[From: *Overcoming the language barrier*, May 1977, vol.2 (München: Verlag Dokumentation, 1978)]

WHERE DOES GETA STAND AT THE BEGINNING OF 1977?

GETA

A group communication presented by

Christian Boitet

Abstract

Automatic translation in Europe certainly encountered a turning point at the beginning of July 1974, with the birth of the Leibniz Group. We shall take this date as the point of departure for explaining the various activities performed since then and the results obtained by the beginning of 1977 in the three main fields involved:

- **study of algorithmic models**
- **implementation of data processing**
- **linguistic utilization**

The studies concerned the models already worked out (ATEF, CETA), *so* as to improve their power and/or their ease of use, and new models. Two of them make it possible to complete the definition of a complete sequence for translation, while the others are the first step towards working out the systems of the next generation.

As regards the computer implementation, the models for transfer and morphological generation have been programmed, and a reorganization of the monitors and part of the software makes it possible to offer the Linguist a complete tool for multilingual translation. In addition, moreover, the study of a new, portable version has been undertaken within the Leibniz Group.

GETA has continued the linguistic activities relating to its Russian-French translation project, together with other related applications, like the analysis of Japanese and a special subset of French. The other applications, of varying size, have been performed outside Grenoble by members of the Leibniz Group and tested at Grenoble for data processing reasons. These are above all analyses (French at Saarbrücken, English at Nancy, Portuguese at Campinas, Italian at Pisa).

1. INTRODUCTION

GETA (Groupe d'Etudes pour la Traduction Automatique) was set up at the end of 1971. Its main objective is the study of automated multilingual translation system for non-literary texts written in natural language. At its beginning, the previous work of CETA (Centre d'Etudes pour la traduction automatique) was available.* CETA, headed by B. Vauquois, had devised an experimentally justified methodology: this relates to the utilization of descriptive levels of language, to the separation between programs and linguistic data written in appropriate meta-languages, and to the search for algorithms general and economic enough for the different processing phases. The experiments, conducted on the analysis of Russian, German and Japanese and on the synthesis of French, reached their culmination between 1967 and 1970, years in which the grammars and dictionaries relating to Russian and French were completed and led to translations of convincing quality of a corpus of Russian scientific articles containing more than 300,000 words,using the IBM 7044 of the University of Grenoble.

---

* presented in (22) .

Nevertheless, these experiments brought to light certain inadequacies of the system. In particular, the above-mentioned methodology had been worked out in the course of the years and could not be completely complied with in all the components of the system. On the other hand, the formal models underlying the two essential phases of syntactical analysis ("increased out-of-context" grammar in normal form) and of "labelling" (changes of structure producing *a* deep representation, or "pivot") proved difficult to handle by the linguists from a certain size of grammar onwards. After all, the "pivot" formalism may have been too ambitious, for, since every trace of surface phenomena was lost at the end of the analysis, it was often necessary to resynthesize information that could have been retained and, a more serious point, it was almost always impossible to translate "bit by bit" a sentence of which only partial analyses were available.

This is why GETA preferred to undertake the study and implementation of new models rather than to transcribe the previous system to the new IBM 360/67 of the University of Grenoble, which would in any case have been neither as simple nor as quick as might have been thought at first, considering the differences between the two computers and the two operating systems.

Determined and written in part by J. Chauché (5),two systems, "ATEF" for morphological analysis and "CETA" for transformation of tree diagrams (usable for both analysis and synthesis), were being completed and had allowed several applications (Russian, Japanese, French) to be undertaken by July 1974.

We shall keep this date as a benchmark, for this was when automatic translation in Europe reached a decisive turning point, with the birth of the Leibniz Group, due principally to the initiative of J.M.Zemb, B. Vauquois, and D. Hérault. Since then, new linguistic applications have been written by several members of the group, and GETA has improved the existing systems, studied and produced new systems so as to have available a complete software for translation, and

embarked on more theoretical studies of "third-generation" systems.

In order to explain the whole of this work, we shall make a distinction between the following three main fields:

- study of algorithmic models
- implementation of data processing
- linguistic utilization

and, for each of these, we shall try to indicate the situation in July 1974, the objectives that have been pursued since then and what has been achieved, together with the prospects at the beginning of 1977, i.e. some time before this conference. Since this article was written at the end of October 1976, we sometimes speak in the present tense of projects merely started on that date but which should soon be completed, without something particularly unforeseen happening, and we certainly intend to give updated information in the oral presentation.

Before embarking on our presentation proper, it may perhaps be useful to define and briefly justify the axes of the above-mentioned methodology.

Levels:

All linguists agree that "levels" should be distinguished in the description of language, though the agreement does not extend either to their definition nor to their content. To be more precise, we might cite Droste (1973): L1 is the grammatical level at which the only concern is the formal properties of the language units: L2 is the conceptual level, which contains a formalized representation of certain relations between the language units resulting from the existence of a universe of reference (the "gnosto-encyclopedic") feature of (15); finally, L3 is the pragmatic level, that of communication in particular situations.

It should not be thought that more usual levels

(phonological, morphological, syntactical) are obtained by a refinement of this classification. In fact the latter relates to the structures of the subject under consideration (language) and the former to the means of realization, (computer independent) . It is therefore not surprising that one can speak of "semantics by syntax" or lead persons astray with a system like ELIZA. Hence the occasional misunderstandings between linguistics and information specialists, the latter considering the means of implementation rather than the linguistic content.

