REQUIREMENTS FOR MACHINE TRANSLATION: PROBLEMS, SOLUTIONS, PROSPECTS

by

Rolf Stachowitz

Linguistics Research Center
The University of Texas at Austin

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 achievement 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:

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

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):

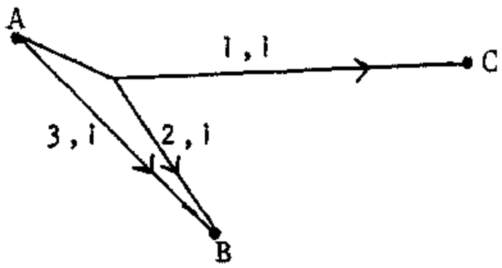


Figure 1

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:

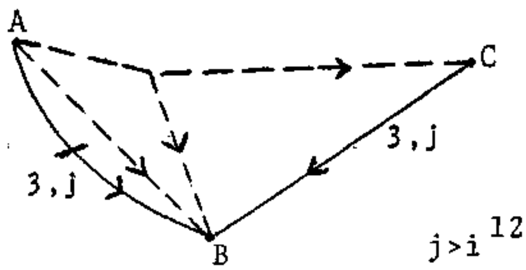


Figure 2

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

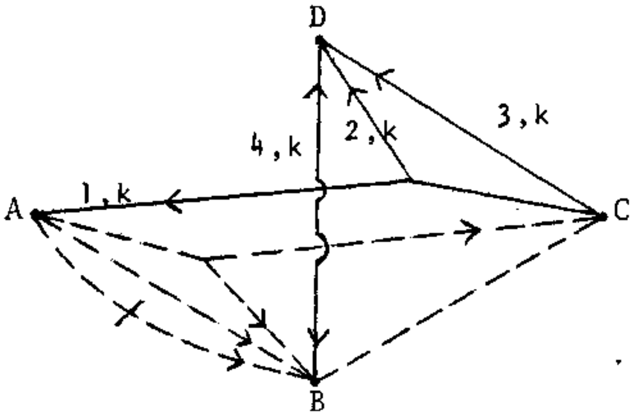


Figure 3

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

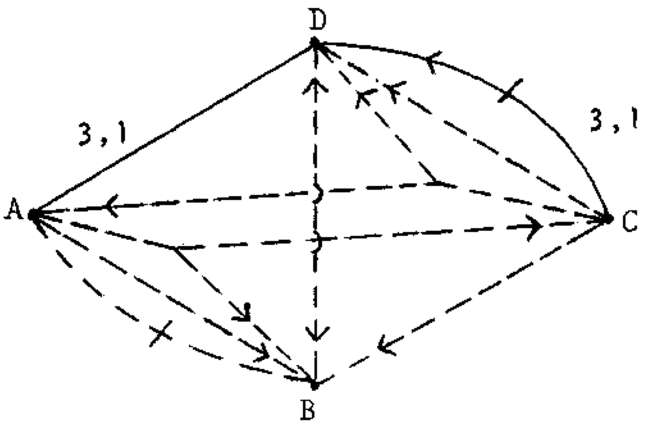


Figure 4

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 logical or.[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

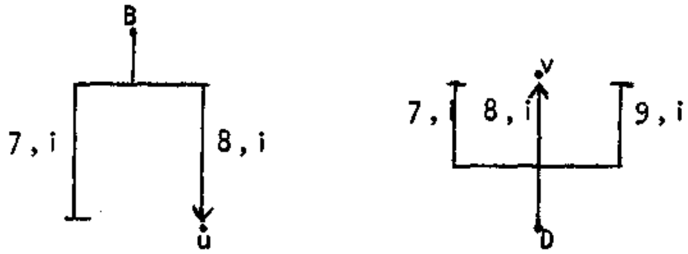


Figure 6

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:

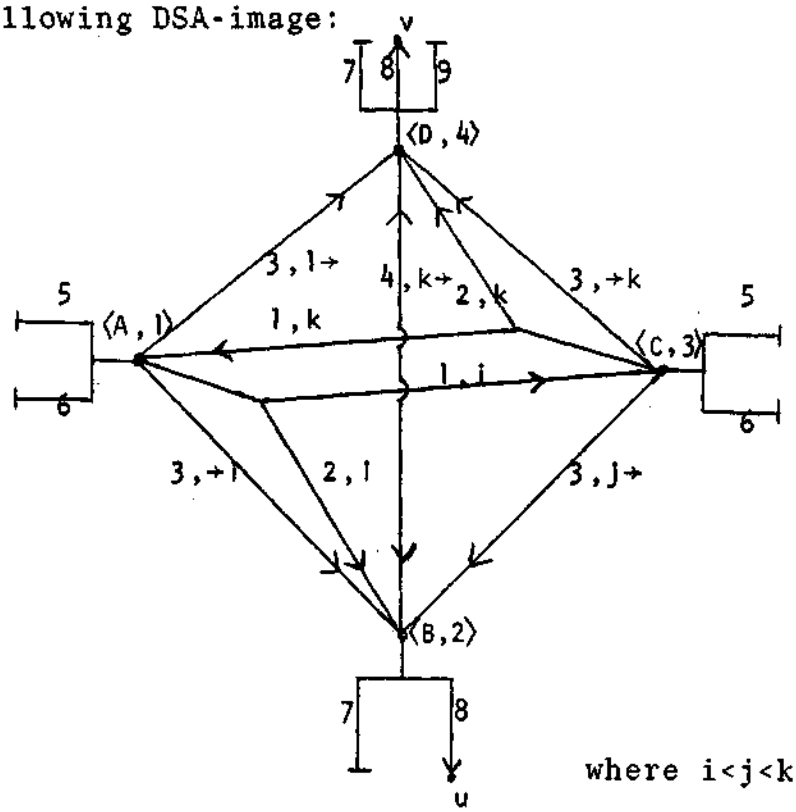
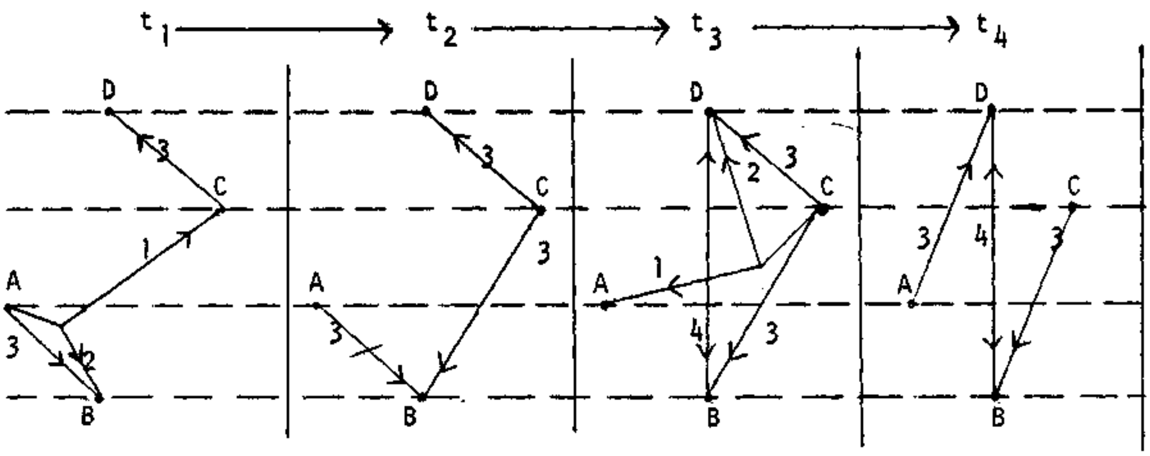


