may occur with a particular meaning.[44]

Polysemic terms are thus represented as one term. The system of rules

```
R 3    C N            *page            : surface rule
       + TY(H,IN)
         :

       N HOTELBOY    R 3              : NF rules
                     $ TY(H)

       N BOOKPAGE    R 3
                     $ TY(IN)
```

expresses the polysemy of the noun *page*. The transformations of the surface component have the effect that *page* is interpreted as *HOTELBOY* or *BOOKPAGE*, or both, in environments as *The page slept* or *The page tore* or *He touched the page*, respectively.

Lexicographic work at LRC (cf. the appendix, for details) has already resulted in word lists containing

a) 10,000 German verbs and 10,000 English verbs, both classified with respect to their object complements;[45] about 2,000 entries of the latter have been classified with respect to subject and adverbial complements. Similar work on the German verbs is in progress.

b) 33,000 German nouns (letters A through K) with about 70,000 English correspondences; the first 7,000 of these German nouns have been classified according to the scheme shown in the appendix;

c) 6,000 German and English verbal phrases (verb-noun phrase and verb-prepositional phrase combinations), classified as to subject, object, and adverbial complement.

Work on adjectives and adverbs is beginning.

Future additional lexicographic work at the Center will be directed towards the establishment of a minimal set of additional semantic features in order to disambiguate verbs which have particular meanings in particular lexical environments, "distinguishers" in the sense of Katz-Fodor.

In view of such combinations as:

<u>*abhaengen*</u> <u>*von*</u>	<u>*depend*</u> <u>*on*</u>
<u>*abhaeng*</u>*ig* <u>*von*</u>	<u>*depend*</u>*ent* <u>*on*</u>
<u>*Abhaeng*</u>*igkeit* <u>*von*</u>	<u>*depend*</u>*ency* <u>*on*</u>

we also plan to reduce the size of the surface dictionary (projected number for German = 80,000 entries, for English = 100,000 entries) by removing productive derivations and compounds from the dictionaries. This will be performed by adding derivational affixes to the surface dictionary and word formation rules to the surface word grammar. In order to facilitate the design of the necessary word formation rules for German and English, programs are presently being constructed to analyze and display in concordance format the analysis of each of the individual entries in the current surface dictionaries by means of the whole surface dictionary (to which all derivational affixes of the language have been previously added).[46]

The listed components, in particular the complete lexical component, give LRS to a great extent the power of the hypothetical translation device T (pages 46 through 49). LRS can meet the requirements a through g:

a) derivation of semantic reading R for sentence t;

b) mapping of semantic reading R into semantic reading R';

c) derivation of sentence t' from semantic reading R';

d) disambiguation of t in context;

e) access to a meaning rule component;

f) generation of syntactic structure of t' which resembles the syntactic structure of t;

g) generation of t' with a stylistic interpretation corresponding to the stylistic interpretation of t.

Though LRS permits the carrying over of lexical ambiguities (requirement h), we feel that this will not be necessary because of the ability to disambiguate in context.

Requirement i): carrying across of non-ambiguity of t into corresponding non-ambiguity of t', can presently only be obtained by re-analyzing standard string t'. This we do not regard as practical. Carrying over of non-ambiguity could be guaranteed by adding diacritics to t' which simulate the labeled bracketing of t'. However, this may not be very convenient for the reader.

Apart from its applicability to machine translation and information retrieval, we assume that LRS also provides reasonable explanations for a number of not easily explainable linguistic phenomena, as for example the occurrence of the underlined *the*'s in the sequence

41. *One of Rembrandt's pictures was sold yesterday.*
42. *The seller was very happy with the price.*
43. *The buyer is probably an American.*

If we represent the sentence 41 by the graph

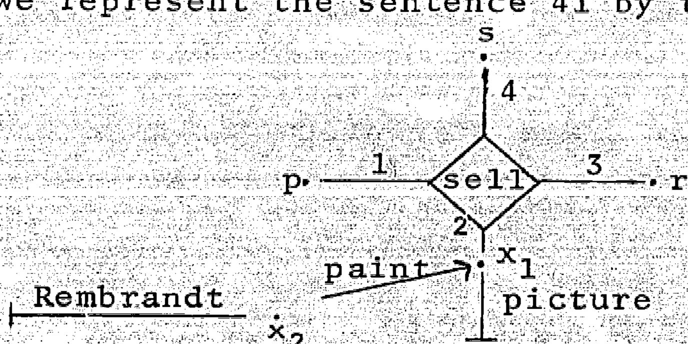


Figure 37

we can explain the occurrence of the definite articles in the sentences 42 and 43 by the fact that the object they refer to (p, r and s) have been implied though not specified in sentence 41.

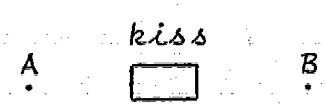
We can also reasonably explain the following "paraphrases" of sentences:

> *A and B kissed. A and B kissed one another. A and B gave kisses to one another. A and B exchanged kisses.*

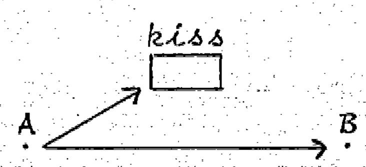
which have complete or partial correspondences in a number of languages such as French, Latin, Serbian, Hebrew, German. Let us represent *A and B kissed* by $A \,.\, \xleftrightarrow{kiss} \,.\, B$ (which can also be read as:

> *A and B kissed one another:* $A \,.\, \underset{kiss}{\overset{kiss}{\rightleftarrows}} \,.\, B$).

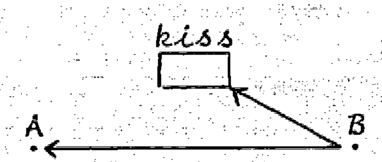
The nominalization of *kiss* results in the following diagram



In order to establish a relation between the three objects, a diagram like



or



is necessary. These graphs permit the interpretation *A gives a kiss to B* or *B gives a kiss to A*.

Since the kiss that B gives to A is identical with the kiss A gives to B, we need to extend the graph to

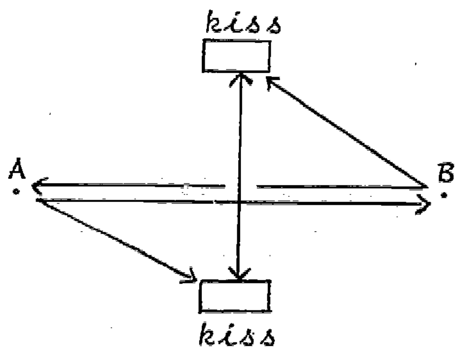


Figure 38

The resulting diagram, as we may observe, is similar to the diagram for "sell" in Figure 7 (page 15), where one of the conditions for the equivalence of the given objects, namely money, has been removed. This exactly describes the actions involved in an exchange of objects thus permitting the interpretations: *A gives a kiss to B* and (simultaneously) *B gives a kiss to A* or *A and B gave kisses to one another* or *A and B exchanged kisses* (or: *a kiss*).

LRS, as we observe, is a complex configuration of components, actually more complex than described in this paper. This complexity, of course, is due to the complexity of the processes occurring during speech recognition and speech production. The question, however, that naturally arises is: How efficiently, i.e. how inexpensively, can mechanical translation be performed with LRS? We will try to answer that question in the next and final chapter.

6. Progress in Hardware Development and the Future of Machine Translation

The criteria according to which the feasibility of machine translation is normally evaluated are: quality, speed, and cost. In this chapter we do not want to deal with the first of these criteria: our demands on the quality of MT output have been stated and the quality of such output can really not be evaluated before the output exists. We also want to ignore speed, since speed is a factor which is normally used in favor of machine translation. As to cost, we want to restrict ourselves to costs arising from computer processing and exclude those costs which might arise through pre-editing and post-editing (though not in LRS, which is conceived as a fully automatic MT system) and key-punching of a text.

Cost of computer time is dependent on mainly two factors: the actual use of central processing time and the use of input-output time. That the central processor can work with immense speed is generally known; it is less known to the non-specialist that input-output operations are by many orders of magnitude slower than the speed of the central processor and that the central processor must stop with its computations for a particular program until the input-output operations for that program are completed.

Machine translation is a process which requires almost constant input-output operations. We can visualize the performance of a computer during machine translation by imagining a human being A who reads a text according to the following conditions: A has available different kinds of information;

a) a dictionary consisting of a number of separate booklets, which contains all paradigmatic, syntactic, and semo-syntactic information pertaining to a word,

b) a grammar which also consists of several separate individual volumes,

c) a dictionary of word definitions or meanings consisting of even more separate volumes than the paradigmatic dictionary,

d) a semantic grammar in several volumes which contains the interpretation rules necessary for the computation of the meaning of a text from the lexical items and syntactic relations.

A has to read the text word by word. He may only continue with the next text word if he has found the word that he is currently looking at in one of the parts of the paradigmatic dictionary. Actually, it must be in that part of the dictionary which he is holding in his hand. If the word occurs in that volume, he may proceed to the next word. If not, he has to put this volume down and pick up another volume and check whether the word occurs in it. By means of an efficient search procedure he repeats this process until he finds the volume which contains the word. He then looks up the word and writes down its part of speech interpretation. Then he proceeds with the next word. To speed up his performance, A keeps the volume which he is currently "processing" in his hand as long as possible because it might be the case that the next word that he reads also occurs in that part. In reality, to decrease the number of volumes of the dictionary, A is not reading whole words but constituents of words.[47] When A has looked up and written down all the paradigmatic information associated with each word constituent, he begins processing the text again, beginning with the first word, this time consulting his grammar books. The procedure is repeated in a similar fashion. Then A starts using his dictionary for semantic analysis, and so on.

The picking up of individual volumes and putting them down again represent the input-output operations of a computer whose central memory is simply not large enough to hold several volumes of a dictionary,or even the whole grammar,since the memory must also hold the programming instructions and the results of the computations.

The advantage of the LRS subscript grammar G is that it represents an abbreviated edition of a multi-volume grammar G'. (Some of the subscript rules represent hundreds, a few even thousands of former context-free rules with simple symbols.) The information in grammar G permits the computer to compute the information contained in Grammar G' and only that information actually needed for the analysis of the particular text sequence currently being processed. And we recall that a computer can compute with extremely high speeds.

In spite of the advantages of the subscript grammar, we observe that the problems pertaining to the recognition of the dictionary items are not alleviated by means of such a grammar since the number of dictionary entries is a given number which cannot be changed. (The conflation of dictionary items, possible with a subscript grammar, still does not change the number of entries.)

Fortunately, a development in computer hardware is in the offing which will have decisive effects on machine translation and other research areas which are forced to deal with large data bases: the holographic memory. (Cf. Peter L. Briggs: "Holographic Memories Could Make Others Obsolete", Part IV of "The Great Memory Debate" in Computerworld, August 26, 1970, page 44.)

"Researchers now working with holographic memories claim that one holographic memory the size of an average office desk will have the capacity of all on-line storage in use in the Western world." and that "The desk-size holographic unit, with several 100 trillion bits of storage, would exceed the capacity of all of the disks, drums, and core memory now in use ..."

(Holographic memory) "will offer users multitrillion character storage at ... prices probably less than one-thousandth of (the current price) for large-capacity disk storage."

The information in such memories can be accessed with the speed of light; "access times below 20 nano-seconds/per character or/word or/whatever (will be) feasible within five years. It is possible that such memories may be sufficiently faster than (the processing speed of) the best central processors, that they can efficiently serve several large CPUs ... or several thousand terminals at once." ... "Users have indicated that they really don't have any idea what impact unlimited memory might have on their DP" (data processing) "applications and system designs, but they all agree that the whole way of using a computer ought to change when the storage of data is no longer a factor, and when the access speeds are as fast as the central processor, itself."[48]

The conclusions for MT are obvious. The speed and, consequently, the cost of machine translation can be considerably reduced because all the dictionaries, syntactic rules, semantic rules, etc., even the processing programs can be stored in a part of a holographic memory. The problems which remain in the production of workable holographic memories, namely to make them erasable, are no real problems for an established MT system since it will be able to operate with a read-only memory. Changes and additions to the grammars which will be necessary because of neologisms that are introduced into a language can always be stored on disk and be read into the central memory before translation is performed.