The fact remains that one cannot perfect so complete and adequate a description of a language without dividing up the difficulties. In the case of an automatic translation system, this division has first to be made in accordance with the means (the components of the system) and then with the levels of realization.

Programs and data:

This leads to the second rule: separate the programs from the linguistic data. And this is not gratuitous Cartesianism. For the groups working on the subject cannot be homogeneous (nobody is a specialist in everything!) and linguists and computer specialists have to collaborate on the determination of systems and work separately to implement and use them. Moreover, this makes it possible to partly verify the coherence of the linguistic data before execution. Finally, it is a prerequisite of the search for "universal" algorithms of adequate performance.

Whilst automatic translation programs of the "first generation" failed to recognize this principle, others may perhaps have applied it too strictly to the extent that the algorithm always performs every possible combination. It seems more sensible to include in the linguistic data certain controls on the execution of a general algorithm. This then allows of eliminating certain possibilities without actually calculating them (7)!

Algorithms:

After all, an automatic translation system has cost and mass constraints. For this reason, it is essential to analyse the various operations to be performed on the various possible representations of their data, in order to determine algorithms that are decidable and of minimum complexity. In the CETA system (22) , for example, the system for transforming tree diagrams was undecidable, hence certain undesirable and unpredictable loops.

It may be tempting to use a single formalism (LISP, Qsystems) to write, realize a whole system of automatic translation, this formalism being put directly at the disposal of the linguist. Apart from the fact that these systems always have the computing power of a Turing machine and are therefore undecidable, this results in considerable space and time being lost for the simpler aspects of processing (analysis, morphological generation), to such an extent that none of the trials in this direction has resulted in systems capable of translating voluminous texts.

1. STUDY OF ALGORITHMIC MODELS

We look at the process of translating a text in the following way: a text is a string of characters, on which a morphological analysis is performed. The result is put in the form of a tree diagram, and the phases of syntactical-semantic analysis, of transfer and of syntactical generation are realized by transforming this tree diagram. The tree diagram obtained must finally be transformed into a string, the output text, by a morphological generation phase.

Several other expressions from our "slang" will be used in the following:

1. A variable is defined by a name and a list of special values. The set of its values is:
   - the set of elements of the list and an "empty" value, if it is "exclusive",

- the set of subsets of the preceding set, if it is "non-exclusive",
- the set of the relative integers of absolute value less than or equal to the single element of the list, if the variable is "arithmetic".

For example, one will write GENDER: = (MASC, FEM, NEUTER). On the other hand, consideration is given to "general variables", or supervariables, which group other variables.

2. A mask of variables, (x) is a combination of values of the usable variables (declared to the system). The labels borne by the structures (strings, trees) will in all cases be masks of variables; moreover, the set of variables may change from one processing phase to another. Contrary to too widespread a habit, the labels are therefore complex ones. This allows of avoiding false problems (artificial discontinuities), and to separate the geometrical properties of the structures (a node that is the "ancestor" of another) from their interpretation.

3. A format is a particular and constant mask of variables which has been given a name and which can be used as a reference in dictionaries and grammars.

4. A form is a succession of non-blank characters bracketed by two blanks.

5. A lexical unit is a value of the LU variable, predefined and exclusive. The lexical units are introduced by the dictionaries and not by the declaration of the variables.

6. A labelled tree diagram is one in which each node bears a set of information presented in the form of a mask of variables.

1.1 THE SITUATION IN JULY 1974

By the above date two models, "string-tree" (ATEF) and "tree-tree" (CETA) had been worked out (5), and their

main ideas were to serve as framework for the study of two other models, "tree-tree" (TRANSF) and "tree-string" (SYGMOR), simpler ones intended for completing the algorithmic set-up underlying the translation system being implemented.

Obviously we cannot engage here in a detailed description of these models, for which we refer to the GETA publications listed in the literature. However, we can try to summarize their main characteristics.

1.1.1. ATEF

This model for text analysis with finite state automaton (Analyse de Textes à Etats Finis) realizes a non deterministic finite state automaton by using a pushdown stack. Its external data comprise:

- declarations of variables and formats
- dictionaries (a maximum of six, plus if necessary a dictionary of fixed idioms), in which each article contains a segment, two format names and possibly a lexical unit (LU)
- a grammar, in which each rule comprises a list of calling formats, conditions and actions.

The system successively analyses each form in the text, examining a priori all the possible analyses. Each stage of a particular analysis consists in cutting up a segment into "what is left" of the form to be analysed and in applying one of the rules referenced by the "morphological" format associated with this segment.

The conditions may relate to the results of the analysis of the four preceding forms, to the accessible strings and to the partial results stored by this analysis. It is also possible to store a condition on the result of analysis of the following form. A particular condition consists in giving a list of "subrules" and requiring that at least one of them applies to the result of the current rule.

There are three types of actions: assignment of values to the "C" mask, representing the "current" result of the analysis, transformations of the string remaining for analysis and special functions. These functions make it possible to:

- check the progress of the algorithm by eliminating certain possibilities and by opening or closing certain dictionaries *
- store the current result (case of a compound word)
- create new lexical units using the form processed.
- take into account certain phenomena of linkage between masks (e.g. in the case of idioms)
- decide that a sentence limit has been reached (this does not result from pre-editing).

In the case of an unrecognized form, the system starts analysis again by connecting it with a special format, which in particular makes a call to an obligatory rule, that of "the unknown word". This rule may have subrules, so that one is not restricted to uniform behaviour. For the analytical sense of each form is fixed by the linguist when the external data are compiled.