Figure 7

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:

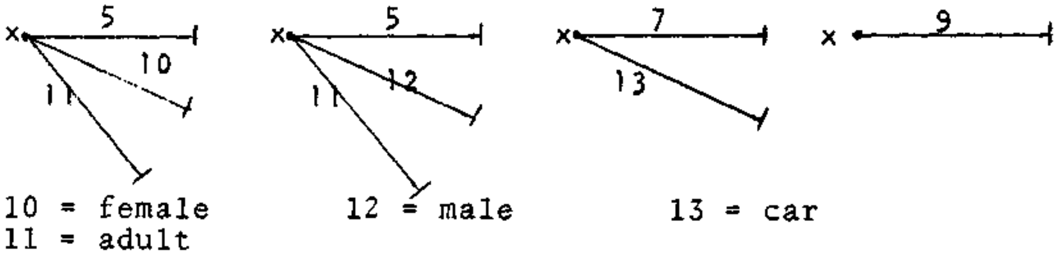


Figure 9

Sentence 11 will result in the following DSA-image:

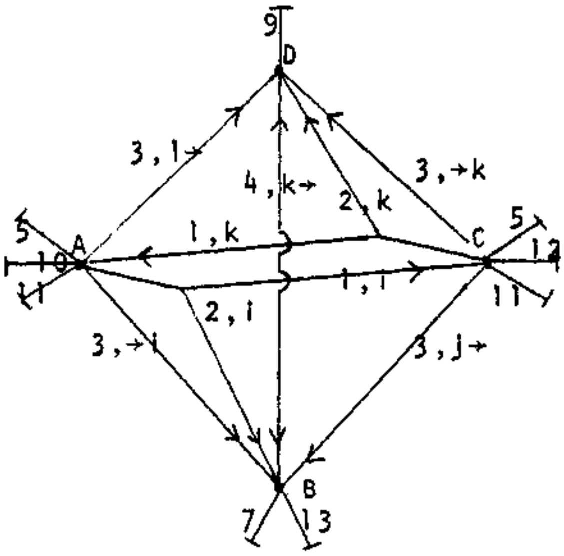


Figure 10

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

Further assume that device K has a meaning rule component with inference rules, statements of definitions, and equivalence 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.)

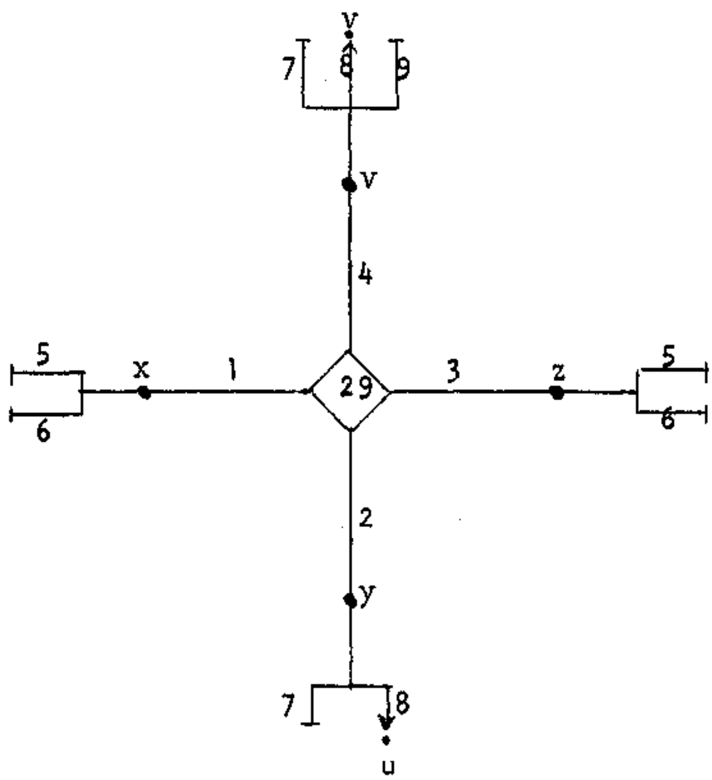
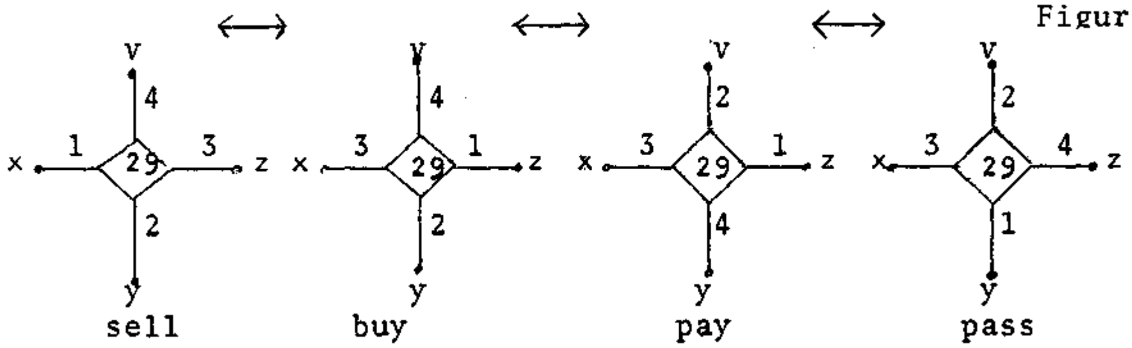


Figure 11

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

Examples of equivalence rules are given in Figure 12.