In our opinion the real importance of such memories lies not so much in the increased speed with which data processing can be performed, but in the completely new methods of processing data and solving problems that such memories will permit.

The various models of human performance that have been constructed in the social sciences: sociology, economy, etc., normally reflect in some way the way we are accustomed to talk about a subject matter. In linguistics we are accustomed to talk about sounds, morphemes, words, syntax, semantics, and even about context and pragmatics. Linguistic models, however abstract, in some way reflect this way of our talking about language. Thus, we have hierarchical phonological, syntactic and semantic "levels" in some models, and phonological, syntactic and semantic "components" in others. The effect of each component or level is twofold:

a) it assigns to the data an interpretation according to its instructions, and

b) it eliminates those interpretations which were well-formed according to the instructions of previous components but which are not well-formed according to its own.

Holographic memories may change our way of constructing models which is based on 19th century investigations and considerations (John Stuart Mill); according to those we assume one, or a few variables for the analysis of a complex phenomenon and keep all other factors invariant. The fact that we speak of several levels or components of "language", like phonetics, phonemics, morphology, lexicon, syntax, semantics, pragmatics, etc., has not been imposed on us because of the nature of language but because it is easier for us to treat individual phenomena by ignoring certain others, especially if those others are very complex and really not quite unders* od. With the capabilities of computers expanded in such a way, we can finally begin to re-introduce the total approaches (ganzheitliche Methoden) by mentioning the conditions for all the variables that we know.

Now, what does that mean for machine translation? Since the projected access time of such memories, about 20 nano-seconds, is shorter than the time needed for a minimal basic computer operation, it means that such a memory can be read by several computers "simultaneously".

We could thus theoretically construct a machine translation system in which one computer performs dictionary analysis; one, word analysis; one, syntactic analysis, etc.: one computer for each component of the system. The intermediate output of each computer could immediately become input for the next "higher" computer, which again would give its output to the one "above" it, etc. At the same time, each computer could return the results of its own computation to the computer working directly "below" it in the hierarchy. Of course, we are not seriously proposing a system consisting of several computers to perform machine translation, but it is generally known that we can simulate on one computer the performance and capabilities of several computers. We can thus write programs which no longer analyze the data in a hierarchical "horizontal" fashion but in a hierarchical "vertical" fashion, which is the way the human brain operates during the understanding and production of sentences. Nobody would seriously assume that semantic interpretation is performed over the output of some type of complete syntactic analysis represented by a tree with the root S. If that were so, strings of words like those underlined in the following sequence:

George said: After I had ... As usual he could not finish his sentence because Mary interrupted him.

could not be understood. And that we really understand sentences sequentially is clear from many observations, like the following: During a conversation between two people A and B, B explains some matter to A and hesitates, grasping for some word that eludes him; A provides the missing words and continues the sentence for B.

It is perfectly possible that mechanical translation performed with such "vertical" model will approximate "simultaneous translation"; that, while the system is still processing source language text on the input side, it is already producing target language translations on the output side.

I may be overly optimistic when I say that eventually the cost of machine translation may depend on two factors:

a) the speed with which the source material to be translated can be read into the computer, and

b) the speed with which the translation can be printed out by computer.

Holographic memories will provide us with the technical capabilities to construct models which are to a high degree representations of the reality which surrounds and which affects us. They will provide us with the means to test our hypotheses, and, if necessary, to modify or even reject them. It is our task to be prepared for these possibilities by performing the necessary research, by collecting the necessary data. This task will not be easy; it will also be expensive; but eventually it will be rewarding, not just as an "intellectual exercise" but as a means to understand ourselves, to become an integrated part of a cybernetic society.

FOOTNOTES

1 There is no need to deal in this paper with certain claims according to which these disciplines are actually sciences.

2 Cf. I.M. Bocheński: Die zeitgenoessischen Denkmethoden, Dalp-Taschenbuecher, Bd. 304; Lehnen Verlag, Muenchen, 1959 (2).

3 "Die schematische Durchfuehrung eines vorgegebenen allgemeinen Verfahrens bietet (nach einigen Proben) offenbar einem Mathematiker kein besonderes Interesse. Wir koennen also die bemerkenswerte Tatsache feststellen, dass ein schoepferischer Mathematiker durch die spezifisch mathematische Leistung der Entwicklung einer allgemeinen Methode den durch diese Methode beherrschten Bereich gewissermassen mathematisch entwertet." Hans Hermes: Aufzaehlbarkeit, Entscheidbarkeit, Berechenbarkeit, Springer-Verlag, Berlin, 1961. The translation of this passage provided in the English translation of this book somehow does not reflect the author's statement.

4 Charles J. Fillmore: The Case for Case in: Universals in Linguistic Theory (eds.: Emmon Bach and Robert T. Harms), Holt, Rinehart and Winston, Inc., New York, 1968.

5 Cf. John Lyons: Introduction to Theoretical Linguistics, Cambridge University Press, Cambridge, 1968.

6 Personal communication with Reed Bates and Emmon Bach.

7 This principle is most often used in dictionary definitions where the meaning of the term defined is a common subset of the meaning of the words linked by "or" in the definiens.

8 Cf. Peters, P. Stanley and Robert W. Ritchie: "A Note on the Universal Base Hypothesis". Journal of Linguistics, Vol. 5, 1969 and "On the Generative Power of Transformational Grammars", to appear in Information Sciences. It is surprising how little impact their results have had on the linguistic community, so far. For the only exception - to my knowledge - cf. Emmon Bach: "Syntax since Aspects" (paper given at the Georgetown Roundtable Conference, March 1971).

9 Cf. the publications in the series: *Transformation and Discourse Analysis Papers*, University of Pennsylvania.

10 Performed in spring 1967 and described in Lehmann-Stachowitz, 1970 and Stachowitz, 1971.

11 Clearly, commands, requests and questions might be reformulated as statements, as for example "Someone orders that S", "Someone requests that S", "Some requests a statement S(x)" such that the variable x is replaced by a constant, where x represents the questioned element in a sentence, as in "Where are you going?" or by an affirmation, negation or modification of certainty or uncertainty as in "Will he come?" "Yes". "No". "Maybe". "Possibly". "Maybe not". etc. We do not have such a reformulation in mind. We argue in the next paragraph of the text that a sentence evokes an image of something. This "something" we want to call a state of affairs.

12 j>i stands for: The point of time represented by j is later in time than the point of time represented by i.

13 Lines which extend from a node represent predications joined by logical and.

14 Clearly, this is a simplified version of the meaning of "sell"(A,B,C,D). We ask the reader to accept our definition.

15 Line 7, representing the property "physical object", may be omitted from Figure 10 if we assume a meaning rule component which contains the meaning rule "For all x, if x is a car, then x is a physical object".

16 The value for n will have to be determined experimentally.

17 If the equivalence relation between "sell" and "pay", and "sell" and "pass" is not regarded as appropriate, the sign for equivalence may be replaced by the sign for inference.

18 Ternary relations are represented by a triangle, binary relations by a cross-section of a lens: 

19 Requirements c and d are possibly too strict to represent actual speech production.

20 We are ignoring in this representation the various time relations as expressed in Figure 7.

21 A leaflet handed out by one of the University of Texas at Austin student groups in 1965 contained as the only statement: "Students should have a voice in decisions that effect them". We assume that the system as well as the reader of this footnote automatically interprets "effect" as "affect"; the system would do this because it becomes "aware" of the absurdity of the statement as it stands, in contrast to the reader, who, normally, only becomes aware of it when the printing error is pointed out. (I owe this example to Professor Norman Martin of the University of Texas at Austin Philosophy Department.)

22 To be exact, the terms "referent", etc. only refer to the objects which are "involved" in states of affairs.

23 We are using the term "synonymous" as a substitute for the term "equi-iconic", which to define would be a further digression; for this term cf. Lehmann-Stachowitz, 1971b.

24 We exclude from this judgment the works of J.A. McConochie Simplicity and Complexity in Scientific Writing: A Computer Study of Engineering Textbooks. Ed.D. dissertation, Columbia University, 1969, and M.L. Gopnik, Linguistic Structures in Scientific Text, Ph.D. dissertation, University of Pennsylvania, 1969; both authors have arrived at results which seem to indicate that the language used in scientific texts is indeed a simpler subset of the regular language.

25 A stylistically correct translation would be "He goofed".

26 The actual percentage is lower since we considered only eight verbs of 15 verbs occurring in that passage. The text, though originally selected at random, is, of course, too short to count as a representative sample.

27 Wildhagen, Karl and Will Héraucourt, English-German German-English Dictionary, Vol. II German-English, Brandstetter Verlag, Wiesbaden, 1953, and Heinz Messinger: Langenscheidts Handwoerterbuch Deutsch-Englisch, Langenscheidt KG, Berlin, 1960 (2).

28 Such an assumption would, of course, mean that there are certain human beings which have learned and can express certain things in their language which no speaker of another language can learn and express. We regard this as impossible.

29 Gruber, Jeffrey S., Studies in Lexical Relations, Ph.D. dissertation, M.I.T., Cambridge, September, 1965.

30 For a comprehensive description, cf. S.R. Petrick, "Syntactic Analysis Requirements of Machine Translation", IBM T.J. Watson Research Center, Yorktown Heights, 1971.

31 Automatic Language Processing Advisory Committee 1966. Language and Machines: Computers in Translation and Linguistics. Publication 1416. Washington, D.C., National Academy of Sciences, National Research Council.

32 Petrick (op. cit.)

33 Immanuel Kants Logik, ein Handbuch zu Vorlesungen in: Immanuel Kant - Werke in zehn Baenden (herausgegeben von Wilhelm Weischedel), Band 5, Wissenschaftliche Buchgesellschaft, Darmstadt, 1968 (pocket book edition of the Kant-Studienausgabe).

34 Catford, K.C., A Linguistic Theory of Translation -- An Essay in Applied Linguistics, London, Oxford University Press, 1965, published as volume 8 in the series Language and Language Learning, R. Mackin and P.D. Strevens (eds.) and Louis Hjelmslev, Prolegomena to a Theory of Language, Baltimore, 1953.

35 This is necessary to insure the eventual well-formedness of the standard string. If more than one string should result, those which most closely correspond in their accidental features to those of the input sentence t can be selected.

36 We have taken these examples from: Langenscheidt's German-English dictionary, cf. footnote 27.

37 The comma has a stronger binding power than the period.

38 We use the arrow to refer to a binary relation which is nominalized.

39 An "and expression" attaches two lines to a node if it is not in the domain of another "and expression"; one branch, if it is.

40 The terms "a" and "the" have really several operation statements associated with them, interpreting such sentences as "A whale is a mammal", "The whale is a mammal", "The United States is a country", and "Whenever John rides a bus, he starts a fight with the conductor".

41 The NF expressions contain the semantic features of the interpreted terminals of the language, which permits the disambiguation of the predications upon x_2.

42 We treat proper names as predications for two reasons: They may refer to more than one object; certain semantic features, like human, male, female, are normally associated with proper names, even size, as e.g. "Haenschen" (little John). In our system, the "proper names" of objects are represented by a subscript of x.

43 Hockett, Charles F., A Course in Modern Linguistics, The MacMillan Company, New York, 1960.

44 Such information includes semantic markers, distinguishers in the Katz-Fodor sense, area of provenience information, and stylistic information.

45 The list of English verbs - taken from Hornby, A.S., E.V. Gatenby and H. Wakefield, The Advanced Learner's Dictionary of Current English, Second Edition, Oxford University Press, London, 1963 - will appear as an appendix to Lehmann-Stachowitz, 1971b, the list of German verbs in Lehmann-Stachowitz, 1971a. The lists are alphabetically arranged according to the following criteria:

a) verbs which are both transitive and intransitive,
b) verbs which are only transitive, and
c) verbs which are only intransitive.

Each list is subdivided into two parts: one with one-word entries, the other with entries consisting of more than one word. The lists of English verbs which take prepositional objects, sorted alphabetically according to various criteria, has appeared as an appendix to Lehmann-Stachowitz 1970, vol. II.