Formally, the output of this system is a graph whose nodes are the masks (or group of masks for compounds) found and where the vertices indicate the compatibility of the analyses with respect to the grammar. Different presentations of this output, tree diagrams or otherwise, are possible (Qgraphs, tree diagrams with or without "homophrases").

---

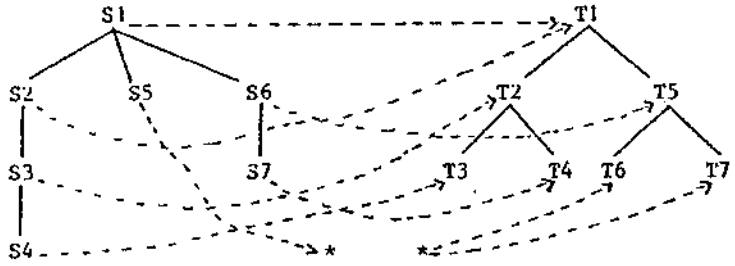
* This is formally done by assigning a value to the special variable DICT.

1.1.2 CETA (5, 6)

This model for control and transduction of tree diagrams (Contrôle Et Transduction d'Arborescences) is based on the notion of transformation due to Gladkij and Mel'tchuk (10) and on the fact that a simple linear writing exists for tree diagrams, for which a given subtree diagram is a sub-string, perhaps discontinuous. It is then possible to identify figure schemata and to make the transformation by means of composed regular pushdown transducers. In total, a tree diagram is processed in a number of steps proportional to its number of nodes.

In theory, but not yet in practice, the system itself allows of describing subtree diagrams specifying or not the order of the "brothers" (the order within a group, or between groups), and if requiring the presence of subordinates at any depth. A transformation is determined by:

- a subtree diagram
- the transformed subtree diagram
- the "transfer function", which allows of transferring into the transform the subordinates of the points of the diagram not present in the diagram
- the assignment of the variables (on the points of the transformed tree diagram).



Let us take an example given by J. Chauché:

```
S1 $B (S2(S3(S4(*)) ; S5, S6 (S7))) / S1 : ETQ -E- A,
S2 : ETQ -E- B, S5 : ETQ -E- B / ETQ (S3) -I- ETQ (S2)
-NE- ETQ0 -ET- ETQ (S4) -E- ETQ (S7)
== T1(T2(T3,T4), T5(T6,T7)) / T1 <— - S1, S2 ;
T2 <—- S3 ; T3.<—- S4 ; T4 <-- S7 ; T5 <-- S6 ;
T6 <--* ; T7 <—-* ; *<—- S5 / T1 : S1 ; T2 : S3 ;
T3 : S4 ; T4 : S7 ; T5 : S6 ;
T6 : S2, ETQ := ETQ(S6) -U- ETQ (S1) ; T1 : S5.
```

Here ETQ is a non-exclusive variable (union and intersection are possible) defined for example by: ETQ= (A,B,C,C1,D). One then has the following schema, in which S2 must precede the group (S5, S6), and in which the masks must meet the Boolean conditions appearing before the ==sign. The transfer function is represented by the dashed arrows in the figure, and the assignment of variables are defined by the last part of the rule. The sign "*" designates the empty node and can be used to request that a node be a "leaf" (in the case of S4) or that it be placed alongside another.

In this system *a* transformational grammar consists of an ordered set of names of such rules. A transformational system is a set of grammars. A grammar can be used in two modes, "unitary" and "exhaustive". In the first way the grammar is applied to a given tree diagram once on the points of the tree diagram. A point is the root of a transformation for a rule of a given grammar if it is the root of a subtree diagram realizing the tree schema of the rule, and if neither itself nor any of its descendants can be the root of a transformation for a rule of inferior rank.

When a point is the root of a transformation, none of its ancestors can therefore be the root of a transformation.

In both modes, an application of the grammar consists in performing all the possible transformations once, with the above restrictions. In the "exhaustive" mode, application of

the grammar is reiterated to the result, with the supplementary restriction that no point that was already the root of a transformation or its descendants can any longer be the root of any transformation. Thus the number of "free" points decreases with each application, and one stops when no rule is applicable any more.

On the other hand, a grammar may comprise "recursive" rules. With the name of such a rule one associates:

- the name of a "recursion" grammar from which to choose the rules participating in the recursion
- the sequence of rules of the grammar mentioned participating in the recursion
- a subtree diagram of the resulting tree diagram (to the right of the == sign) of the recursive rule: this subtree diagram must have a number of points less than the number of points of the schema.

A recursive rule is then applied in the following way:
the rule is applied if it is applicable, then the "recursion" grammar is applied to the result, using only the rules mentioned. The important point is that this grammar is applied to the subtree diagram which root is the point corresponding to the root of the "recursion" subtree diagram. Because of the condition regarding the number of points, the number of "free" points decreases also in this case, and the algorithm stops necessarily.

Indeed, the grammars are organized in a highly flexible way: at the end of each of them appears the list of chaining possibilities, in the form of a sequence of pairs (condition, name of the grammar). Equipped with the chaining relation, the set of these grammars forms an hierarchy, or in other words, it is impossible to "cycle". A condition is a figure schema of exactly the same type as that used in the rules themselves. A reserved symbol, &NUL, makes it possible to stop processing. Chaining is attempted only if the grammar

has been applied at least once.