Figure 12



These graphs represent the meaning rules:

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

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.)

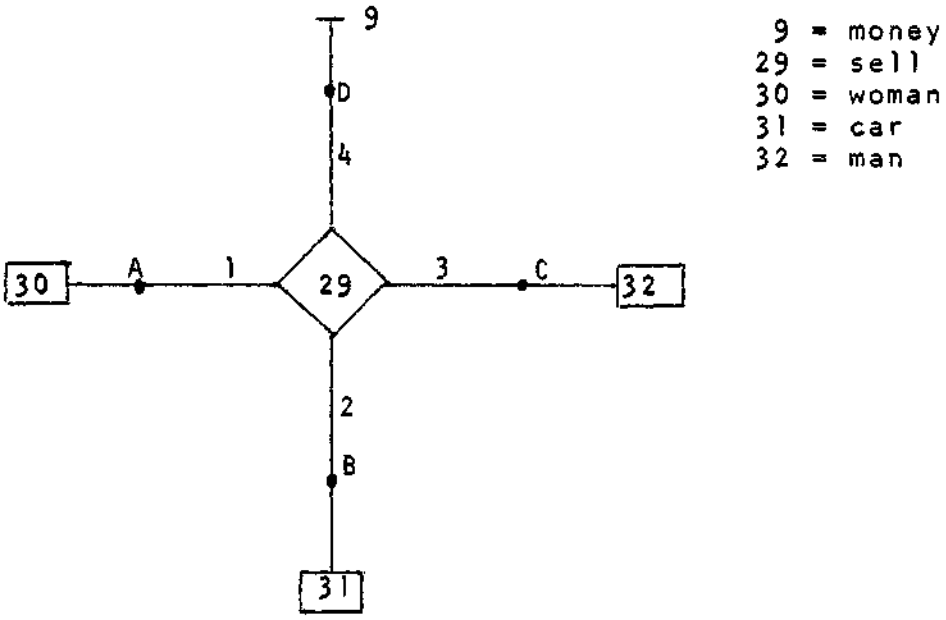


Figure 13

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 *language data processor*; we shall call K a model of comprehension or a device with *rudimentary artificial intelligence*.

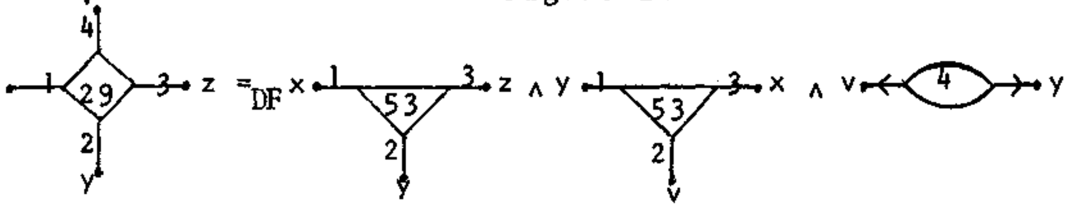
A slightly more intelligent version of K, having generated the DSA-images of SA 20 (or SA 21), will analyze the DSA-images $x \underset{15}{\bullet\!\!\relbar\!\!\relbar\!\!\dashv}$ (and $x \underset{15}{\bullet\!\!\relbar\!\!\not\relbar\!\!\dashv}$) by means of an operation rule; ($\underset{15}{\bullet\!\!\relbar\!\!\relbar\!\!\dashv}$ represents the predicator "sentence"). This operation rule, a subroutine called by $\underset{15}{\bullet\!\!\relbar\!\!\relbar\!\!\dashv}$ (or $\underset{15}{\bullet\!\!\relbar\!\!\not\relbar\!\!\dashv}$), 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 <u>logical and</u>.[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 <u>ihren</u> Anfang mit ...*
   The development began with ...

7. *Der Aufstand nahm <u>seinen</u> Anfang mit ...*
   The revolution began with ...

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

9. *Er schoss einen <u>gewaltigen</u> Bock.*
   He shot a <u>tremendous</u> buck.
   He made a <u>tremendous</u> mistake.

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

11. *Den <u>festen</u> Entschluss fassen, etwas zu tun.*
    To seize the <u>firm</u> decision to do something.
    To decide <u>definitely</u> 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* [*ihrer*$_1$ / *seiner*$_2$] *Raeder*$_3$ *zerbrochen. Wir werden* [*sie*$_1$ / *ihn*$_2$ / *es*$_3$] *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 $\left\{\begin{matrix} \textit{machine's}_1 \\ \textit{engine's}_2 \end{matrix}\right\}$ wheels$_3$ broke. We will send the $\left\{\begin{matrix} \textit{machine}_1 \\ \textit{engine}_2 \\ \textit{wheel}_3 \end{matrix}\right\}$ 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 we 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 one 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 addition, 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, ...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.

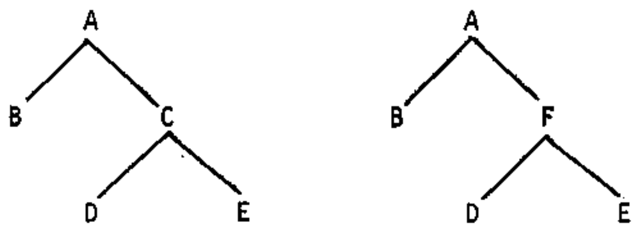
Each rule schema of each grammar consists of a syntactic part and an optional transformational part. For surface and standard grammar the syntactic part of each rule schema consists of context-free rewrite rules. The transformational part contains only transformations whose structural description is 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 $\underset{+up}{\textit{look}}$ *something*.

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.



).