46 The results will be published as derivational dictionaries of German and English, sorted according to affixes and stems.

47 This look-up procedure is actually more efficient than generating a glossary of the text and analyzing each word only once.

48 I would like to thank Bary Gold for calling my attention to this article and for discussing some of the technicalities and my conclusions with me.

APPENDIX

Lexicographic Work at the Linguistics Research Center

Lexicographic work at the Center is performed in five stages:

a) the copying of lexical material from dictionaries, such as Wildhagen, cf. footnote 27, and Hornby, cf. footnote 45. Information pertaining to distinguishers and area of provenience is copied as given in the dictionaries;

b) the addition of syntactic and semo-syntactic features to the obtained items according to the classification scheme given in the following pages;

c) the establishment of equivalence relations or inference relations between syntactic and/or semo-syntactic features of all entries or large subsets of entries. (Features that can be predicted from the occurrence of other features need not occur in the dictionary; they can be introduced by means of redundancy rules during actual analysis);

d) mechanical conversion of the established lists to the LRS dictionary format.

e) conflation with the current LRS dictionaries which contain for each item a subscript pertaining to paradigmatic information and, in the cases of allomorphs, a subscript with the information on how to generate the lemma. German nouns contain gender information; all adjectives contain information about their attributive and/or predicative use.

Stages a and b represent the descriptive phase; stage c, the interpretative phase. Lexicographic work on German and English adjectives, adverbs and nouns is in stage a, work on verbs and a subset of nouns in stage b. During stage c, we plan to introduce additional semantic features required because of the distinguishers associated with some lexical items. (Area of provenience information is handled as one of the accidental features of a lexical entry).

The following pages are a copy of the coding instructions for the LRC lexicographers. Note that some semo-syntactic features occur - to facilitate encoding - as syntactic features, cf. the subscript RL under nouns. During the conversion to LRS format, the features will receive their "correct" interpretation.

VERB FEATURES

TY (VT, VR, VI, VTC, NP, NG*$_E$)

TS (HU, AL, PL, IN, AB, PO, AN, BP, MS, CN, CO, NM, UN, QU, MA, E, P)

FS (NP, IT, TH, MI, FT$_E$, GR$_E$, ICL, IMI$_E$, II*$_G$)

DS$_G$ (G, D, A)

OB (G$_G$, D$_G$, A$_G$, O$_E$, all PREP's, TH, CL, MI, FT$_E$, GR$_E$, ICL, IMI$_E$, PAPL, II$_G$, BC, CM, NC, NA, AC, I)

TO (HU, AL, PL, IN, AB, PO, AN, BP, MS, CN, CO, NM, UN, QU, MA, E, P, R, RCC, IT)

RA (TIM, PNC, EXT, SIM, PRI, POST, LOC, DIR, ORN, MAN, MOD, CAUS, MSR, DEG, FRQ, PRB)

OA (DOR)**

Subscript Definitions:

TY = type of verb
TS = type of subject; always code one of the underlined values for TS and TO; code values without underline only if subject or object is restricted to that value
FS = form of subject
DS = deep subject; mark only if English translation is nominative, e.g. *es friert mich*; do not mark *es gehoert mir*
OB = form of object; for 2 objects with +, the order is: O + PREP, O + CLS; PREP + PREP reverse order given in dictionary. English: Only one object: NP or refl. is not marked. Adjust G order to E order
TO = type of object; code TO values even for object clauses and phrases
RA = requires adverb; e.g. *put* RA(DIR). He put the book on the table, but *He put the book.
OA = optional adverb

Value Definitions:

TY
VT = takes at least one object which is not a reflexive pronoun
VR = takes at least one object which must be a reflexive pronoun

TY	VT,VR = takes at least two objects, one which is reflexive and one which is not a reflexive pronoun VI = intransitive VTC = takes a cognate object only; we define cognate object as the true cognate and all nouns subsumed under that term, e.g. *einen Tanz (Walzer, Regentanz) tanzen* NP = no passive NG = no progressive
TS	HU, AL, etc. as defined for noun features E = entia (any noun, PO or AB) P = plural
FS	NP = noun phrase; code only if another FS value is present IT = it, es; no TS information is required TH = that-clause MI = marked infinitive FT = for-to construction GR = gerund ICL = interrogative clause IMI = interrogative pronoun + marked infinitive II = interrogative pronoun + infinitive
DS	G = genitive D = dative A = accusative
OB	O = NP object Th, MI, etc. as defined above for FS CL = main clause PAPL = past participle BC = takes *be* + NP or ADJ (*think*) CM = takes optional *be* + NP or ADJ NA = takes NP or ADJ complement without *be* NC = takes NP complement without *be* (*elect*) AC = takes adjective complement without *be* I = infinitive
TO	HU, AL, etc. as defined for noun features E = entia (any noun, PO or AB) P = plural R = reflexive RCC = reciprocal (*aneinander geraten*)

RA

TIM = time
PNC = punctual
EXT = extensional
SIM = simultaneous with point of reference
PRI = prior to point of reference
POST = later than point of reference
LOC = location
DIR = direction to
ORN = direction from
MAN = manner
MOD = modality
CAUS = causality
MSR = measure
DEG = degree
FRQ = frequency
PRB = degree of certainty

OA

DOR = direction or origin, i.e. adverb of directionality

Case ambiguity in German prepositions: 1 = acc., 2 = dat.
Example: AN1, AN2

* Subscript E: relevant for English verbs only
Subscript G: relevant for German verbs only

For the descriptors TS and TO, one of the underlined features must be coded for each verb; values without underline can be optionally added.

NOUN FEATURES

TY (<u>HU</u>, <u>AL</u>, <u>PL</u>, <u>IN</u>, <u>AB</u>, <u>AN</u>, <u>PO</u>, MA, BP, MS, CN, CO, NM, <u>UN</u>, QU)
OB (all prepositions)
TO (HU, AL, etc.)
TA (ZU, CL, TH, DIR)
SX (MA, FE)
RL (WO; WOHIN; WARUM; OB; WIE; ALS)
DF (VT, VI, A)
FM (A)

<u>Subscript Definitions</u>:

TY = type of noun
OB = object
TO = type of object
TA = takes attribute
SX = sex
RL = relative pronoun
DF = derived from
FM = form

<u>Value Definitions</u>:

TY

HU = human
AL = animal
PL = plant
IN = inanimate
AB = abstract
AN = animate
PO = physical object
MA = machine which can perform human activities
BP = body part
MS = mass (homogenous, occurs without article in sg: milk, sand)
CN = count
CO = collective (components can be counted; can be used with *disperse* (*group, herd, government*))
NM = proper name
UN = unit (ADV/QU + ____; e.g. Meter, Jahr)
QU = quantity (___ + (of) NP; e.g. group, glass, half, dozen, %)

In this set, one of the underlined values must be coded for each noun; values without underline are optionally added as appropriate.

TA
- ZU = zu-infinitive
- CL = main clause
- TH = that-clause
- DIR = direction (e.g. Flucht nach Italien, zu den Indern)

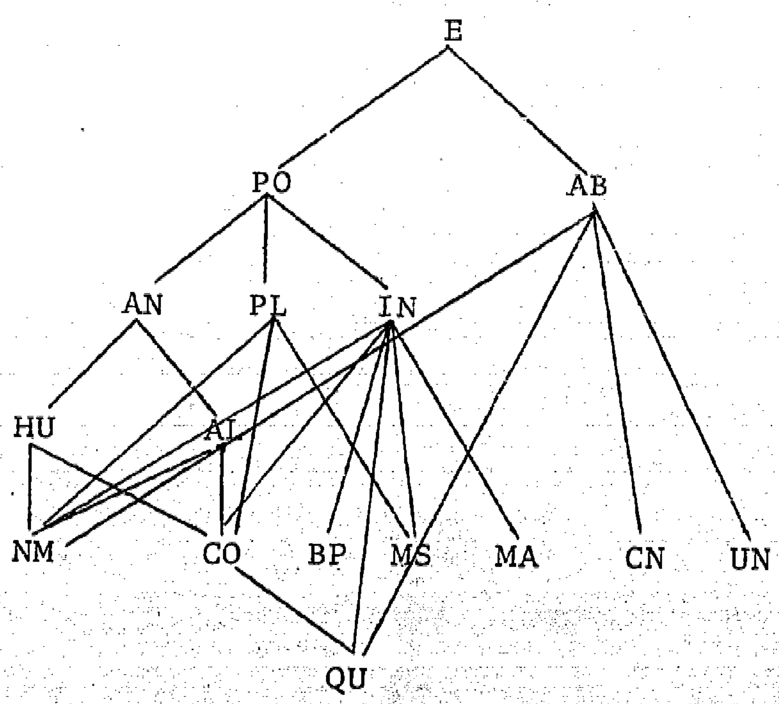
SX
- MA = male
- FE = female

DF
- VT = transitive verb
- VI = intransitive verb
- A = adjective

FM
- A = adjective (e.g. "Abtruennige(r) is coded as a noun: ABTRUENNIG TY(HU) FM(A))

Compounds:

BAUM + WOLL + FABRIKANT



ADJECTIVE FEATURES

MD (HU, AL, PL, IN, AB, PO, AN, E, TH, PLU)

FM (PRPL, PAPL)

TY (MSR, TM)

RA (TIM, PNC, DUR, PLC, LOC, DIR, ORN, MAN)

OB (G, D, A, PREP's)

TO (HU, AL, PL, IN, AB, PO, AN, E)

Subscript Definitions:

MD = the adjective modifies nouns of the specified type
FM = the adjective has the form of a participle
TY = type of adjective
RA = the adjective requires an adverb (e.g. wohnhaft)
OB = object
TO = type of object

Value Definitions:

MD
- HU, AL, etc. as defined for noun
- TH = that-clause
- PLU = plural noun or collective or mass noun

FM
- PRPL = present participle
- PAPL = past participle

TY
- MSR = measurable (wide, old; e.g. five years old, five men strong)
- TM = may undergo tough movement (hard, easy)

RA TIM, PNC, etc. as defined for adverbs

OB
- G = genitive
- D = dative
- A = accusative
- AN1 = an with accusative
- AN2 = an with dative
- other government-ambiguous prepositions are coded analogously

TENTATIVE ADVERB FEATURES

TY (TIM, PNC, EXT, SIM, PRI, POST, LOC, DIR, ORN, MAN, MOD, CAUS, MSR, DEG, FRQ, PRB)

MD (A, AV, V, N, S)

Subscript Definitions:

TY = type of adverb
MD = modifies

Value Definitions:

TY

TIM = time
PNC = punctual
EXT = extensional
SIM = simultaneous with point of reference
PRI = prior to point of reference
POST = later than point of reference
LOC = location
DIR = direction to
ORN = direction from
MSN = manner
MOD = modality
CAUS = causality
MSR = measure
DEG = degree
FRQ = frequency
PRB = degree of certainty

In this set, one of the underlined values must be coded for each adverb; values without underline are optionally added.

MD

A = Adjective
AV = Adverb
V = Verb
N = Noun
S = Sentence

Bibliography

Carnap, Rudolf. 1928. Der logische Aufbau der Welt. Weltkreis-Verlag, Berlin.

Carnap, Rudolf. 1934. Logische Syntax der Sprache. Julius Springer Verlag, Wien.

Chomsky, Noam. 1965. Aspects of the Theory of Syntax. M.I.T. Press, Cambridge, Massachusetts.

Frege, Gottlob. 1879. Begriffsschrift in: From Frege to Gödel (ed. Jean van Heijenoort). Harvard University Press, Cambridge, Massachusetts, 1967.

Harris, Zellig S. 1957. "Co-occurrence and Transformation in Linguistic Structure", Language, Vol. 33, No. 3. Also in: The Structure of Language (eds. J.A. Fodor and J.J. Katz), Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1964.

Harris, Zellig S. 1962. String Analysis of Sentence Structure, Mouton & Co., The Hague.

Harris, Zellig S. 1965. "Transformational Theory". Language, Vol. 41, No. 3.

Harris, Zellig S. 1968. Mathematical Structures of Language, John Wiley & Sons, New York.