1.1.3 THE CONSIDERED ORGANIZATION

As CETA does not use dictionaries, it was necessary to modify its conception or to create a new model to realize the "transfer" between two languages. The second solution has been chosen. On the other hand, a tree-string model was lacking for the morphological generation. The organization planned in July 1974 was as follows (the hatched parts represent the software and the others the linguistic data):

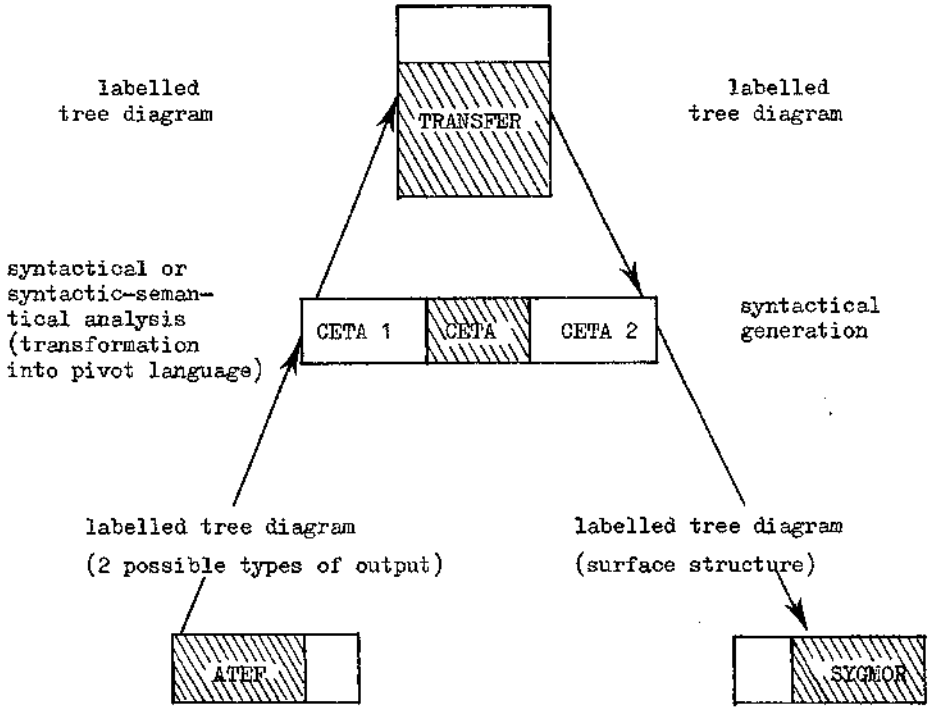


Figure 1

Organization planned in July 1974

1.2 WORKING GUIDELINES

The studies performed since July 1974 may be divided into three groups:

- studies for improving the existing models
- studies aiming to implement new models
- longer-term studies

1.2.1 EXISTING MODELS

1. ATEF

So as to be able to process unknown proper nouns more easily, a new special function, TRANSA, has been introduced. This makes it possible to construct a new lexical unit using what remains to be analysed at a certain point of the analysis, and no longer using the whole form, as in the case of TRANS. This is particularly useful in Russian, in which most proper nouns are declined.

On the other hand, the search for fixed idiomatic phrases has been modified so that priority is given to the longest, which is more natural.

Finally (2), the algorithm has been slightly modified to make it possible during execution to give control to the user in certain conditions, in order to correct an unrecognized form and to reanalyse it, and/or to introduce new items into the dictionaries for the ongoing execution.

2. CETA

Studies have been performed to define "dictionary procedures", so as to fill the gap mentioned above and to process the unfixed idiomatic expressions in a less complex way, as now each requires a rule. However, the implementation would have been too long and delicate on account of the complexity of the available system.

However, it was possible to make the model more flexible by allowing to write "conditional assignments" of the

"if-then-else" type. This study was performed by P.Guillaume and K. Quézel-Ambrunaz (12), who were also responsible for implementation. A modification of this kind, which may seem a minor one, is in fact highly important to the linguists, since it makes it possible to process in only one rule a whole family of discontinuous idiomatic phrases; for instance, the points associated with a phrase will be replaced by a single point bearing a lexical unit associated with the whole phrase so as to facilitate further analysis and the transfer.

1.2.2 NEW MODELS

These are SYGMOR and TRANS, which appear in Fig. 1 without further explanation.

1. SYGMOR (18, 21)

This model is a composition of two transducers: the first, "tree-string", "flattens" the labelled tree diagram of the input, in accordance with a rule chosen by the user, in order to produce a string of masks; the second transforms this string into a string of characters, using external data consisting of:

- the declaration of variables, formats and procedures of condition and assignment
- dictionaries (with direct addressing by the values of certain variables, one at least being referenced by the LU)
- a grammar

Each item in the dictionary is a list of condition/assignment/string triplets, where the last triplet must have an empty condition. For instance * , one will have:

in a basic dictionary:ULHIBOU == /DES-E-PLURX/HIBOU.

in a dictionary of endings: GROUP1 ==ISP3PP/ / ENT;
MODE-E-PART-ET-TPS-E-PASS//E≠≠;

---

x The examples in this section have been taken from B.Thouin (1975)

where ISP3PP is a condition defined by: ISP3PP:=
= PERS-E-3-ET-NBR-E-PLUR-
ET-TPS-E-PRES-ET-MODE-
DANS-IND-U-SUBJT.