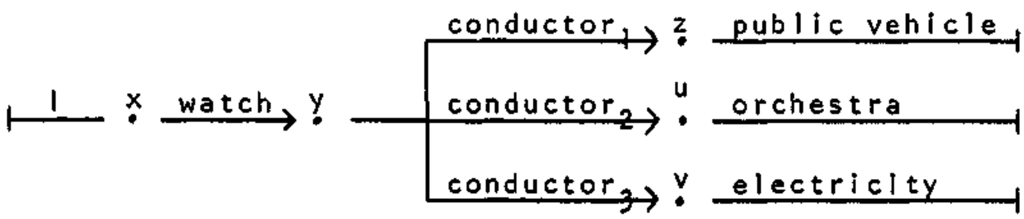
The normal forms of an ambiguous sentence t may be connected by means of "or" links, resulting in one connected normal form. Assume that the normal form of each of the following sentences is represented by the associated graph.

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.

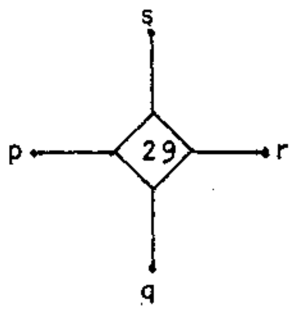


Figure 15

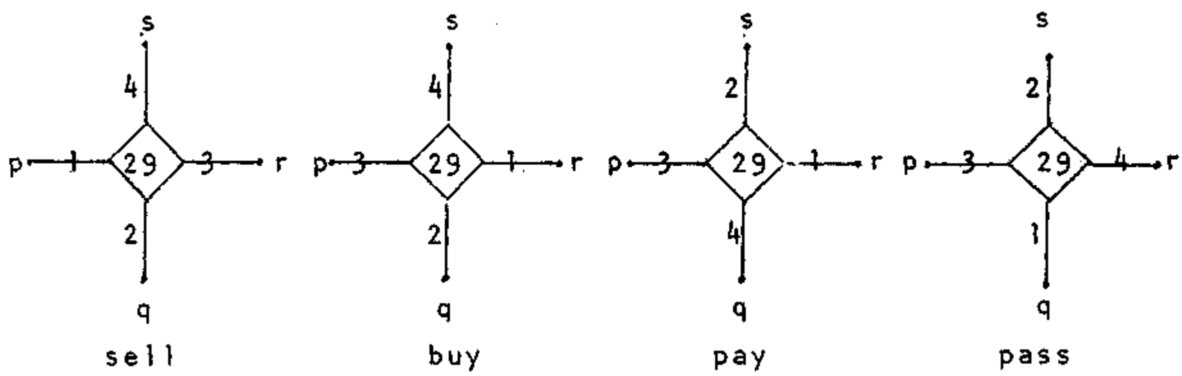


Figure 16

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:

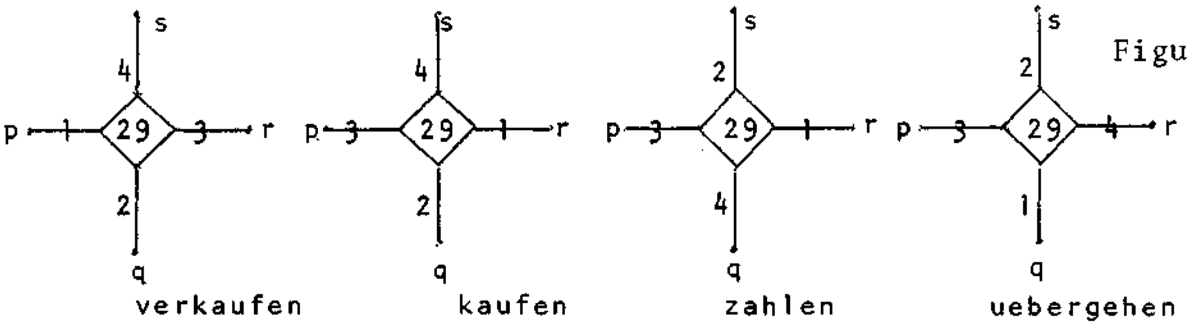


Figure 17

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:

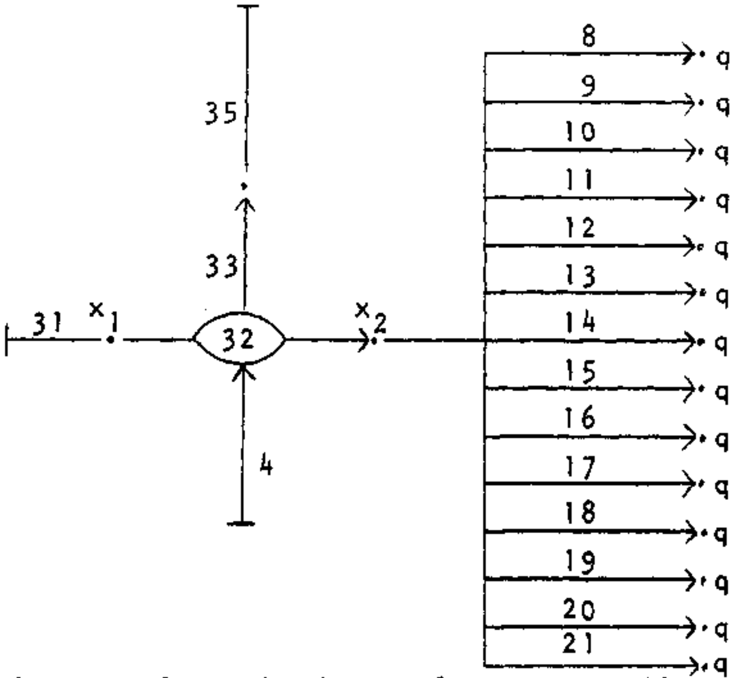


Figure 18

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"):

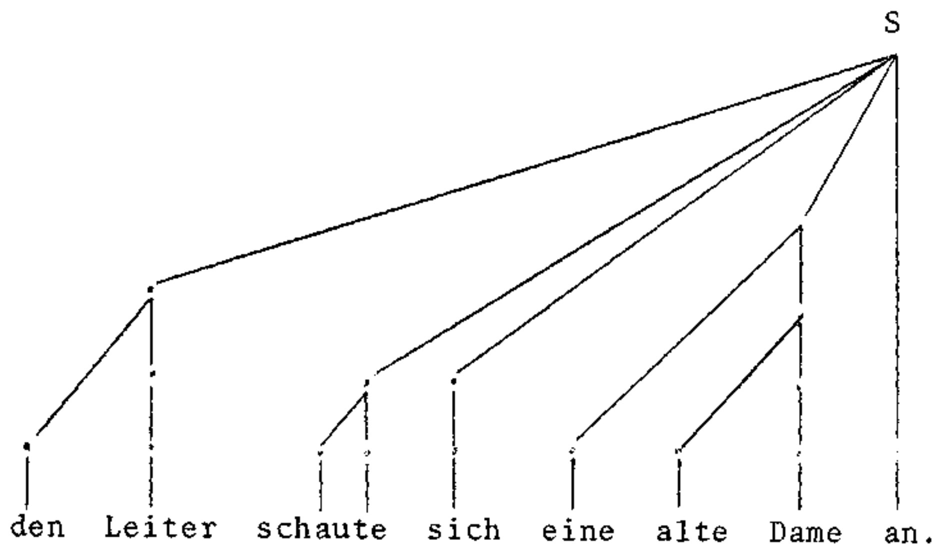


Figure 19

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.