Kamlah, Wilhelm and Paul Lorenzen. 1967. Logische Propädeutik oder Vorschule des vernünftigen Redens. Bibliographisches Institut, Mannheim.

Kaplan, Abraham. 1964. The Conduct of Inquiry. Methodology for Behavioral Science. Chandler Publishing Company, San Francisco.

Katz, Jerrold J. and Jerry A. Fodor. 1963. "The Structure of a Semantic Theory". Language, Vol. 39, No. 2. Also in: The Structure of Language. Readings in the Philosophy of Language. (eds. Jerry A. Fodor and Jerrold J. Katz), Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1964.

Katz, Jerrold J. and P.M. Postal. 1964. An Integrated Theory of Linguistic Description, M.I.T. Press, Cambridge, Massachusetts.

Katz, Jerrold J. 1966. The Philosophy of Language, Harper & Row, New York.

Lehmann, Winfred P. and Rolf Stachowitz. 1970. Research in German-English Machine Translation on Syntactic Level, Vol. II. Linguistics Research Center, The University of Texas at Austin, Austin, Texas.

Lehmann, Winfred P. and Rolf Stachowitz. 1971a. Development of German-English Machine Translation System. Linguistics Research Center, The University of Texas at Austin, Austin, Texas.

Lehmann, Winfred P. and Rolf Stachowitz. 1971b. Normalization of Natural Language for Information Retrieval. Linguistics Research Center, The University of Texas at Austin, Austin, Texas.

McCawley, James D. 1968a. "The Role of Semantics in a Grammar", in: Universals in Linguistic Theory, (eds. Emmon Bach and Robert T. Harms), Holt, Rinehart and Winston, New York.

McCawley, James D. 1968b. "Concerning the Base Component of a Transformational Grammar". Foundations of Language, Vol. 3, No. 3.

Morris, Charles. 1946. Signs, Language and Behavior, New York, Prentice-Hall, Inc.

Petrick, Stanley Roy. 1965. A Recognition Procedure for Transformational Grammars, Ph.D. dissertation, M.I.T.

Reichenbach, Hans. 1947. Elements of Symbolic Logic, Collier-MacMillan Ltd., London. (First Free Press Paperback edition 1966).

Stachowitz, Rolf. 1970a. "The Construction of a Computerized Dictionary". Paper presented at the Modern Language Association Lexicographical Conference, Columbus, Ohio.

Stachowitz, Rolf. 1970b. "A Model for the Recognition and Production of Synonymous Expressions With Different Deep Structures". Paper presented at the Conference on Linguistics, University of Iowa, Iowa City.

Stachowitz, Rolf. 1971. "Lexical Features in Translation and Paraphrasing: An Experiment". Linguistics Research Center, The University of Texas at Austin, Austin, Texas.

Tesnière, Lucien. 1966. Eléments de Syntaxe Structurale. Librairie C. Klincksieck, Paris. (deuxième édition revue et corrigée).

Weinreich, Uriel. 1966. "Explorations in Semantic Theory". Current Trends in Linguistics 3. Theoretical Foundations. (ed. Thomas A. Sebeok), Mouton & Co., The Hague.

The Current Status of Computer Hardware and Software as it Affects the Development of High Quality Machine Translation

by

D. Walker

The Mitre Corporation

532a

The developments in computer hardware and software over the past ten years have gone a substantial way toward satisfying the needs specified in the early '60's as prerequisites for effective machine translation programs. In particular, the storage capacities and processing speeds of current computers far exceed some of the stipulated requirements established during that period. Increases in sophistication of programming systems have paralleled hardware developments as evidenced in operating systems like OS for the IBM 360 and 370 series and Tenex for the PDP-10, to name only two. Compiler technology also has advanced markedly during the period, particularly as elaborations of the syntax-directed techniques introduced about ten years ago. Programming languages as well have increased in breadth, flexibility, and power, so that, although assembly language coding certainly still would reduce run-time, it no longer is a cost-effective alternative. As a result it seems reasonable to say that hardware and software considerations no longer constitute major obstacles to machine translation, at least according to strategies that are currently being pursued.

In spite of the conciseness--and, I believe, the essential correctness--of the foregoing summary statement, two observations need to be made before considering the implications I believe can be drawn from it. There are no systems for machine translation that I am aware of which use algorithms designed specifically to take advantage of recent computer capabilities. That is, the strategies currently pursued are those established in the early

1960's. While hardware and software may not be obstacles, it is not clear that they have been used to full advantage. However, looking at the other side of the issue, it also is not clear that new approaches, particularly those motivated by the recent concerns with semantic processing, might not result in specifications for machine architecture or programming that cannot be met by existing equipment and procedures.

Whatever importance is assigned to these observations, it is clear in any case that the problems of mechanical translation at this stage are primarily of two kinds, linguistic and algorithmic. That is, the responsibility for establishing hardware and software requirements depends on the design specifications for a mechanical translation system. And these specifications entail a knowledge of the grammars of the language involved, a strategy for analyzing them, and a procedure for relating the analyses. Until we can resolve these matters satisfactorily, any prescriptions for hardware and software are purely speculative.

In spite of these uncertainties, one class of computer capabilities should be stressed in this context both because of its potential use in the process of mechanical translation and because of the role it may play in grammar development and in the formulation of algorithms for linguistic analysis. I am referring to interactive capabilities that allow for on-line access to the computer. Although it is only recently that such man-machine systems have become cost-effective, it is still worth asking why machine-aided approaches to machine translation were not proposed and pursued in earlier years. Logically, they would seem easier to implement than would fully automatic approaches. Again, I suspect that the problem here as before is the lack of understanding about grammar, linguistic analysis, and translation algorithms. However, there has been a substantial amount of work now with grammar testers and with systems for handling personal files, work that should be extended into the mechanical translation arena.

Equivalents and Explanations in Bilingual Dictionaries

by

L. Zgusta

Linguistics Research Center
The University of Texas at Austin
and
Department of Linguistics
The University of Illinois

536

EQUIVALENTS AND EXPLANATIONS IN BILINGUAL DICTIONARIES.[1]

L. Zgusta

The task of the bilingual lexicographer is to find such lexical units in the target language as are equivalent to the lexical units of the source language, and to coordinate them. We call "lexical equivalent" a lexical unit of the target language which has the same lexical meaning as the respective lexical unit of the source language. The definitional requirement is that the identity should be absolute: the equivalent should have the same polysemy, the same stylistic value, etc. But such absolute equivalents are rather rare. In the majority of instances, the lexical meaning of the respective lexical unit of the target language corresponds only partly to that of its counterpart in the source language. If we wish to be very precise, we therefore speak about partial equivalents, but normally, we use the term "equivalent" knowing that the majority are partial.

Before starting the search for equivalents, we must compare the structures of the two languages in order to decide which grammatical categories will be considered reciprocal. This is easy in languages which have similar categories of lexical units, or, traditionally, similar parts of speech, e.g. there is no problem in deciding that the French equivalent of an English noun will first be sought among French nouns. But one must not stick to this principle too strictly. For instance, German *Handarbeit* (subst.) has a good equivalent in English *hand-work* (subst.), but if it is used as a label on wares, the English equivalent is *hand-made*, because the English substantive denotes only the process, not its results. Usually, there are not only such isolated points of trouble, but also discrepancies rooted in the system. It is easy to decide that English substantives and adjectives will be considered equivalent to Czech substantives and adjectives, and to indicate pairs like Czech *nebe* : English *heaven*, Czech *nebeský* (adj.) : English *heavenly*, *celestial*, in a Czech-English dictionary. But there will also be pairs like Cz. *cihla* : Eng. *brick*, Cz. *cihlový* (adj.) : Eng. *brick* (as in *a brick wall*). The second pair of equivalents can be left without comment if the Czech user of the dictionary is supposed to have a fair knowledge of English. If this is not true, the entry of the second pair should contain an indication of how the equivalent is construed, e.g. by giving an example (*brick wall*). The example used here is easy to handle, but the real life of the lexicographer poses more difficult problems of this type. The main thing seems to be to see these discrepancies before one begins the concrete work and to decide on their solution in general, so that the individual instances are treated in a unified way in the whole dictionary.

The equivalent should be a real lexical unit of the target language which occurs or can occur in real sentences. (We shall see later that this requirement must be limited, but it is valid for the majority of instances.) The usual procedure is for the lexicographer to collect a broad range of typical contexts in the source language in which the respective lexical unit occurs. (It goes without saying that this can be shortened by using native speakers as informants, or by using one's own competence; but at least some collections of contexts - not necessarily long ones - usually are essential.) The lexicographer then tries to translate all these typical contexts into the target language, using in each instance the prospective equivalent of the target language. If the prospective equivalent fits into all these contexts, it is an absolute one; if not, it is partial and the entry will have to indicate some other (partial) equivalent(s) to cover the whole range of the lexical meaning of the entry word. The way the lexicographer presents the data in the dictionary is largely governed by the purpose of the prepared dictionary. Let us discuss some examples.

German *heiraten, sich verheiraten* "to marry" are usually considered equivalents of Chinese *xu jia*. One of the differences between them is that the Chinese lexical unit is used in reference to women only. In a dictionary whose only purpose is to help native speakers of German to understand Chinese texts the entry could have the basic form

xu jia : "heiraten, sich verheiraten"

The two equivalents are applicable in all contexts, so that it is not necessary to state the restriction of the Chinese lexical unit; and the German user needs no information about the German equivalents. But if the dictionary is to be more descriptive, and is to give the German user more information about Chinese lexical units, the Chinese semantic restriction will have to be indicated:

xu jia (von Frauen) : "heiraten, sich verheiraten"

If, on the other hand, the dictionary is intended to help the Chinese user produce German texts, it is necessary to indicate the difference between the two German partial equivalents, so that the user can make the right choice:

xu jia : "heiraten" ("to take in marriage"), "sich verheiraten" ("to get married")

(The English words in quotation marks symbolize an indication which would have the form of a gloss or of an explanation, either in German or in Chinese, in a real dictionary.)

A combination of the intentions mentioned requires, then, an entry of a form like

xu jia (von Frauen) : "heiraten" ("to take in marriage"), "sich verheiraten" ("to get married")

Another type of entry can be discussed with the help of the following example: *beinahe, fast* "almost, nearly" can be considered the German equivalent of Chinese *xian xie*. The Chinese contexts are roughly of the type: He nearly stumbled, fell, starved, died, knocked someone down, poisoned someone, etc. Let us, therefore, suppose that a Chinese-German dictionary is prepared which should also have some descriptive power. The entry then would have to contain a gloss stating the applicational restriction, for instance of the following form:

xian xie (bei negativen Ereignissen) "beinahe, fast"

In a Chinese-English dictionary, the entry could have the form

> *xian xie* (referring to negative events) "almost, nearly"

The applicational restriction could be stated in the form of an example or of some examples; the advantage of this method of presentation is that the information is more immediate, and, additionally, that it is less explicit than the gloss.

Let us now consider the English equivalents. They both have multiple meaning. If we accept Hornby's description of their meaning, we see that *almost* has two senses, viz.: (1) as in *He almost fell* (*almost* is replaceable by *nearly*), (2) as in *Almost no one believed her* (*almost* is not replaceable by *nearly*). The other equivalent, *nearly*, has (according to Hornby again) three senses, viz.: (1) as in *It is nearly 1 o'clock* (replaceable by *almost*), (2) as in *I have $20, but that will not be nearly enough for my journey* (not replaceable), (3) as in *nearly related* (not replaceable).

If we quote *almost, nearly* together as equivalents of the Chinese lexical unit, they disambiguate each other, because every user will assume that only that sense applies which is common to both of them.

On the other hand, if we consider the German equivalents *beinahe, fast*, we find that they are as close synonyms as possible, because a difference in their meaning is almost imperceptible. If this is so, we can ask why both of them should be quoted. There are two arguments in favor of citing both. First, the indication of synonyms in the target language helps the user to find various expressions he can use, if only for stylistic variation. And second, imperceptible as the difference is, there usually is some slight difference between the meaning of even such close synonyms, so that if both are indicated, the information is richer and the user is inspired to imagine yet other possible translations and synonyms. But in any event, even a large dictionary should not indicate too many synonyms of this type, and a small one can omit them.