At any point during processing, certain quantities are accessible and are referenced in the metalanguage by special symbols:

- 1, 2,... n are the numbers of the dictionaries($n \leq 8$).
- C and P are the"current" and "preceding" masks.
- T and S are two strings, T being the one currently processed and S the one previously output. They are also the names of two associated masks.
- G,D and M are three origins, left,right and middle, in T, and define the points where a string can be inserted (given by consulting a dictionary).

A rule of grammar comprises:

- a condition of application relating to the masks and accessible strings
- a part for reading one or more dictionaries and manipulating the accessible strings
- a part in which one can assign values to the C,P,T masks using the dictionary, and to the P,S, masks using C and T,
- a "string transformation" part to realize replacements of substrings in T
- finally, a part indicating a set of possible continuations in the grammar after application of the rule.

Hence, for example:

The rules between parentheses are optional in the indicated continuation. It should be noted that, unlike ATEF, SYGMOR realizes a finite-state deterministic automaton thus reflecting the lesser complexity of the synthesis process. To process a mask, SYGMOR thus searches for the first applicable rule (at least one rule must have an empty condition),

applies it and follows the continuation indicated until it finds an inapplicable obligatory rule.

2. TRANSFER

We mentioned the problem of transfer earlier: it is necessary at the same time to allow of consultation of dictionary and transformations of structure. For the "multi-lingual" philosophy of the system makes it necessary, in order to benefit benefit from it, to arrive at "pivot" structures of such a kind that one does not have a system of syntactico-morphological generation for each pair of languages but for each language. The problem has received a preliminary solution, consisting in making the transfer in two phases:

- TRANSF: consultation of the dictionary, allowing possible to transform each point of the "source" tree diagram into a "target"subtree diagram. This dictionary is directly addressed by the source LU.
- CETA2; transformations by a CETA phase to attain the pivot form.

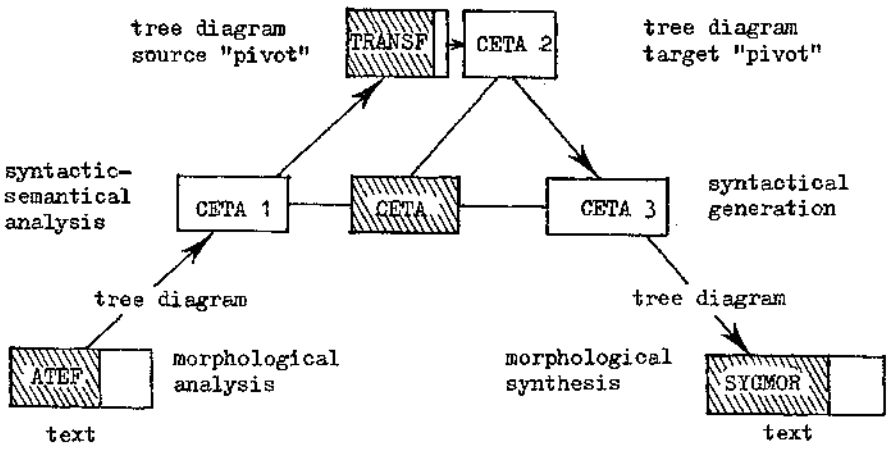


Fig. 2. Current organization

TRANSF (17) uses the following external data:

- declarations of variables of assignment formats, and of condition and assignment procedures (this is the only model which uses at the same time two sets of variables, "source" and "target")
- transfer dictionary, in which each item associates to a source LU a list of condition/target diagram/ assignment triplets, the last triplet having an empty condition. Here is an example taken from a version of the Russian-French transfer:

```
"DRUG" ==$SN5J/1(2) / 1 :'AUTRE'; 2: 'UN' /
                /     / 'AMI'.
```

N5J is a condition procedure, relating to the LU source mask 'DRUG', and 1 (2) represents a diagram with two points. In the case where the node doesn't come from the Russian expression for "one and the other", N5J is not verified and TRANSF substitutes for the source node a target node where the LU is 'AMI' and where, in this case, the other variables are empty if they are special to the target or equal to their old values if they have been declared common to the source and the target.

It may be remarked that this rather simple model would make it possible to raise the level of analysis by bringing about "semantic expansions" so as to allow removal of ambiguities at a "referential" semantic level (2) , using a dictionary of the previous type as data base. However, it is not planned to implement this idea in the immediate future.

1.2.3 LONGER TERM STUDIES

These are mainly located at two levels, that of the general organization of a translation system and that of the models that may constitute it. However, they all fit into a longer-term project aimed at creating a "third-generation" multilingual translation system in a high-level language.

A system of this kind would in our opinion have several interesting characteristics:

- generalized use of heuristic methods in each "module" of the system (as is already the case) and at the level of their chaining.
- introduction of "self-adaptation" properties into certain modules (e.g. grammars with modifiable weight so as to allow of a certain "tuning" during processing.
- definition of a metalanguage allowing to call the various modules and thus to write a whole class of "monitors" corresponding to different translation strategies.
- use of data bases to remove the ambiguities resisting purely syntactical processing.
- (non-constraining)use of the possibilities presented by conversational operating systems as regards the introduction and debugging of linguistic data as well as the intervention by the user during processing (cf.1.2.1) and the revision of the result.
- use of more general working structures than the tree diagrams and common to all the modules, for instance networks analogous to Qgraphs, with a view to greater homogeneity, and starting from greater flexibility.

Some of these ideas have been developed in (2); all should contribute to the determination of the new "third-generation" system.

