[From: *Overcoming the language barrier*, 3-6 May 1977, vol.1
(München: Verlag Dokumentation, 1977)]

THE SAARBRÜCKEN AUTOMATIC TRANSLATION SYSTEM (SUSY)

H. D. Maas

Sonderforschungsbereich 100
"Elektronische Sprachforschung"

(Electronic linguistic research)

Abstract

The SFB 100 research programme comprises four subsidiary projects which are examining the following problems:

1. **Automatic identification of lemmata in German texts, in other words reducing the word forms appearing in the text to their basic forms by means of a mechanical lexicon containing 100 000 entries.**
2. **The compilation of multilingual dictionaries. The fundamental theory aimed at automatic syntactical/ semantic analysis of sentences, mainly of German sentences, and translation is yet to be fully developed.**
3. **Deep structures in grammar.**
4. **Automatic translation from Russian into German. The models mentioned above are being algorithmed and tested using Russian-German as an example. A French-into-German variant is being worked on. As a private initiative experiments with English into German and Esperanto into German are being carried out.**
5. **Syntactical analysis of French.**
6. **Development of a linguistically orientated programming language.**
7. **Study of old Icelandic legal texts with the aid of a computer.**

All these specific projects are being carried out with a view to making automatic multilingual translation possible along the lines of the Russian into German project. They comprise:

1. **Analysis of the source text with six operators:**
   a) **reading in of the text**
   b) **inflectional analysis and classification of the grammatical data**
   c) **the elimination of most of the ambiguous word forms**
   d) **the breakdown of complex sentences into simple, smaller units**
   e) **detection of noun phrases**
   f) **detection of verb phrases and their structure. Classification of noun and phrase-type verbal complements, transition to the deep structure.**
2. **Translation of source language lemmata into target language lemmata.**
3. **Generation of the synthesized target text, in other words the linearization of the deep structures by the application of the transformational grammar of the target language.**

1. Objectives of the special unit on electronic linguistic research.

The special research unit on electronic linguistic research in Saarbrücken ("SFB - 100" for short) comprises four subsidiary projects all aimed at the syntactical and semantic analysis of texts in various languages by computer methods.

The research programme comprises four subsidiary projects, which are known as A,C,E and F. A brief description of the fields they cover is given below:

The project is pursuing four objectives:

(1) Automatic identification of the lemmata in German texts. The aim is to produce dictionaries(possibly in the form of indexes) for any German text which classifies the basic form ('lemma name') for any given word form. The main problems occur in the automatic analysis of inflection and the elimination of homographs. A syntactical description of the sentences in the text is not required.

(2) Compilation of dictionaries. Automatic identification of lemmata and translation require machine-readable dictionaries containing morphological, syntactical and semantic information on each entry. The subsidiary project in question is investigating what information should be available in lexica of this type and their status in the grammar on which it is based.

(3) Deep grammatical structures. The basic theory for automatic sentence analysis is being established here by determining and formally describing grammatical structures, in particular in German. The model used is a dependent grammar which describes the basic structures and transformations.

Since the results of the analysis are to be used in an automated translation process, some of the work covered contrastive aspects.

(4) Automatic translation of Russian into German. This project represents the test case for the subsidiary projects mentioned above since the theoretical knowledge acquired is transformed to algorithms and applied to Russian-German. This enables the adequacy of the rules of analysis and synthesis to be tested but also shows up problems of interlingual transfer.

Project C is pursuing objectives similar to those of A while also investigating the specific problems of present-day French and is closely allied to translation project A. It is currently being investigated which modifications need to be incorporated into the existing analysis algorithms which were largely designed jointly for Russian and German.

Although it is possible to program algorithms for linguistic purposes using existing programming languages, it is, however, a very awkward and time-consuming process and so subsidiary project E (information science) is working on the definition and implementation of a linguistically orientated programming language which will incorporate the results of experience with SFB - 100.

2. The SUSY translation system

Having outlined the research objectives of SFB-100 I should like to give you a brief description of the SUSY translation system which consists of eight operators, six of which are assigned to analysis, one to transfer and one to synthesis. In addition there are a number of modules for, among other things, dictionary compilation and display of analysis results.

(1) <u>The LESEN operator;</u> The purpose of this programme is to read the text to be translated, the input being

via either punched cards or screen. The words in the text are identified, numbered and tagged with the serial number of the sentence. This programme is not bound to any one language and displays its results in a file (known as X15)

(2) The WOBUSU operator; This operator carries out the analysis of the inflections with the aid of which each word in the text (from file X15) is provided with its basic forms and grammatical data. A word form index is used for this which covers about 50% of the text and generally only contains functional words while the main words (noun, verb, adjective) can be found in a dictionary of roots. There is a special WOBUS0 operator for each language and various possible dictionary structures and methods of inflectional analysis are examined. At present there are WOBUS0 operators for German, English, Esperanto, French and Russian. They all have one thing in common and that is that fixed sequences of word forms which really only represent one word are grouped together as one unit.