In sum, we have discussed three types of indication of partial equivalents and synonyms:

(1) *heiraten; sich verheiraten* : a rule (semantic or grammatical) of the target language makes it predictable which of the two will be used;

(2) *almost, nearly* : both can be used, but only in those senses of their multiple meaning which overlap;

(3) *beinahe, fast* : either can be used, and the two taken together make the information somehow richer.

Although there are many borderline cases between these types, it is useful to know them; but it is above all types (2) and (3) which are difficult to distinguish. In type (1), it is preferable to put a semicolon between the two partial equivalents; in types (2) and (3), a comma is generally used.

Another type of problem can be illustrated by the following example. The German equivalent of Chinese *jin* is *alt* "old". When the lexicographer analyzes the contexts of the source language, he will perceive that they belong roughly to the three following groups: (1) old edition of a book, an old malady recurs, old society, old ideology, old dwelling, old job; (2) old method, old custom, old dream, old archive; (3) old equipment in industry, old material, old clothes. Unless the dictionary belongs to the smallest type, without any generative power in the target language, it will not be sufficient to state simply *jin : alt*, but it will be felt necessary to give richer indications. It will also be essential to indicate that the German equivalent must not be taken in one of its senses as in *Er ist 10 Jahre alt* "He is ten years old". If the dictionary proceeds, as usual, by the indication of synonyms, one can suppose an entry of approximately the following type:

> *jin* : (1) "alt, früher, ehemalig" (that is, say, "old, former, previous");
> (2) "alt, schon lange bestehend" ("old, existing for a long time");
> (3) "alt, gebraucht, durch langen Gebrauch abgemacht" ("old, worn out by long use").

When we consider these indications, we see that an equivalent like *alt* "old" undoubtedly is a lexical unit which can be immediately inserted into a German sentence, whereas *schon lange bestehend* or *durch langen Gebrauch abgemacht* are somehow felt as non-minimal, as expansions of what the simple *alt* can convey. But these non-minimal expansions have the advantage that they, when we see them in isolation, give more information about the lexical meaning of the source language. Equivalents of the first type are usually called translational equivalents, those of the latter type explanatory or descriptive ones. Naturally there are many equivalents which combine both advantages; for instance *gebraucht* "used" seems to be, in the example given, a good translational equivalent with great descriptive power.

Very frequently, it is necessary to give a translational equivalent and an explanatory one, or only an explanatory one. For instance, an English-French dictionary can hardly proceed by giving a simple equivalent of English *boyhood*, because there is no really good one. The explanatory equivalent would probably be something like *état de garçon*. But this cannot be inserted into sentences (or translation of sentences) like *In his boyhood, he . . ..* A more translational equivalent like *adolescence* or *jeunesse* is indicated. But these words are not restricted to male children in French, as the English word is. And so the entry would probably have to make a compromise and indicate, say,

boyhood : "adolescence, période de jeunesse" (d'un garçon)

The explanatory equivalent has the advantage of being very general, because it is situated rather on the notional than on the purely linguistic level. If the user grasps what is indicated, and if he knows French well, he will be able to understand many different English sentences, and he will feel free to adapt his French translations as need be. In various contexts, he may say, "Au temps d'adolescence ...", "Dans sa période de jeunesse ...", "Quand il était jeune garçon ...", but possibly and simply also "Quand il était jeune ...", etc.

The explanatory equivalent works particularly well if the target language is the user's native one, because it makes considerable demands on his knowledge. The advantage of the translational equivalent is that it is purely linguistic and that it offers the user directly an expression that can be used. But apart from the fact that it frequently conveys less information, the translational equivalent can cause a good deal of trouble to the lexicographer. Let us discuss an example. We said that Chinese *jin* has a good equivalent in German *alt* "old". The subsequent discussion has shown that the lexicographer will probably feel it necessary to add some further equivalents. This can be pushed too far. For instance, the lexicographer may find Chinese contexts in which the best translation would be "preceding, foregone, past, obsolete"; there will be contexts in which "ancient, antique, archaic" seem to fit well, etc. But to indicate all this would mean that the bilingual dictionary would grow into a synonymic dictionary of the target language. The lexicographer's task is to indicate the most general translational equivalents which have a broad range of application. And so the explanatory equivalent and the translational one are not so much opposed as one would think: they both act as representatives of groups of synonyms and near-synonyms, out of which the user may choose the most suitable one (if he knows them, or if he is able to use a monoglott or a synonymic dictionary of the target language). The difference between the two types is that the translational equivalent is always a possible choice for application in a sentence and sometimes the best one.

But while the lexicographer tries to indicate the best equivalents, he frequently is faced with the fact that he does not find any. For instance, it is hard to find an English equivalent for the Russian preposition *po* in such everyday expressions as Russian *po pribytii ego* "on his arrival", *po ukhode ego* "after he had gone, after his departure", *po okončanii* "as soon as he had finished". Grammatical information must be supplied instead of lexical equivalence, or in combination with it.

The so-called culture-bound words pose another problem, because they frequently have no lexical equivalent in the target language. There are basically three types of solution: (a) The lexicographer may try to create a translational equivalent by borrowing the respective word into the target language, frequently in a phonemically adapted form. (b) He may try to create a translational equivalent by coining a loan-translation, or by coining a new expression in the target language. (c) He may try to find an explanatory equivalent in the target language (with the eventual hope that it may become a translational one, if used frequently in future). If we take examples from a less known language, the three types are:

(a) Ossetic *ælam* : Eng. "alam" (borrowing)
(b) Oss. *ironvandag* : "Ossetic way" (new coinage by loan-translation)
(c) Oss. *ziw* : "collective help" (explanatory equivalent)

It is clear that the explanatory equivalent (c) gives the richest information; types (a) and (b) can be chosen only if it is expected that the respective words will have a high frequency in translated texts (where there will be explanatory notes, etc.). But for a real understanding, we need an explanation in all three types, for instance:

(a) *ælam* : "alam" (fruit and candy bound on a twig and carried by mounted participants at a funeral feast)
(b) *ironvandag* : "Ossetic way" (an ancient Ossetic funeral ritual)
(c) *ziw* : "collective help" (socially expected help, above all in agricultural work, organized within or by a group of people)

It depends on the lexicographer's decision (and this, in its turn, on the type and purpose of the dictionary), whether his explanations will be minimal (as here, type b), or whether they will verge on the encyclopedic types a, c); but they should have a uniform style through the whole dictionary.

The difference between what we call an explanatory (or descriptive) equivalent and an explanation is that the explanatory equivalent tends to be similar to a translational equivalent. If stabilized and accepted into the language, it can become a lexical unit of the target language. But an explanation tends to be very similar to a lexicographic definition (or is even identical with it) as used in monoglott dictionaries, and usually cannot aspire to becoming a lexical unit. But there is no need, I think, to stress that there are a great number of borderline cases.

And so we see that the bilingual lexicographer works basically with translational equivalents, synonyms, mutually disambiguating synonyms, mutually complementing synonyms, explanatory equivalents, and explanations. All of them have the purpose of informing the user about the meaning of the lexical unit of the source language, of supplying him with lexical units of the target language which can be used in source-language sentences, and of inducing in him a recollection of other suitable, near-synonymic lexical units of the source language even if they are not directly indicated.

A good entry of a bilingual dictionary also needs information usually supplied by illustrative examples (quoted or coined), by glosses, labels and similar means; but a discussion of this type of information would require another paper.

FOOTNOTE

1 This article is based on a section of my *Manual of Lexicography* (forthcoming). I wrote that book in co-operation with several colleagues who supplied material and examples from various languages. Full acknowledgement of those examples will be found in the book itself.

The Shape of the Dictionary For Mechanical Translation Purposes

by

L. Zgusta

Linguistics Research Center
The University of Texas at Austin
and
Department of Linguistics
The University of Illinois

548a

THE SHAPE OF THE DICTIONARY FOR MECHANICAL TRANSLATION PURPOSES.[1]

L. Zgusta

A dictionary of the type we have in mind here should contain the lexical units of the source language, selected according to the needs of the type of texts to be translated. Lexical equivalents of the target language should be coordinated with these lexical units in such a way that the choice is as precise and as automatic as possible. Great difficulties are caused in this task not only by the polysemy and homonymy of the lexical units of the source language, but also by the fact that the equivalents usually cannot be coordinated in a one-to-one way. We call "lexical equivalent" a lexical unit of the target language which has the same lexical meaning as the respective lexical unit of the source language; that means the equivalent should have the same polysemy, the same stylistic value, etc., as the lexical unit of the source language. However, this is seldom the case, and, consequently, more than one equivalent is often needed to cover the lexical meaning of the source word. We should, then, make the distinction between absolute equivalents, which comply with the definitional requirement of a one-to-one correlation in lexical meaning, and partial equivalents; but general usage allows us to

---

[1] Work on this article was performed at the Linguistics Research Center, University of Texas at Austin, Texas, under Rome Air Development Center contract. The paper profited from a discussion of its contents with M. Kay and R. Stachowitz.

speak about equivalents when it is usually the partial ones we have in mind.

The present article is not primarily concerned with the problem of (partial) equivalents, their choice, their mutual disambiguation and the delimitation of their applicability in an entry.[2] This article concentrates on problems of choice from among more than one (partial) equivalent within the entry of a lexical unit of the source language. The point of view taken here is that, on the one hand, the more we can rely on simple formal indications of the source language the better, but that, on the other hand, such simple formal indications do not always exist; and that one of the cardinal difficulties with which we have to cope is that the selection of a suitable (partial) equivalent is to be made by an agent which is by far less imaginative than the human mind.

Semantic difference in the source language (and, therefore, the necessity of a certain selection among the partial equivalents) is frequently indicated by some difference in form. The situation is rather simple if the difference of form is easy to detect. It is easy to change an entry of the type

German *in* → English *in, into*

into the following shape

German *in* + *Dat.* → English *in*

in + Acc. → English *into*

---

2 See "Equivalents and Explanations in Bilingual Dictionaries", to be published in .

Since we envisage, for the moment, only basic translational needs, this form of the entry should suffice to guarantee a good selection of the equivalent in sentences like German *in dem Wald gehen* - English *to walk in the wood*, and German *in den Wald gehen* - English *to walk into the wood*, that is, given the ability to recognize which German substantive is governed by *in* and whether it is dative or accusative.

The example just discussed is one of the simplest ones. It can be said that the recognition of semantic difference and the choice of the equivalent entailed by it are not difficult if the semantic difference is indicated by a clear morphological difference.

The formal distinction, however, does not necessarily have to be a morphological one; the main thing is that the distinction should be clear in itself and non-ambiguous. For instance, it should be easy to discern the polysemy of German *handgreiflich*, because in one of its senses, it is used exclusively with the forms of *werden*: *handgreiflich werden* "to use physical force". In its other sense, it is used with *machen*, *sein* and a few other verbs: *handgreiflich machen* "to make available", *handgreiflich sein* "to be available".

Perhaps more complicated is the following type of case. If we simplify the multiple meaning of German *ableiten*, we can construct an entry of the following type:

German *ableiten etwas* → English *to lead away*

ableiten etwas von etwas → *to derive*

(As in German *den Strom ableiten* - English *to lead the current away*, German *die Adjektive aus den Substantiven ableiten*, English *to derive adjectives from substantives*.) It would seem that it should not be too difficult to distinguish the two types of rections quite automatically, and make the choice accordingly. The next example will, however, be more complicated.

The simplest way to construct the entry of German *beraten* seems to be

German *beraten jemanden*	→	Eng. *to advise*
beraten etwas	→	*to deliberate (upon)*
beraten ueber etwas	→	*to deliberate (upon)*

The last two German rections are different in their grammatical form, but there is no semantic difference. On the other hand, there is no grammatical difference between the two first rections, but there is a semantic difference which entails a different choice of English equivalent. The abstract expression of the two rections in the lexicographic entry (*jemanden :: etwas*) is rather simple, and no human user of a dictionary could have difficulty with it. Still, for the purpose of automatic recognition and choice, the presence of this entry in the dictionary entails the necessity of indicating in the lemma of each substantive whether it belongs to the category "*jemand*" or "*etwas*". This should not be too difficult a task; let us, however, discuss yet another type of situation.