1.3 CONCLUSION: STATUS OF THE WORK AT THE BEGINNING OF 1977

This may be concisely summed up by saying that the objectives explained in 1.2.1 and 1.2.2 have been attained, and that it has been possible to test the linguistic adequacy of the underlying ideas by daily use on numerous languages. It is moreover clear that certain inadequacies have been noted and that efforts will be made to remedy them in the future system.

As for the longer-term studies, the detailed determination of the new system has started, and we shall no doubt be able to specify the progress made during the oral presentation.

2. IMPLEMENTATION

2.1 THE SITUATION IN JULY 1974

During the whole period under consideration we worked under the CP/CMS system of virtual machines available on the IBM 360/67 of the University of Grenoble. Our software may in principle be broken down in the following way:

- utility programs
- compilers for the metalanguages of the various modules
- programs for execution of the various modules
- trace programs for the various modules
- loading routine(s) for the programs and the assembled data
- supervising routine(s)
- guiding programs enabling the linguists to use the above tools without knowing anything more than the conversational edit program EDIT of CMS.

In July 1974 the software for the ATEF and CETA systems was available. The guiding programs were written in EXEC 360 and all the rest in assembler 360. The software for CETA, which was very recent, was not yet very reliable. On the other hand, that of ATEF had already been debugged during a fairly long period and was functioning without major hitches.

2.2 THE OBJECTIVES PURSUED

There were four of these: creation of the new modules SYGMOR and TRANSF, perfecting the data processing sequence at all levels (cf. Fig 2), improving ATEF and CETA and moving towards greater portability.

2.2.1 NEW MODULES

The programming of SYGMOR, entrusted to T. Thouin, assisted by N. Adamopoulos, was done in EXEC for the guiding program and in PL/1 for the rest, an innovation which moreover fits in the framework of the longer-term project mentioned above. This programming and the studies made for it have formed the subject of several publications (1,18,19,20/21). It came into use at the end of 1976.

The study, programming and insertion of TRANSF, entrusted to M. Quézel-Ambrunaz, were terminated more quickly, since the whole system, up to and including the CETA 3 phase, was ready for use in the spring of 1976.

2.2.2. PERFECTING THE PROCESSING SEQUENCE

The report programs loading and executive routines available in July 1974 did not allow of going any further than the TRANSF phase. It was necessary to remodel these programs quite considerably to obtain a complete software, the first part attending to analysis, transfer and syntactical synthesis, and the second dealing with morphological generation. Some idea of the complexity of such a system may be given by mentioning a few details:

- as far as possible, work is done in central memory during the execution. Input-outputs are limited to the traces and the possible outputs of intermediate results intended for further treatment of the same texts. It is therefore necessary to load all the necessary linguistic data and to have access to them at the right moment (case of CETA, used several times).
- one "source language" * can derive towards several target languages and one "target language", and con-

---

* The terms "source language" and "target language" are shorthand for "the linguistic data associated with the analysis/synthesis of a language, known to the system through a special 'language code' ".

versely one "target" must be usable while coming from several "sources". Hence problems of independence at compilation time and of correspondence at execution time, to mention only those.

- the linguistic data can be shared among several virtual machines so as to allow several linguists to work at the same time.

2.2.3 IMPROVEMENTS TO ATEF AND CETA

As regards ATEF, the modifications relating to interaction with the user at execution time (2) were completed relatively quickly, using assembler 360 and a structured language of the same level. PL360. For a long time they remained experimental before being integrated with the system normally used.

As regards CETA, the departure of J. Chauché and the complete absence of documentation on his programs (execution) allowed of only partial completion. However, it has proved possible for P. Guillaume and M. Quézal-Ambrunaz to program the principal improvement, consisting in permitting conditional assignments sufficiently independent of the execution programs. It should be available at the beginning of 1977. However, we should point out that this system, in its present state, has proved capable of use for numerous linguistic applications (cf.111), even if certain of its theoretical possibilities are not completely implemented.

2.2.4 TOWARDS GREATER PORTABILITY

Since computers are outdated roughly speaking every 10 years, we had to envisage reconversion to more recent equipment. The first objective was to adapt our software to the VM/CMS 370 system running on IBM 370/158 and upwards, the first trials made it possible to give a demonstration during the symposium organized in September 1976 by the SFB 100 at

the University of the Saarland, Saarbrücken.

The second objective was to study more closely the high-level languages capable of being used in a fairly easy way, and the choice has fallen on PL/1, in which SYGMOR is written. More precisely, it is a subset of PL/1 determined in collaboration with other members of the Leibniz Group to be implemented on the maximum of installations accessible to the Group, and moreover, to be sufficiently restrictive in order to prohibit certain over-expensive programming techniques.

2.3 CONCLUSION; STATUS AT THE BEGINNING OF 1977

Complete software for multilingual translation is available under CP/CMS 360, and minor modifications to certain utility programs and to the guiding programs enable it to be transferred under VM/CMS 370. Moreover, a demonstration has been suggested to the EC if this is possible on the latter's equipment.

As we shall see in the following part, this software has been used by several teams of linguists within the Leibniz Group for working on several languages by developing linguistic data of considerable volume and performing trials first on artificial corpuses and then on real ones.

3. LINGUISTIC UTILIZATION

3. 1 THE SITUATION AT THE BEGINNING OF 1974

The only linguists using the GETA software were those of the group. The main project was the production of a Russian-French translation system, and consequently a complete "right-left" morphological analysis had been written and tested on one of our corpuses ("A 12"). The corresponding dictionaries contained about 1600 LU, and the grammar nearly 150 rules. Moreover, more voluminous dictionaries, containing more than 10,000 LU (of which 8500 for the verbs), had been compiled for use after the debugging phase. Finally, writing of the syntactical analysis was beginning (CETA 1 phase).