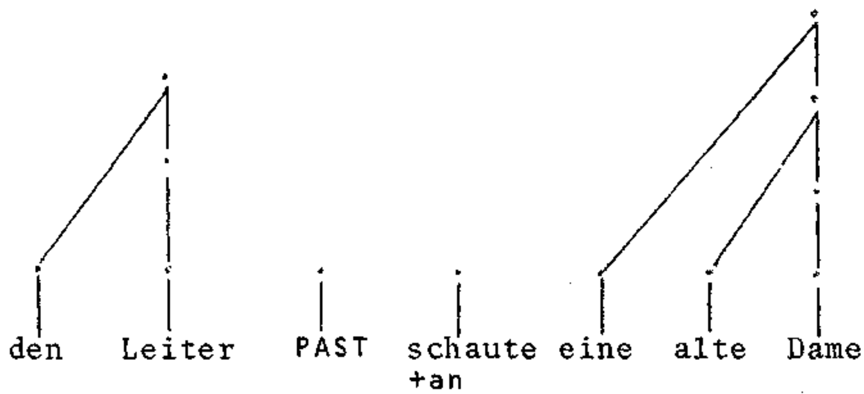


Figure 21

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

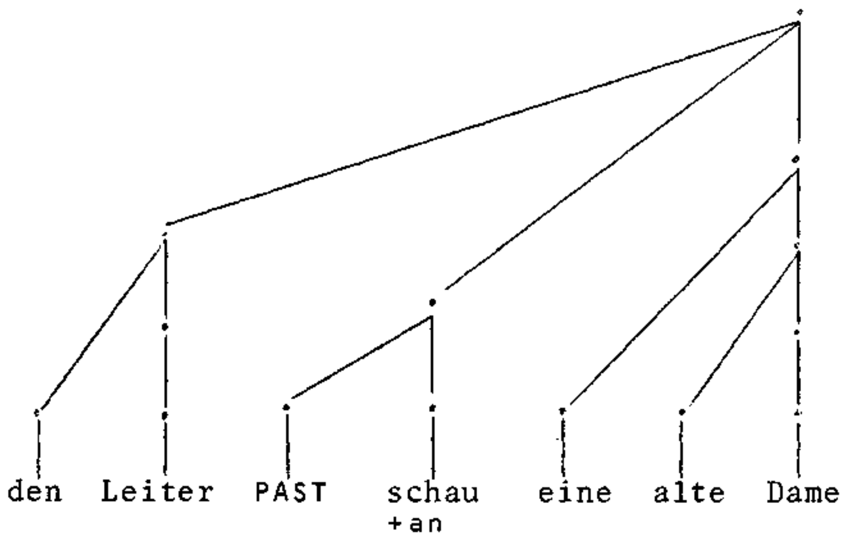


Figure 22

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.