The entry of German *abhalten* can be constructed in the following way:

German *abhalten* (1) *jemanden von sich* → *to hold off*

(2) *jemanden von etwas* → *to hinder, prevent*

(3) *etwas* → (a) *to keep out*
→ (b) *to hold*

We see that within one rection, (3), there are two choices (a), (b) which are semantically governed: (a) is chosen if the object (represented by *etwas*) is, e.g., *Wasser, Naesse, Zugluft, Regen*; (b) is chosen if the object is, e.g., *Sitzung, Wahlen, Gericht*, etc.

Another example of this type is German *auslegen*. One of its rections (the most frequent one) is *auslegen etwas*. The respective part of the entry would have to have a form similar to the following one:

auslegen (1) *etwas* (a) [*im Ladenfenster*] → *to display*
(b) [*Geld*] → *to pay provisionally*
(c) [*Texte*] → *to interpret*

In a case like this, the really important indication is the one contained in brackets. And as every lexicographer knows, to construct these restrictive (or semantic) glosses (as they are frequently called in lexicographic theory) belongs among the most difficult tasks because it is hard to find the real limits of the restriction. This is, however, a purely lexicographic task which every good lexicographer is accustomed to coping with. It is not without significance that in the compilation of a dictionary of a living language, it is nearly always native informants who are used for this task. But in the situation envisaged in this article we try to count with an automatic choice from among the equivalents, and this causes much trouble. The reason is that every human user of a dictionary will immediately understand that an indication like [*im Ladenfenster*] is simply an example since goods can be displayed also on stands within a shop or in the market, and so on. Not only that; the human user will also understand that the restrictive gloss [*im Ladenfenster*] is, at the same time, the representative of a certain type of situation, since one can speak about somebody displaying his goods without mentioning where and how, and choice (a) is then entailed. Therefore, this part of the entry could also have the following form:

auslegen (1) *etwas* (a) [*Waren*] → *to display*

This restrictive gloss would have other difficulties of its own. We mention it to show that restrictive glosses have to be chosen from among various possibilities inherent in the facts of language.

In the same way, [*Geld*] is both an example and the representative of a class of synonyms, near-synonyms, and semantically related words (*eine Summe auslegen*). In (c), [*Texte*] would seem to be simply the hierarchically higher notion (*Oberbegriff*) comprising singularia like Bible, Homer, 6th Amendment to the Constitution, etc., or any text(s); but in reality, it must be understood as a representative of other expressions, too. There is no need to go as far as poetical language to have a sentence like *Falls die Daten=verarbeitungsmaschine den gestrigen Verkauf von Papieren auf der New Yorker Boerse falsch auslegt, dann . . .*

The difficulty of this problem is obvious. One of the easiest answers would be that we should increase the number of concrete examples quoted in the restrictive glosses. For instance, one could imagine the following form for the entry quoted above:

abhalten (3) *etwas* (a) [*Wasser, Naesse, Fluessigkeiten, Regen, Hagel, Wind, Zugluft*] → *to keep out*

The increase in the number of concrete examples in the restrictive glosses would be an enormous gain; but we should count with dozens and perhaps hundreds of them in one gloss. It does not seem to me, however, that the more or less exhaustive enumeration of examples could be a real solution. Let us discuss the following example. That part of the entry of German *verjuengen* which is concerned with technical terminology could have the following form:

verjuengen (1) *etwas* (a) [*Maßstab*] → *to reduce*
(b) [in biology][3] *to rejuvenate*

The restrictive gloss pertinent to (a) could be expanded by an enumeration of examples. I cannot, however, see that choice (b) could be governed by the indication of concrete examples. First, because the area of objects of rejuvenation, attempted or real, is rather vast; still, one can imagine a restrictive gloss with perhaps hundres of examples, e.g. [*Gewebe, ...Knochen, ...Zellen, ...Greise, ...Reflexfaehigkeit, ...Regenerationsfaehigkeit, ...etc.*]. But the second difficulty seems to be more grave. The area of objects of rejuvenation is not only vast; it is always getting more vast, and the very

---

[3] We do not take into consideration that the theory of lexicography usually distinguishes indications of this type from the restrictive (or semantic) glosses. Indications of this type are usually called labels.

purpose of science is to render it more vast. Consequently, one must take into consideration that after we have established our set of examples in the restrictive gloss, there will be biological texts reporting new investigations, discoveries, etc., concerned with new objects not stated among our examples; which would make a correct choice of the equivalent impossible. And since the main purpose of machine translation is to translate recent reports on new discoveries, etc. quickly, we can conclude that the choice of the equivalent cannot be based on an exhaustive enumeration of contextual examples (understood as key words), lest we block our way to the very goal we are trying to reach.

It seems that what is needed is a classification of all entry-words selected for the future dictionary into classes constituted by the restrictive glosses and the semantic criteria contained in them. For instance, since the correct choice of an equivalent in some entries depends on whether the object is a person or not, this category should be indicated in the lemma of each substantival entry-word; since a correct choice in another entry depends on whether the context is a biological one or not, the pertinent indication should be a part of the lemma of the respective entry-words. This should be done with all the restrictive glosses involved in the corpus of entries. It would require further researches, but it seems that the number of different restrictive glosses could be slightly reduced if they had, when possible, the form of hierarchically higher notions (*Ober=begriffe*), or if they indicated terminological areas (such as "biology", "chemistry", etc.). In this way, though the automatic procedure would certainly not command an abstractive ability of its own, it would possess a rich repertory of coherently constructed criteria for the necessary choices, applicable not to a broader semantic range of texts but at least to a much larger corpus of them than that on which the original investigations were based.

What has been discussed up to now is certainly no panacea.[4] There will be bases which will resist a generalization. For instance, another rection of German *auslegen* (not mentioned above) is *auslegen* (2) *etwas mit etwas*. In German, contexts characterized by this formal feature are not only clearly differentiated from the contexts of the type *auslegen* (1) *etwas*, but they also form a unified group, with a unified if general meaning. But there is no general equivalent in English, and the choice of the partial ones is governed by the object of the action. Consequently, we have to imagine that this part of the entry could have a form similar to the following one:

auslegen (2) *etwas mit etwas*	(a) [*Teppiche*]	→ *to cover with (carpets), to carpet*
	(b) [*Zement*]	→ *to line with (cement)*
	(c) [*Elfenbein*]	→ *to inlay, encrust with (ivory)*

On the other hand, the semantic classification is necessary even in cases where one would not immediately suspect it. We

[4] The method proposed here has some similarity to the method using so-called "semantic parameters". (Cf., e.g. Ubin, expression of the parameter Magn in Russian, in: Mašinuyj perevod i prikladnaja lingvistika, 11, 1969, p. 60 ff.; Šaljapina, Ways of expressing semantic parameters in English, ibid. p. 106 ff.) The difference, however, is that whereas the search for semantic parameters leads to establishing more or less purely notional frameworks and constructions, the present approach tries to remain as close to really occurring contexts as possible.

have stated above that it is relatively easy to find a solution for those cases in which a difference in meaning is indicated by a difference in form, preferably in morphology. The dictionary can make use of such differences. Sometimes, the morphological distinction alone is sufficient to indicate the difference in meaning. For instance, the series of German forms *die, der, der, die Diaet, die, der, den, die Diaeten* can be seen as a normal paradigm of a feminine substantive. There is, however, the semantic difference that the forms of the singular require the English equivalent "diet, regimen", whereas those of the plural require the equivalent "daily allowances". Such a situation is easy to solve. Probably every lexicographer will take *Diaet* "diet" as singulare tantum, and *Diaeten* "daily allowances" as another word, a plurale tantum; and such a solution is undoubtedly even more practical for an automatic procedure.

But not all cases are as beautifully clear-cut as this. A morphological difference is sometimes of only partial value. For instance, if we try to find an English equivalent for German *Ort*, in its application as a technical term, we can arrive at the following result:

Ort (1) [in geography] → *place*
(2) [in geometry] → *locus*

It is usually maintained that the two are sufficiently differentiated by the fact that *Ort* (1) [geogr.] has the plural *Orte*, whereas *Ort* (2) [geom.] has the plural *Oerter*. This morphological distinction is fully sufficient for the plural; if we had to deal with pluralia tantum, this part of the reduced entry could have the form:

Orte → *places*
Oerter → *loci*

Since the singular is not morphologically differentiated, a semantic (that is, contextual) differentiation is necessary.

The same situation can be observed in *die Mutter*, plural *Muetter* "mother"; *die Mutter*, plural *Muttern* "nut": the singular is not morphologically differentiated. On the contrary *das Erkenntnis* "decision, judgment, sentence" and *die Erkenntnis* "comprehension, perception, cognition" are well differentiated in the singular, but since they have identical forms in the plural, *die Erkenntnisse*, they should be semantically differentiated as a juristic, and a psychological and philosophical term, respectively.

Cases like those just discussed are particularly disagreeable if there is a semantic difference only in a small part of a paradigm. Let us discuss an example. We can imagine a strongly reduced form of the entry of German *erledigen* as follows:

erledigen	(1)	*etwas*	→ *to finish, arrange, settle*
	(2)	*jemanden*	→ *to dispose of*

This German verb has the normal participle *erledigt* which has the same polysemy: *Das ist erledigt* "That's settled", *Durch die naechste Saeuberung wird er erledigt (werden)* "He will be disposed of by the next purge". This form, however, has other senses of its own, so that an eventual entry could have the following form:

erledigen	(1)	*etwas*	→ *to finish, arrange, settle*
	(2)	*jemanden*	→ *to dispose of*
erledigt	(1)	participle to *erledigen* (1), (2)	
	(2)	*[Person]*	→ *done for, finished*[5]
	(3)	*[Stelle, Posten]*	→ *vacant*[6]

5 As in: *Nach disen Strapazen bin ich erledigt* "I'm finished after these labors".

6 As in: *Jede erledigte Stelle* "Each vacant situation".

Cases like this are rather treacherous. Dictionaries are normally built on the principle that the form of the source language in which the entry-word is indicated and to which the equivalent is coordinated (the so-called canonical form) is a representative of the whole paradigm of the entry-word, that is, if the source language happens to be a language with paradigms. Therefore, before the inclusion of a word, with its equivalent(s), into the dictionary, its whole paradigm should be checked, and the more important semantic peculiarities of its single forms should be duly noted.

If polysemy needs semantic differentiation by the context, we can expect that the same will be true of homonymy (overlapping as the two notions are). The situation is basically identical, so there is no need to discuss special examples. There is, however, a special type of situation, in which a homonymous pair or polysemous meanings are differentiated by the form. German *Abrede* generally has the meaning of "understanding, agreement"; but the set expression *in Abrede stellen* means "to disavow, to dispute". This expression being rather frequent, the reduced entry could have a form like:

Abrede (1) → *understanding, agreement*
(2) *in Abrede stellen* → *to disavow, to dispute*[7]

This brings us to a topic which I shall mention only briefly, namely the fact that there are combinations of words which are set, which have a unified meaning, and which even

[7] This type overlaps with the type of *handgreiflich werden* as discussed above.

function as a lexical unit of a language.[8] There are many various types of them. A dictionary of the type under consideration here, prepared for coping with texts of a limited range only, will hardly select many colorful idioms such as *Das Hasenpanier ergreifen* "To fly away". But it will have to list frequently occurring set expressions like *in Abrede stellen*, particularly when their meaning is not predictable from that of their individual parts. Also, a dictionary of our type will probably select many technical terms which consist of more than one word. The technical terminology of any science gives many examples of the type *leichte Infanterie, schwere Infanterie,* etc. The situation in German is particularly easy, because a large number of such terminological coinages have the form of compound word, cf. *Dampflokomotive* "steam engine". Still, there is no predictable regularity in this, cf.