On the other hand, the writing of morphological analyses of French, Portuguese, German and Japanese had been undertaken. In the last case the problem was much more complex, since Japanese words are not separated by spaces and a "form" no longer corresponds to a word but to a whole sentence. The corpus of Japanese had been supplied by Prof. Ishiwata of the NLRI, Tokyo, in the form of a magnetic tape on which all the characters (kanjis, katakanas, iraganas and Latin ones) are coded on one or more bytes.

3.2 THE WORK UNDERTAKEN SINCE

3.2.1 GENERAL

As stated at the beginning of this paper, the Leibniz Group came into being in Grenoble in July 1974. It was an occasion to group the efforts of several groups working on automatic translation in Europe and Canada. Most of them, unlike GETA, are more oriented to linguistics than towards the creation of data processing models. And, though our systems are conceived from a multilingual point of view, we have only really used them for a full-scale application within the Russian-French project.

Cooperation with other groups therefore presented a mutual interest, since it allowed us the test of number and gave our partners access to software whose utilization demands only knowledge of the metalanguages and of the operating principles of the models, without presuming any training in data processing.

From the linguistic point of view, confrontation with the practical multilingual problem has led to fruitful collaboration, culminating at the beginning of 1975 (meetings in Pisa and Lugano) in the determination of a "pivot" formalism (11) in which the results of analysis are supplied to the transfer and those of the transfer (for us, the output of CETA 2) to synthesis. This is a language of labelled tree diagrams equipped with a formal syntax that utilizes a certain number of common "pivot" variables. Its principal interest derives from the fact that the structure itself is merely a bracketing and that the significant information (syntactical class, syntactical function, logical relation etc.) is given in the masks borne by the nodes. Thus one and the same graph may represent different levels in the analysis of a statement, which makes it possible no longer to have to practise an "all or nothing" policy, but on the contrary to use in each case the most profound result obtained by the analysis.

3.2.2 THE FAIRLY LARGE-SCALE PROJECTS

These are projects relating to Russian, French, English and Portuguese.

1. Russian

The work has consisted in recasting the morphological analysis and in writing the syntactical analysis and the transfer. For strategic reasons of utilization of the ATEF system, preference has been given to rewriting the analysis in a left-right direction. At the end of 1976, the grammar obtained comprised no more than a hundred or so rules. The syntactical analysis has passed through several successive versions. The status of this work in spring 1976 has

incidentally been presented by N.Nedobejkine in (16).

Finally, transfer occupied the second semester of 1976, at the end of which the dictionary comprised about 2000 Russian LU. We may recall that an LU may cover several roots and their derivatives (loi, légal, légaux,légaliser etc.) .

2. French

Several projects have been undertaken on this language: three analyses and one synthesis. The first two analyses were performed by the SFB 100/C of Saarbrücken, in collaboration with B. Vauquois and D. Froment of GETA. That by J. Weissenborn and M. Belin (23) (morphological and syntactical) is meant for integration in a translation system, while that by E. Stegentritt (morphological only) is oriented more to the implementation of linguistic hypotheses, without any great concern with effectiveness or integration.

The third analysis and the synthesis have been undertaken by GETA. The analysis (morphological and syntactical) is intended for integration in a system for graphic manipulation in natural language. (9),created in collaboration with the Artificial Intelligence Group of the University of Aix-Luminy. When B. de la Fayolle's thesis was published, there were about 300 LU, some thirty ATEF rules and 120 CETA rules, tested on a small corpus of graphic instructions in French. For the needs of this project, it was of course necessary to determine an appropriate "pivot" used as input to the graphic system proper (13).

The synthesis was undertaken at the end of 1976 within the framework of the projects, and is therefore too recent for significant figures to be produced.

3. English

This is a morphological and syntactical analysis performed by the Automatic Translation Group of the University of Nancy and undertaken at the end of 1974. It was able to make quite rapid progress, since this group already had great experience with analysis of English and therefore merely had to adapt its presentation to our equipment. By the end of 1976 the morphology had been finished and the syntactical analysis was beginning to give encouraging results.

4. Portuguese

A morphological and syntactical analysis has been undertaken by P. Daun Fraga of the University of Campinas (Brazil) in collaboration with B. Vauquois. In a few quite short stays in Grenoble, its author succeeded in constructing an analysis comparable in volume and in performance to that of (23) on French. It should be noted that the University of Campinas is also a member of the Leibniz Group.

3.2.3. SMALLER-SCALE PROJECTS

First of all there is an analysis of Italian, a long-term undertaking by the CNUCE of the University of Pisa, and using ATEF and CETA. For various reasons this project has made less progress than expected.

Other, more marginal applications have related to the morphological analysis of German, Polish and Quetchua.

3.3. CONCLUSION: STATUS AT THE BEGINNING OF 1977

1. Translations

The first results of the Russian-French project should appear at the beginning of 1977, using the analysis, transfer and synthesis devised by GETA.

Likewise within the group, work on Russian-English has been started by P.Paul of the University of Melbourne and

on Portuguese-French by P.Daun Fraga. Finally, a French-English project is under study, in cooperation with the Centre du Documentation of the CNRS.