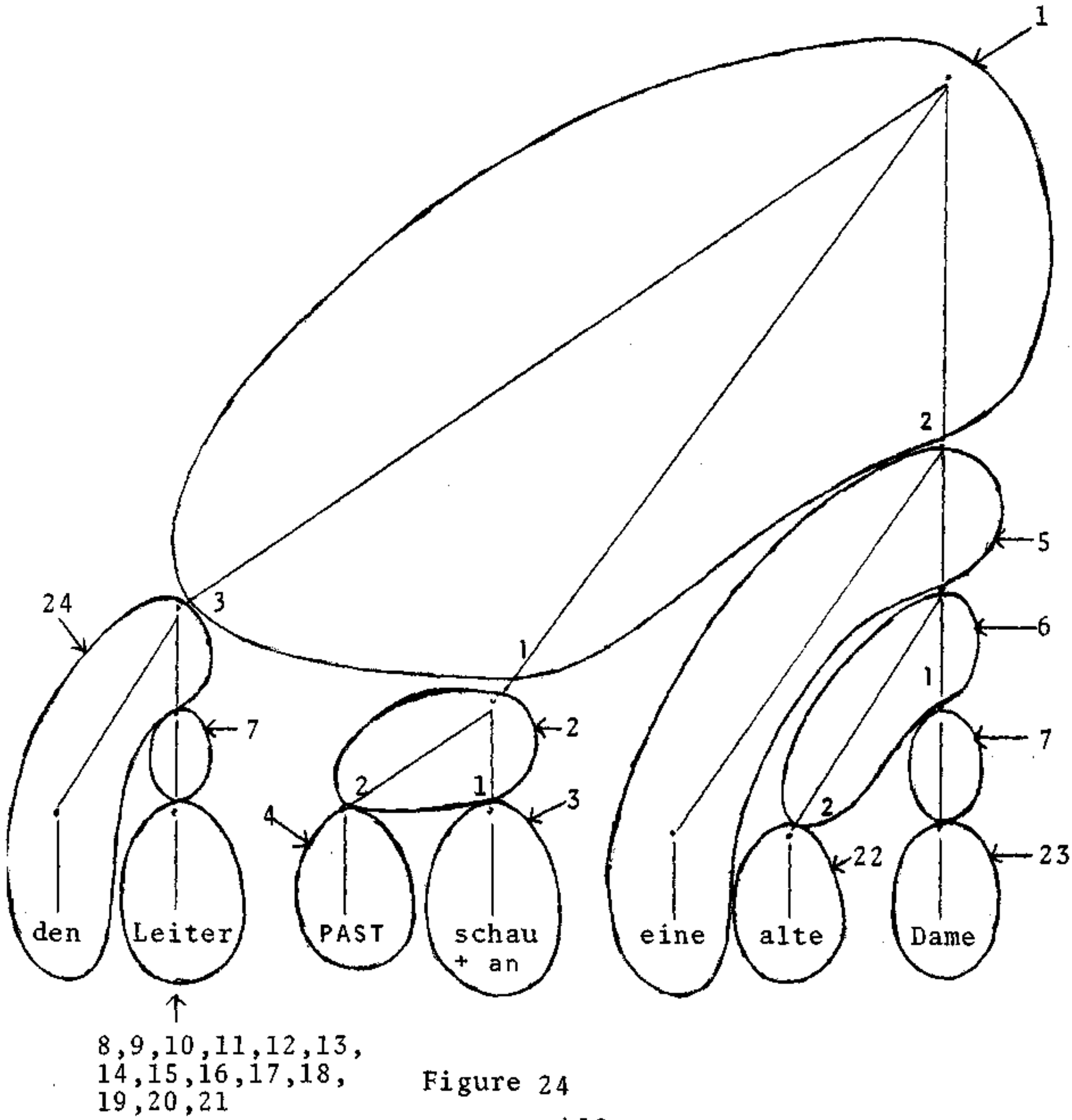


Figure 24

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

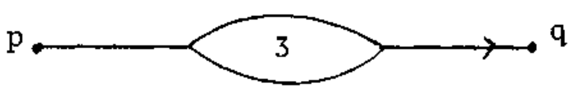
The normal form of a sentence is processed by the DSA-image construction component, which ignores all items which are not of degree 0, or which do not have an operator statement (indicated by light-face print), or which do not have an identifier (indicated by a "+"). For each non-ignored NF expression,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

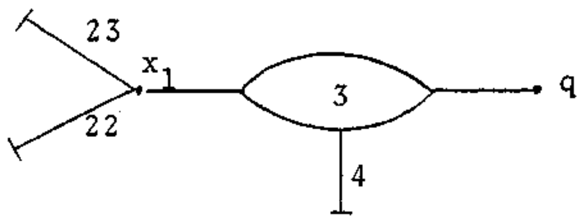


Figure 25

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:

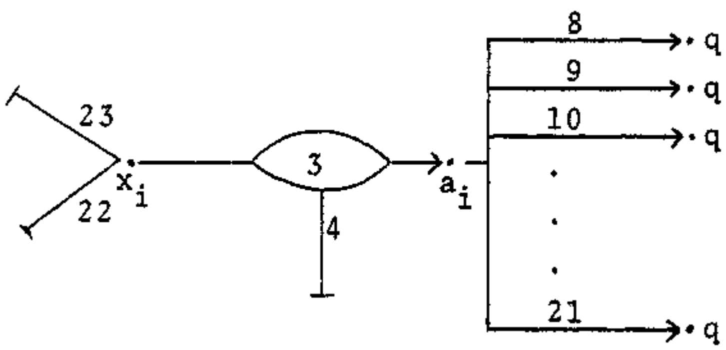


Figure 26

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.

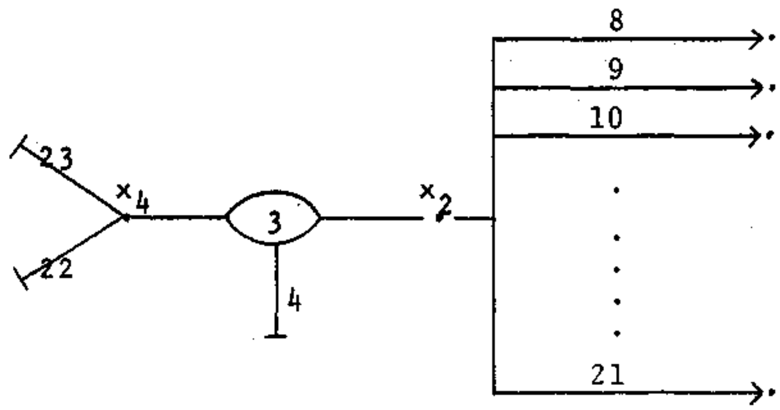


Figure 27

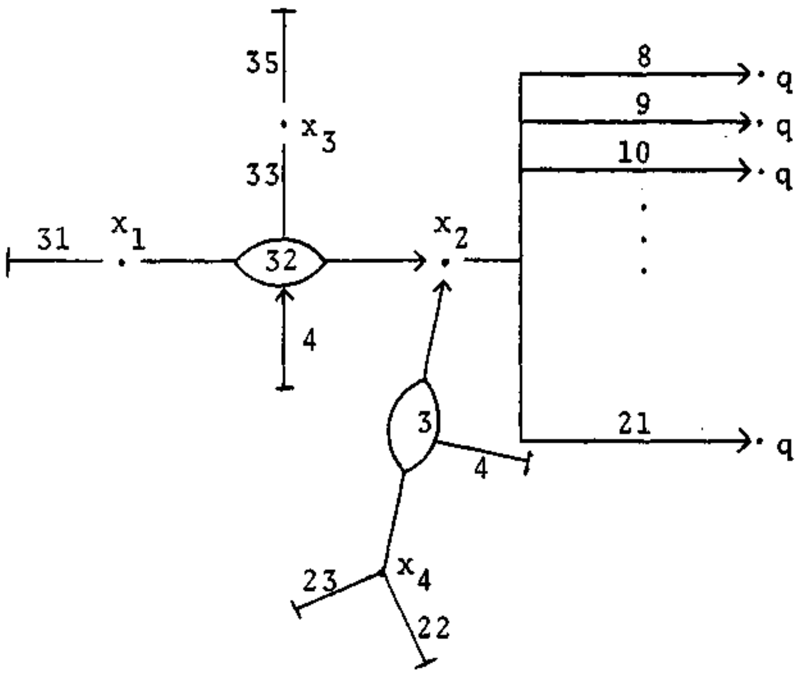


Figure 28

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

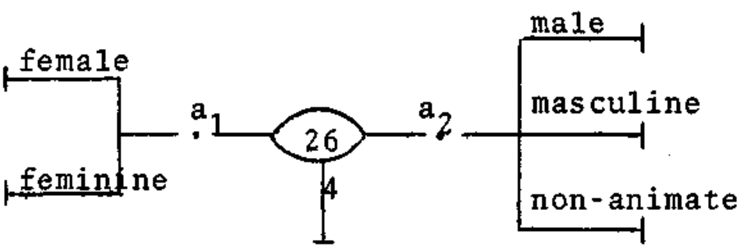


Figure 29

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.