Sauerkraut "pickled cabbage", but
saure Gurken "pickled cucumber",

so that the lexicographer has to check the whole semantic area carefully. It will also be necessary to have the productive parts of compound words listed in the dictionary as entries of their own if they have a regular effect on the meaning of the whole compound. With real compound words, this is not too frequently the case, but affixes and elements which approach the status of affixes can be treated this way. For instance, German *ur-* → "proto-"; *pseudo-* → English "pseudo-", etc. Such an indication has the big advantage that it is so to say productive: it can take care of newly coined expressions (assuming they are coined regularly), unknown at the moment of the compila-

[8] On these, see my "Multiword Lexical Units", linguistic studies presented to A. Martinet I, p. 578 ff.

tion of the dictionary.

There are some points which may deserve to be mentioned. Many a dictionary tends to forget that we find multiword lexical units not only among the denotative words. But the inclusion into the dictionary of expressions like German *ab und zu* "from time to time", or German *auf und ab* "up and down" is useful. And again, we will have to put into the dictionary indications of how to discern polysemy. Consider the difference between German *von (heute, nun, jetzt, gestern,* etc.) *ab,* English "from (today, now, yesterday, etc.) onwards", and German *vom Bahnhof ab (geht die Straße bergab)* "the street begins to go downhill at the station". Therefore, a strongly reduced part of the entry should have the form:

von .... ab (1) [*Zeitangaben*] → *from ... onwards* .
(2) [*Ortsangaben*] → *at, beyond*

A particularly obnoxious type of set expressions are those which allow a certain variation. For instance, German *es (tut, schadet, macht) nichts* has a good English equivalent in "it does not matter". It would seem that there is no complication in this. Let us, however, consider the following sentences: *Es tut nichts.* "It does not matter". -- *Er tut nichts.* "He is doing nothing". This shows us that a set expression may have parts which allow some variation, but again, it has parts that do not. Therefore, a good dictionary will have to contain indications of the following type:

es (tut, schadet, macht) (nichts, wenig) →
→ *it does not matter.*

It can be said that the most difficult problem will be how to guarantee that an automatic device will make the correct choice from among the partial equivalents of the target language. This task is so difficult in itself that we should not make it even more difficult by indicating too many (partial) equivalents of the target language. Let us consider some entries discussed above. An entry of the type

auslegen (2) *etwas mit etwas*
(a) [*Teppiche*] → "*to cover with* (*carpets*), *to carpet*"
(c) [*Elfenbein*] → "*to inlay, encrust with* (*ivory*)"

does not strike us as unusual. The verbs *to inlay* and *to encrust* are synonymous for all practical purposes. Every human user of a dictionary is accustomed to understanding an indication like this, so that he is free to use either one or another synonym.

On the other hand, if we take a part of the entry of *erledigen* discussed above

erledigen (1) *etwas* → *to finish, arrange, settle*

we see that it has the same form, but the difference is in the fact that the English verbs are rather mere near-synonyms than full synonyms. Again, a human user is accustomed to seeing entries of this type in any dictionary. Some dictionaries try to distinguish the synonyms from the near-synonyms by using commas in the first and semicolons in the second case: *to inlay, encrust with*, but *to finish;* (*to*) *arrange;* (*to*) *settle*. It is, however, extremely difficult to make the distinction in a systematic way, there being probably more borderline cases than clear-cut ones. And then again, a human user does not need a typographical indication of the distinction so badly: if he is a native speaker of the target language, he knows the distinction anyhow; if he is a speaker of the source language, he may make an error in his choice, but an error which will not be too grave, and with growing knowledge of the target language, he will also acquire the "feeling" for when to use one or another of the near-synonyms.

This is how bilingual dictionaries, particularly the smaller ones, operate: they rely on the abilities and knowledge of the human user. The indications of such dictionaries very frequently have the main purpose of triggering in the human user thinking and imaginative processes which make him recollect words and expressions not immediately indicated in the dictionary.[9] We cannot rely on all these abilities when we construct a dictionary for mechanical use. Therefore, the rule should be that there should be no unspecified indication of synonyms as partial equivalents: if there is more than one partial equivalent, they should be accompanied by the necessary restrictive glosses which will show which to choose. If both equivalents can really be used unrestrictedly, i.e. if they are fully synonymous, it is possible to indicate only one of them (preferably the more frequently occurring one) or to indicate the possibility of free variation, e.g. for stylistic purposes.

[9] This statement is focussed particularly on bilingual dictionaries of living languages for general use. Large philological dictionaries of dead languages are of a different type: they frequently contain an enormous mass of quoted contexts with concrete translations and thus make the information given quite factual and concrete. The human user, however, will still tend to go beyond the indications of this dictionary, since, after all, the indications of a dictionary and an adroit translation of a text are always two things.

To prepare a dictionary which will reach this degree of explicitness and accuracy is an extremely difficult task. Moreover, I am afraid that even when all this is done, there will still occur situations in which the automatic device will not be able to make a choice. This may occur, for instance, in any text where the relevant context is not close to the passage which needs disambiguation. It would seem that in such a situation no random choice should be made but both (or all) possible equivalents should be printed in the output with a sign showing their mutual complementarity.

A similar but much worse situation will occur when the automatic device is faced with a neologism, i.e. with a genuinely new expression or with an "old" expression used with a new sense. To discuss this difficulty, however, is quite a different task, because an attempt at the solution of this problem would require an investigation of the regularity of new coinages. For instance, new terminological coinages tend to have a high degree of regularity. In any case, a discussion of these problems must be reserved for another occasion.

BIBLIOGRAPHY

ARAPOV, M. V.

1967 *Sintaksicheskaya model' yazykov* (A syntactic model of languages) [with V. B. Borshchev]. Moscow: Nauka.

Avtomatizatsiya perevoda tekstov (Automatization of the translation of texts). Editorial introduction to a partial translation of the ALPAC report *Language and Machines, Computers - Translation and Linguistics*. NTI, ser. 2, No. 8.

BACH, EMMON

1971 "Syntax since *Aspects*", in *Report of the 22nd Annual Round Table Meeting*. Washington, D.C.: Georgetown University Press.

BAR-HILLEL, YEHOSHUA

1964 *Language and Information*. Reading, Mass.: Addison-Wesley Publishing Co..

1967a "Dictionaries and meaning rules". *Foundations of Language* 3:409-14.

1967b "Measures of syntactic complexity" (with A. Kasher and E. Shamir], 29-50 in *Machine Translation*, A.D. Booth, ed. Amsterdam: North-Holland. New York: John Wiley & Sons.

1967c "Review of *The Structure of Language: Readings in the Philosophy of Language* (ed. by J. A. Fodor and J.J. Katz)". *Language* 43:526-50.

1968 "Universal semantics and philosophy of language: Quandaries and prospects", in *Substance and Structure of Language*, J. Puhvel, ed., Berkeley and Los Angeles, California: University of California Press.

1969 "Formal logic and natural languages: A symposium". *Foundations of Language* 5:256-84.

1970 "Some reflections on the present outlook for high-quality machine translation (Position paper on machine translation in 1970)". Austin, Texas: Linguistics Research Center, The University of Texas at Austin (appended).

BORSHCHEV, V. B.

1967 *Dispozitsii, algoritmi i porozhdayushchiye procedury* (Dispositions, algorithms, and generative procedures) [with Yu. A. Shreider]. Moscow: Nauka.

FILLMORE, CHARLES

1968a "The case for case", 1-88 in *Universals in Linguistic Theory*, E. Bach and R. Harms, eds.. New York: Holt, Rinehart & Winston.

1968b "Lexical entries for verbs". *Foundations of Language* 4:373-93.

1968c "Types of lexical information", 65-103 in *Working Papers in Linguistics No. 2*. Columbus, Ohio: Ohio State University. [Also to be in *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics, Anthropology, and Psychology*, D. Steinberg and L. Jakobovits, eds.. Cambridge, Mass.: Cambridge University Press; and in *Proceedings of the Balatonszabadi Conference on Mathematical Linguistics*, F. Kiefer, ed. Dordrecht, Holland: D. Reidel.]

1969 "Verbs of judging: An exercise in semantic description", 91-117 in *Papers in Linguistics* 1:1. Tallahassee, Florida: Florida State University.

1970: "On a fully developed system of linguistic description". Austin, Texas: Linguistics Research Center, The University of Texas at Austin.

FRASER, J. BRUCE

1965 *An Examination of the Verb-Particle Construction in English*. Ph.D. dissertation. Cambridge, Mass.: M. I. T.

1966a "Some remarks on the verb particle combination in English", 45-61 in *Report of the 17th Annual Round Table Meeting*, C.I.J.M. Stuart, ed.. Washington, D.C.: Georgetown University Press.

1966b "Survey of automated language processing 1966" [with D.G. Bobrow and M.R. Quillian]. Cambridge, England: Bolt, Beranek & Newman.

1969 "An augmented state transition network analysis procedure", 557-67 in *Proceedings of the International Joint Conference on Artificial Intelligence*. Bedford, Mass.: The MITRE Corporation.

FRASER, cont'd

1970 "Idioms within a transformational grammar". *Foundations of Language* 6:22-42.

GARVIN, PAUL L.

1966a *Predication Typing--A Pilot Study in Semantic Analysis* [with J. Brewer and M. Mathiot]. Canoga Park, California: Bunker-Ramo Corporation.

1966b "Some comments on algorithm and grammar in the automatic parsing of natural languages". *Mechanical Translation* 9:2-3.

1967 "Machine translation - fact or fancy?" *Datamation*, April.

1970 "Operational problems of machine translation: a position paper". Austin, Texas: Linguistics Research Center, The University of Texas at Austin. (appended).

JOSSELSON, HARRY H.

1967 "Lexicography and the computer", 1046-1059 in *To Honor Roman Jakobson*. The Hague, Holland: Mouton.

1968a "The lexicon: a matrix of lexemes and their properties", in *Proceedings of the Balatonszabadi Conference on Mathematical Linguistics*, F. Kiefer, ed.. Dordrecht, Holland: D. Reidel.

1968b *Research in MT: Russian to English. Ten Year Summary Report*. Detroit, Michigan: Wayne State University.

1969a "The lexicon: a system of matrices of lexical units and their properties", paper presented at the International Conference on Computational Linguistics. Stockholm, Sweden: KVAL.

1969b "A linguistic interpretation of machine translation in the sixties", I/1-72 in *Research in Computer-Aided Translation, Russian - English*, 11th Annual Progress Report. Detroit, Michigan: Wayne State University.

1970 "The matrix as a concept for structuring the lexicon". Detroit, Michigan: Wayne State University, Xerox.

KARTTUNEN, LAURI

1967 "The identity of noun phrases". Santa Monica, California: The RAND Corporation.

1968a "What do referential indices refer to?" Santa Monica, California: The RAND Corporation.

1968b "What makes definite noun phrases definite?" Santa Monica, California: The RAND Corporation.

1969a "Discourse referents", Preprint No. 70, International Conference on Computational Linguistics. Stockholm, Sweden: KVAL

1969b "Problems of reference in syntax". Austin, Texas: The University of Texas at Austin, mimeograph.

1970a "The logic of English predicate complement constructions". Austin, Texas: Linguistics Research Center, The University of Texas at Austin (appended)

1970b "On the semantics of complement sentences", 328-39 in *Papers from the 6th Regional Meeting*. Chicago, Illinois: Chicago Linguistic Society.

to appear "Implicative verbs". *Language*.

KAY, MARTIN

1967 "Experiments with a powerful parser", paper No. 10 in *Proc. Deuxième Conference Internationale sur le Traitement Automatique des Langues*, Grenoble, France. Santa Monica, California: The RAND Corporation.

1969 "Computational competence and linguistic performance". Santa Monica, California: The RAND Corporation.

1970a "The MIND system: The morphological-analysis program" [with G. R. Martins]. Santa Monica, California: The RAND Corporation.

1970b "The MIND system: The Structure of the semantic file" [with S. Y. W. Su]. Santa Monica, California: The RAND Corporation.

1970c "Performance grammars". Santa Monica, California: The RAND Corporation.

KULAGINA, O. S.

1969 "Eshche raz k voprosu o realizatsii avtomaticheskogo perevoda (Once more on the problem of the realization of automatic translation)" [with I. A. Mel'chuk and V. Yu. Rozentsveyg]. NTI, ser. 2, No. 11.