2. On the various languages

The analysis of Russian and the Russian-French transfer have been fairly stable since the autumn of 1976. Only the experimental dictionary has grown somewhat (about 2000 LU) and the CETA 2 phase has been completed.

The analysis of French presented in (23) has been reworked and completed. It comprises about 350 formats and 70 rules in ATEF, 130 rules in CETA, with a dictionary of 1200 LU, and has been developed on a text of approx.2500 words.

The analyses by B. de la Fayolle and E. Stegentritt (cf. 3.2.) have remained stable.

The morphological analysis of Japanese (14),from right to left, used about 200 formats and 50 rules, and has been produced on a text taken from the corpus supplied by Professor Ishiwata. The dictionaries comprise 400 LU.

As for the other languages, we shall have sufficient information by the oral presentation.

CONCLUSION

Of course, we have been able to present only a quick survey of the various activities of our group. We should like to stress the fact that, in the linguistic applications, the part played by the information specialists of the group is a purely informative one: that is to say, the essential work is done by the linguists. Thus the applications performed in collaboration with other members of the Leibniz Group are the fruit of their work and not of ours.

Moreover, by way of conclusion, we may return to the triple relationship of the field of automatic translation to algorithmic theory, data processing and linguistics; whilst the first researchers may have thought that the decreasing order of difficulty was data processing-linguistics-algorithmic theory, the present tendency would instead be to say: Linguistics-algorithmic theory-data processing. For linguistic phenomena beyond our control have proved as difficult to grasp as to formalize. Moreover, their instability (if one considers several corpora) induces one to conceive of algorithmic models of greater flexibility so as to be more adequate.

REFERENCES

Sous la forme:

(1) N.ADAMOPOULOS (1975)
Contribution a l'ecriture d'un compilateur pour un dictionnaire dans un système de génération morphologique de langues naturelles.
Rapport de stage, GETA, Grenoble,1975.

(2) Ch.BOITET (1976 a)
Un essai de réponse à quelques questions théoriques et pratiques liées à la traduction automatique. Définition d'un système prototype.
Thèse d'Etat, Grenoble, avril 1976

(3) Ch.BOITET (1976 b)
Problemes actuels en traduction automatique. Un essai de réponse.
COLING-76, Ottawa, Preprint N°33, juillet 1976

(4) Ch.BOITET (1976 c)
Méthodes sémantiques en TA.
TA information n° 1, 1976, 3-42

(5) J. CHAUCHE (1974)
Transducteurs et arborescences.(Etudes et réalisation de systèmes appliqués aux grammaires transformationnelles).
Thèse d'Etat, Grenoble, 1974

(6) J. CHAUCHE(1975)
Présentation du système CETA
Document G-3100-A, GETA, Grenoble, 1975

(7) J. CHAUCHE, P. GUILLAUME, M.QUEZEL-AMBRUNAZ (1973)
Le système ATEF
Document GETA, G-2500-A, Grenoble, 1973

(8) F. G. DROSTE (1973)
Model theory, logic and linguistics.
Linguistics, n° 105, june 73, 5-34

(9) B. DE LA FAYOLLE (1976)
Analyse du Français comme langage de commande dans un système de construction graphique.
Thèse. 3ième cycle, Grenoble, 16 novembre 1976

(10) A.V.GLADKIJ et I.A.MEL'TCHUK
Tree grammars
Septembre 1969, Congrès ICCL,AL 32, Sanga Säby, Sweden

(11) Groupe LEIBNIZ
Projet de représentation des structures de phrase au niveau du transfert.
Lugano, Mars 1975

(12) P.GUILLAUME et M.QUEZEL-AMBRUNAZ (1976)
Etude et réalisation de l'insertion d'affectations conditionnelles dans le système CETA
GETA, document interne, décembre 1976

(13) LALOUM (1977)
Thèse 3ième cycle, Grenoble, à paraître

(14) A. LAURENT (1977)
Analyse morphologique du japonais au moyen du système ATEF.
Thèse de 3ième cycle, Grenoble, à paraître

(15) I.A. MEL'TCHUK et O.S.KULAGINA (1967)
AT: Some theoretical aspects and the design of a translation system.
In "Machine Translation".137-173, A.D.Booth, ed. North-Holland, 1967

(16) N.NEDOBEJKINE, L.TORRE, M.AXTMEYER, J.GAUDEY (1976)
Données linguistiques de l'analyse automatique de russe.
Congrès COLING-76, Ottawa, juillet 1976

(17) M.QUEZEL-AMBRUNAZ (1976)
Le système TRANSF
GETA, document interne,juin 1976

(18) B.THOUIN (1975)
Systèmes informatiques pour l'analyse et la génération de langues naturelles
DEA d'informatique, Grenoble, septembre 1975

(19) B.THOUIN (1976 a)
Présentation et utilisation du système SYGMOR
Document GETA, 1976, à paraître

(20) B. THOUIN (1976 b)
SYGMOR : un système informatique pour la génération morphologique de langues naturelles.
AJCL, 1976, à paraître

(21) B. THOUIN (1977)
Thése de 3ième cycle, Grenoble, 1977, à paraître

(22) B. VAUQUOIS (1975)
La traduction automatique à Grenoble
Document de linguistique quantitative, n° 24, Dunod, 1975, 184 p.

(23) J. WEISSENBORN et M. BELIN (1976)
Arbeiten zur automatischen Analyse des Französischen SFB 100/C, Universität des Saarlandes, Saarbrücken, Sept. 1976, 200p.