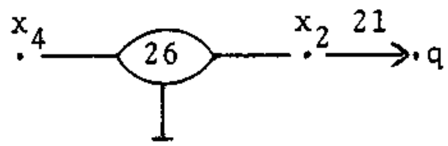


Figure 30

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.

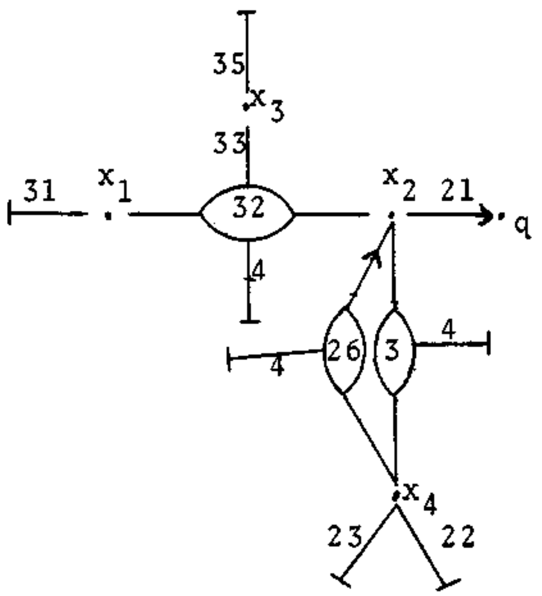


Figure 31

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

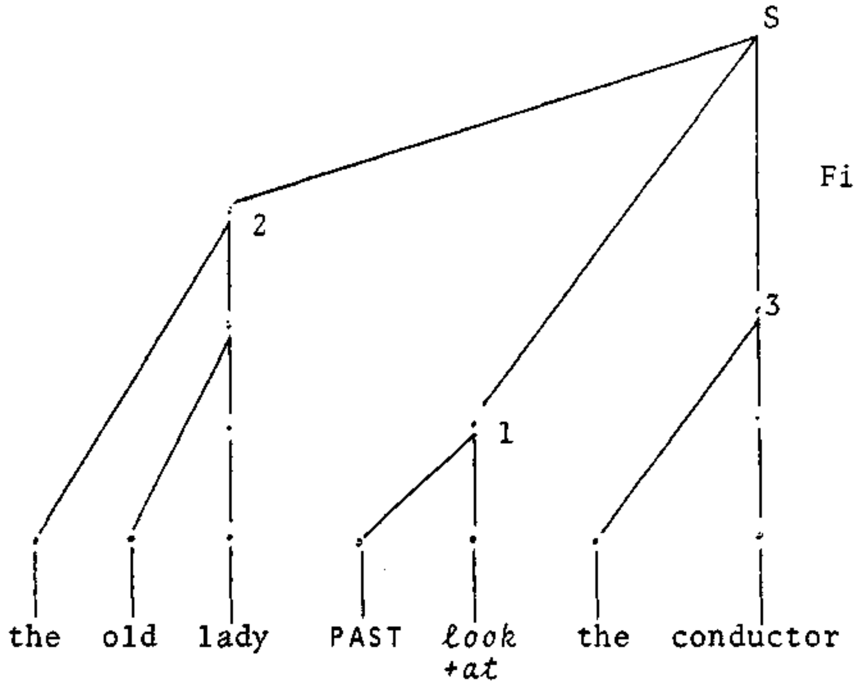


Figure 32

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

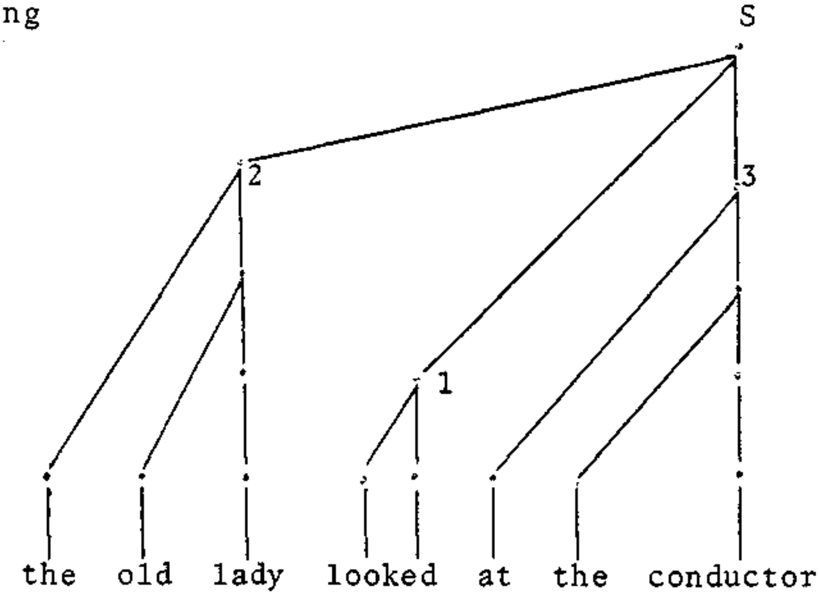


Figure 33

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:

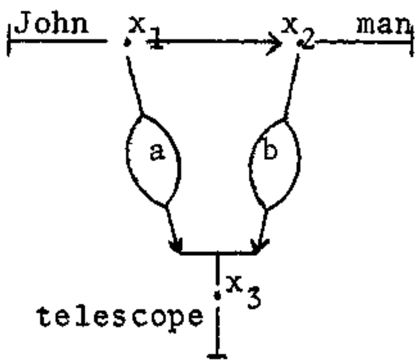


Figure 34[42]

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

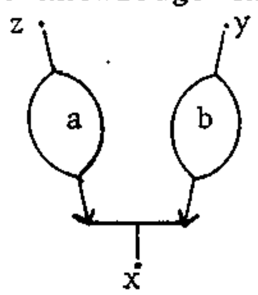


Figure 35

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)