LAKOFF, GEORGE

1965 *On the Nature of Syntactic Irregularity*, Ph.D. dissertation. Bloomington, Indiana: Indiana University. *Mathematical Linguistics and Automatic Translation Report No. NSF-16.* Cambridge, Massachusetts: Harvard Computation Laboratory. [Published as *Irregularity in Syntax*. New York: Holt, Rinehart & Winston, 1970.]

1968a "Deep and Surface grammar". Bloomington, Indiana: Indiana University Linguistics Club, mimeograph.

1968b "Instrumental adverbs and the concept of deep structure". *Foundations of Language* 4:4-29.

1968c "Is deep structure necessary?" [with J. R. Ross]. Bloomington, Indiana: Indiana University Linguistics Club, mimeograph.

1968d "Pronouns and reference". Bloomington, Indiana: Indiana University Linguistics Club, mimeograph.

1969 "On derivational constraints", 117-39 in *Papers from the 5th Regional Meeting, Chicago Linguistic Society*. Chicago, Illinois: Department of Linguistics, University of Chicago.

1970 "Natural logic and lexical decomposition", 340-62 in *Papers from the 6th Regional Meeting, Chicago Linguistic Society*. Chicago, Illinois: Chicago Linguistic Society.

to appear "On generative semantics", in *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics, Anthropology, and Psychology*, D. Steinberg and L. Jakobovits, eds.. Cambridge, England: Cambridge University Press.

LAKOFF, ROBIN

1969 "Some reason why there can't be any some-any rule." *Language* 45:608-15.

LAKOFF ROBIN, cont'd

1970 "Tense and its relation to participants". *Language* 46:838-49.

to appear "Questionable answers and answerable questions", in *Papers in Linguistics in Honor of Henry and Renee Kahane*, B. Kachru, et al., eds.

LEHMANN, WINFRED P.

1969 *Research in Russian-English Machine Translation on Syntactic Level*, vols, 1 & 2 [with L. W. Tosh, R. R. Macdonald, and M. Zarechnak]. Austin, Texas: Linguistics Research Center, The University of Texas at Austin.

1970 *Research in German-English Machine Translation on Syntactic Level*, vols. 1 & 2 [with R. Stachowitz and L. W. Tosh]. Austin, Texas: Linguistics Research Center, The University of Texas at Austin.

LYONS, JOHN

1967 "A note on possessive, existential, and locative sentences". *Foundations of Language* 3:390-96.

1970 "The feasibility of high-quality machine translation". Austin, Texas: Linguistics Research Center, The University of Texas at Austin (appended).

McCAWLEY, JAMES D.

1968a "Concerning the base component of a transformational grammar". *Foundations of Language* 4:243-69.

1968b "Lexical insertion in a transformational grammar without deep structure", 71-80 in *Papers from the 4th Regional Meeting, Chicago Linguistic Society*. Chicago, Illinois: Department of Linguistics, University of Chicago.

1968c "The role of semantics in a grammar", 124-69 in *Universals in Linguistic Theory*, E. Bach and R. Harms, eds.. New York: Holt, Rinehart & Winston.

1970a "Semantic representation", 227-48 in *Cognition: A Multiple View*, P. Garvin, ed.. New York and Washington, D.C.: Spartan.

1970b "Where do noun phrases come from?" in *Readings in English Transformational Grammar*, R. Jacobs and P. Rosenbaum, eds.. Waltham, Massachusetts: Ginn-Blaisdell.

MEY, JACOB

1969 "Syntax or semantics: Some controversial issues in computational linguistics". *Norsk Tidsskrift for Sprogvidenskap* 23:59-70.

1970 "Toward a theory of computational linguistics", paper presented at the Annual Meeting of the Association for Computational Linguistics. Austin, Texas: Linguistics Research Center, The University of Texas at Austin.

NATIONAL ACADEMY OF SCIENCES - NATIONAL RESEARCH COUNCIL

1966 *Language and Machines: Computers in Translation and Linguistics*, a report by the Automatic Language Processing Advisory Committee Division of Behavioral Sciences, Pulbication 1416, Washington, D.C.

NIDA, EUGENE A.

1969 *The Theory and Practice of Translation* [with Charles R. Taber]. Leiden: Brill.

PETERS, P. STANLEY, Jr.

1969a "Ambiguity, completeness and restriction problems in the syntax-based approach to computational linguistics" [with R. Tabory]. *Linguistics* 46:54-76.

1969b "A note on the universal base hypothesis" [with R. W. Ritchie]. *Journal of Linguistics* 5:151-52.

PETRICK, STANLEY R.

1965a "On the relative efficiencies of context-free grammar recognition" [with T. V. Griffiths]. Bedford, Massachusetts: USAF Cambridge Research Laboratories.

1965b *A Recognition Procedure for Transformational Grammars*. Ph.D. dissertation. Cambridge, Massachusetts: M. I. T.

1966 "A program for transformational syntactic analysis". Bedford, Massachusetts: USAF Cambridge Research Laboratories.

1967 "Syntactic analysis" [with S. J. Keyser]. Bedford, Massachusetts: USAF Cambridge Research Laboratories.

PETRICK, cont'd

1969 "On coordination reduction and sentence analysis" [with P. M. Postal and P. S. Rosenbaum]. *Communications of the ACM* 12:223-33.

1971a "On the use of syntax-based translators for symbolic and algebraic manipulation", 224-37 in *Proceedings of the Second Symposium on Symbolic and Algebraic Manipulation.* New York: Association for Computing Machinery.

1971b "Syntactic analysis for transformational grammars". Austin, Texas: Linguistics Research Center, The University of Texas at Austin.

1971c "Syntactic analysis requirements of machine translation". Austin, Texas: Linguistics Research Center, The University of Texas at Austin (appended).

PUMPYANSKIY, A. L.

1966 *Upryazhneniya po perevodu nauchnoy i tekhnicheskoy literatury* (Studies on the translation of scientific and technical literature). Moscow: NAUKA.

ROSS, JOHN ROBERT

1965 "Underlying structures in discourse" [with T. G. Bever], VIII/1-12 in *Proceedings of the Conference on Computer-Related Semantic Analysis.* Detroit, Michigan: Wayne State University.

1967 *Constraints on Variables in Syntax.* Ph.D. dissertation. Cambridge, Massachusetts: M. I. T. Bloomington, Indiana: Linguistics Club, mimeograph, 1968.

1970a "On declarative sentences", in *Readings in English Transformational Grammar,* R. Jacobs and P. Rosenbaum, eds.. Waltham, Massachusetts: Ginn-Blaisdell.

1970b "Gapping and the order of constituents", in *Proceedings of the 10th International Congress of Linguistics,* Bucharest.

SIMMONS, ROBERT F.

1966 "Storage and retrieval of aspects of meaning in directed graph structures". *Communications of the ACM* 9:211-15.

SIMMONS, cont'd

1967 "Answering English questions by computer", 253-89 in *Automated Language Processing*, H. Borko, ed., New York: John Wiley & Sons. [Also in *Communications of the ACM* 8 (1965):53-70.]

1970a "Generating English discourse from semantic networks" [with J. Slocum]. Austin, Texas: Computer Assisted Instruction Laboratory, The University of Texas at Austin.

1970b "Natural language question-answering system". *Communications of the ACM* 13:15-31.

STACHOWITZ, ROLF

1969 *On the definition of the term "discontinous constituents" in context-free phrase structure grammar.* Ph.D. dissertation. Austin, Texas: The University of Texas at Austin.

1970a "The construction of a computerized dictionary", paper presented at the Modern Language Association Lexicographical Conference, Columbus, Ohio. Austin, Texas: Linguistics Research Center, The University of Texas at Austin.

1970b "A model for the recognition and production of synonymous expressions with different deep structures", paper presented at the Conference on Linguistics at the University of Iowa, Iowa City. Austin, Texas: Linguistics Research Center, The University of Texas at Austin.

1971a "Lexical features in translation and paraphrasing: an experiment". Austin, Texas: Linguistics Research Center, The University of Texas at Austin.

1971b "Requirements for Machine Translation: Problems, Solutions, Prospects". Austin, Texas: The University of Texas at Austin (appended).

to appear *Normalization of Natural Language for Information Retrieval* [with W. P. Lehmann]. Austin, Texas: Linguistics Research Center, The University of Texas at Austin.

Development of German-English Machine Translation System [with W. P. Lehmann]. Austin, Texas: The University of Texas at Austin.

SWANSON, ROWENA

1967 *MOVE THE INFORMATION... A Kind of Missionary Spirit.* Arlington, Virginia: USAF Office of Aerospace Research.

1970 "Trend in information handling in the United States". Arlington, Virginia: USAF Office of Scientific Research.

WALKER, DONALD E.

1965a "English preprocessor manual". Bedford, Massachusetts: The MITRE Corporation.

1965b "The MITRE syntactic analysis procedure for transformational grammars" [with A. M Zwicky, J. Friedman, and B. C. Hall], 317-26 in *Proceedings of the 1965 Fall Joint Computer Conference.* New York and Washington, D.C.: Spartan.

1966 "Recent developments in the MITRE syntactic analysis procedure" [with P. G. Chapin, M. L. Geis, and L. N. Gross]. Bedford, Massachusetts: The MITRE Corp.

1967a "On-line text processing: Introduction and overview". Bedford, Massachusetts: The MITRE Corporation.

1967b "SAFARI, an on-line text-processing system", Vol. IV, 144-47 in *Proceedings of the American Documentation Institute.* London, England: Academic Press.

1969a "Computational linguistic techniques in an on-line system for textual analysis". Bedford, Massachusetts: The MITRE Corporation.

1969b "on-line computer aids for research in linguistics" [with L. N. Gross], 1531-36 in *Information Processing 68.* Amsterdam: North-Holland.

1970 "The current status of computer hardware and software as it affects the development of high quality machine translation". Austin, Texas: Linguistics Research Center, The University of Texas at Austin (appended).

WINOGRAD, TERRY

1970 *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language.* Ph.D. dissertation. Cambridge, Massachusetts: M. I. T. [Revised version published 1971, Cambridge, Massachusetts: M. I. T., Project MAC.]

ZGUSTA, LADISLAV

1970a "Equivalents and explanations in bilingual dictionaries", paper presented at the Modern Language Association Lexicographical Conference, Columbus, Ohio. Austin, Texas: Linguistics Research Center, The University of Texas at Austin (appended).

1970b "The shape of the dictionary for mechanical translation purposes". Austin, Texas: Linguistics Research Center, The University of Texas at Austin (appended).

to appear *Manual of Lexicography*. Prague, Czechoslovakia: Academy of Sciences.

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
University of Texas Linguistics Research Center Austin, Texas 78712	UNCLASSIFIED
	2b. GROUP N/A

3. REPORT TITLE

FEASIBILITY STUDY ON FULLY AUTOMATIC HIGH QUALITY TRANSLATION
Volume II

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Final Report 1 February 1970 - 30 June 1971

5. AUTHOR(S) (First name, middle initial, last name)

Dr. Winifred P. Lehmann
Dr. Rolf Stachowitz

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
December 1971	252	108

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
F30602-70-C-0129 Job Order No. 45940000	None
	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) RADC-TR-71-295, Volume II (of two)

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
None	Rome Air Development Center (IRDT) Griffiss Air Force Base, New York 13440

13. ABSTRACT

This report presents the results of a theoretical inquiry into the feasibility of a fully automatic high quality translation (FAHQT), according to Bar-Hillel's definition of this term. The purpose of this inquiry consisted in determining the viability of the FAHQT concept in the light of previous and projected advances in linguistic theory and software/hardware capabilities. The corollary purpose was to determine whether this concept can be taken into consideration a a legitimate and justifiable objective of R&D. The effort was supported by 20 expert consultants from the various universities and research centers in the U.S.A. and abroad. Conclusions and recommendations are presented on pages 44-50 of the report. Individual contributions of participants and consultants reflect a wide range of opinions concerning the prospects of FAHQT in intermediate and long range of R&D.

DD FORM 1 NOV 65 1473

UNCLASSIFIED
Security Classification

UNCLASSIFIED
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Linguistic Theory Computational Linguistics Machine Translation R&D Research in Syntax/Semantics Lexicography in Machine Translation Software/Hardware Capabilities						

UNCLASSIFIED
Security Classification