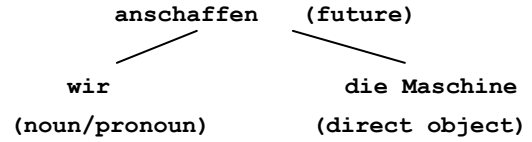
(3) The LEMMAT operator; Syntactical ambiguities within a sentence do not in theory impede successful sentence analysis; if they appear in any great number, the computation time for the analysis of a sentence increases out of all proportion. Ambiguous word classes are thus reduced on the basis of the context with a view to removing every ambiguity as far as that is possible. This simplification procedure is carried out by the LEMMAT operator which is partly controlled by control systems (it is thus not bound to any particular language for the performance of this task) and which partly uses language-dependent subprogrammes. With a view to obtaining multilingual translation a more extensive standardization of the subprogrammes is being attempted

which would enable one-language programme components to be dispensed with as far as possible. At the present time the SFB-100 has LEMMAT operators for German/English and Russian. The Russian operator can be used for Esperanto while the new method is being tested for French, which is controlled by control systems exclusively. We are thus attempting to obtain algorithms which are completely independent of grammar.

(4) The SEGMENT operator: This algorithm breaks down complex sentences into smaller units, which are called segments, and which may be considered as subsentences or parts of subsentences. Algorithm and grammar are separate, the analysis being carried out as automatic push-down. The result shows the coordination and subordination of subsentences and removes discontinuities.

(5) The NOVERA operator; This operator groups consecutive noun-type words into noun phrases. In the process the relationships of the phrases to one another are revealed, in other words sequence and subordination (attributive relationships) are determined. The inventory of each subphrase is recorded, adjectives being understood as predicates which govern noun phrases. The following type of structure can then be obtained from a sentence such as:

'die von uns anzuschaffende Maschine'



It is thus quite easy to generate a number of constructions in the target language in accordance with the syntactical potential of that language:

(a) die von uns anzuschaffende Maschine
(b) die Maschine, die wir anschaffen werden (sollen, müssen)

(c) die Maschine, die von uns angeschafft wird (or werden soll...)

(d) die Maschine, die von uns anzuschaffen ist

(e) die Maschine, von uns anzuschaffen (a structure almost impossible in German) etc.

Conversely, all these constructions are considered as being equivalent at deep structure level for the purposes of analysis and thus they are given a common structure.

The NOVERA operator also prepares the verbal analysis by inventorizing all the verbal words of each subsentence.

NOVERA was designed for Russian and German and now also works in English, Esperanto and French.

(6) The SYNAN Operator; this algorithm has a whole range of tasks to carry out which we can only briefly outline here:

(a) classification of noun groups under a verbal node and determination of the nature of the relationship (subject, object, independent complement).

(b) Classification of subsentences under a verbal node (as subject, object, independent complement) or under a noun node (attributive).

(c) Verbal analysis: identification of the surface constituents, determination of the verbal elements and noun phrases in the deep structure to be reconstructed (where sentences are in a sequence, infinitive constructions etc.)

(d) Determination of the reference of pronouns to noun phrase (carried out only for relative pronouns).

The most problematical aspect of this work is the attempt to reconstruct at deep structure level the elements which have disappeared as a result of cancelled sequences. The synthesis often reveals too much while on the other hand there are examples which could hardly be translated without these reconstructed

elements. The determination of the reference of the pronouns (third person pronouns) is equally difficult and so far no algorithm for this has been produced.

(7) The TRANSFER Operator; The main function of this operator is to search for the target-language equivalents of the lemmata in the source language. A translating dictionary is used for this which lists a target-language lemma for each original lemma and includes target-language data. Currently, work is being carried out with a view to making the selection of translated equivalents take semantic data into account. During the transfer the source-language case is changed into the target-language case. It is assumed that certain transformations are normal (e.g. Russian nominative becomes German nominative). Where deviations from this rule occur the dictionary must be annotated accordingly. The transfer does not include structural changes, only a few parts being common to the two languages. Transfer operators are currently available for Russian-German, English-German and Esperanto-German.

(8) The SYNTHESE Operator: Since our design for automatic translation requires analysis of the source text to be independent of the synthesis of the target text, the job of the Synthese algorithm is to transform the deep structures into surface structures by the application of transformation rules of the target language and then display the latter as an appropriate sequence of work forms (on a printer, a display screen or some type of file).

3. Further development envisaged.

The SFB-100 translation system as it stands enables translations to be produced from Russian, English and Esperanto in German, the translation from English and Esperanto being attributable to private initiative, which thus enables us to draw conclusions with regard

to the scope of application of the system as a whole. This aspect of wide applicability is a very important one in view of the objective of multilingual translation. The success of the trend towards general applicability can be seen for example in the fact that the analysis of French inflection, i.e. the French variant of WOBUSU, was able to be in service within a week and simple French sentences are already being analysed in full.
The next step will be the incorporation of specific French analysis rules together with the compilation of a French-German transfer dictionary. It can be anticipated that within a few weeks the first test French-German translations can be started. In general terms the next point to be examined will be the way in which the synthesis algorithm can be modified and made generally applicable so that it can generate not only German texts but also English and French texts.

4. Scope.
In the light of experience to date the working group engaged on the production of algorithms and programmes for SFB-100 has decided to write all algorithms in a simplified version of the programming language and disregard special features offered by the computer used, a TR 440. An extensive programming library for linguistic applications has been set up which makes programming considerably easier and faster. Frequently used programme modules are available in two versions: the form of a Fortran programme to guarantee compatibility and in the form of a TR 440 assembler programme to save computing time.
It's too soon to know anything about the problems which would arise during the transition to a new computer system, but we are sure that no major obstacles would arise if a new system were used.