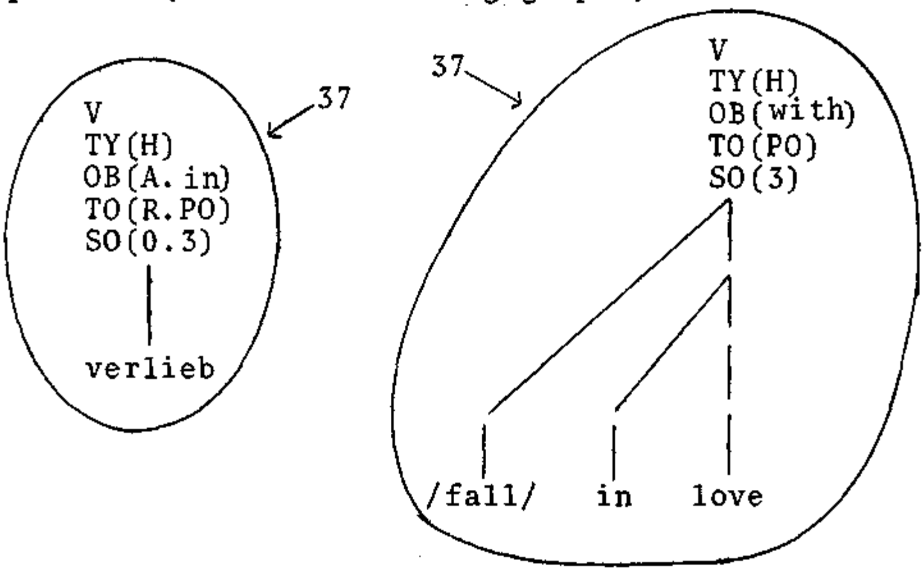


Figure 36

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:

abhaengen von	*depend on*
abhaengig von	*dependent on*
Abhaengigkeit von	*dependency on*

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

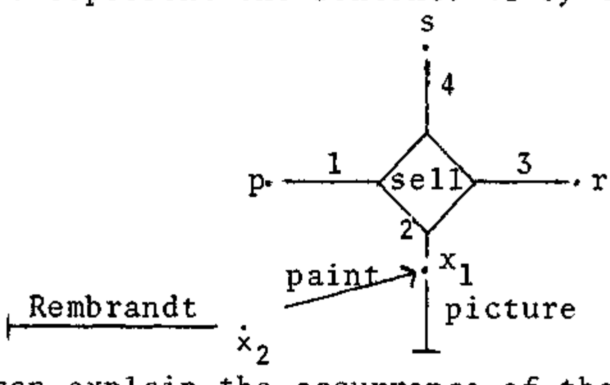


Figure 37

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:

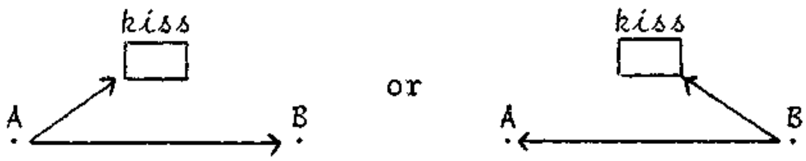
> *A and B kissed. A and B kissed one another. A and B gave kisses to one another. A and B exchanged kisses.*

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



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

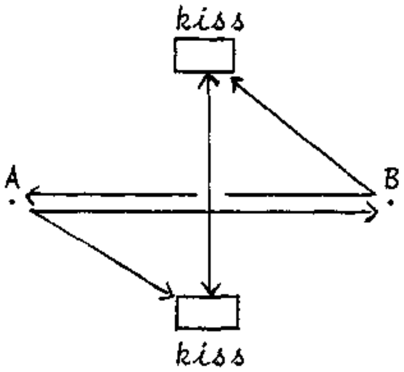


Figure 38

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 understood. 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 (<u>HU</u>, <u>AL</u>, <u>PL</u>, <u>IN</u>, <u>AB</u>, <u>PO</u>, <u>AN</u>, BP, MS, CN, CO, NM, UN, QU, MA, <u>E</u>, 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 (<u>HU</u>, <u>AL</u>, <u>PL</u>, <u>IN</u>, <u>AB</u>, <u>PO</u>, <u>AN</u>, BP, MS, CN, CO, NM, UN, QU, MA, <u>E</u>, P, <u>R</u>, <u>RCC</u>, <u>IT</u>)

RA (TIM, PNC, EXT, SIM, PRI, POST, LOC, DIR, ORN, MAN, MOD, CAUS, MSR, DEG, FRQ, PRB)

OA (DOR)**

<u>Subscript Definitions</u>:

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

<u>Value Definitions</u>:

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 (HU, AL, PL, IN, AB, AN, PO, MA, BP, MS, CN, CO, NM, UN, 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)

Subscript Definitions:

TY = type of noun
OB = object
TO = type of object
TA = takes attribute
SX = sex
RL = relative pronoun
DF = derived from
FM = form

Value Definitions:

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 <u>nach</u> Italien, <u>zu</u> den Indern) }

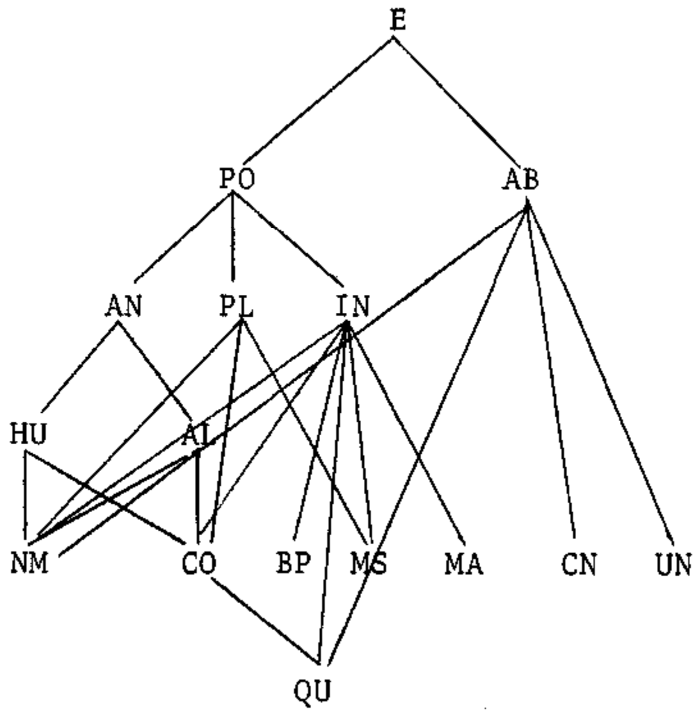
SX { MA = male
FE = female }

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)

<u>Subscript Definitions</u>:

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

<u>Value Definitions</u>:

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 <u>old</u>, five men <u>strong</u>)
	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.