Chapter 13: Transfer systems since 1970

13. 1: University of Montreal, TAUM (1965-1981)

Research on MT in Canada began in 1965 with the sponsorship by the Canadian Research Council of projects at the University of Saskatchewan, at the Cambridge Language Research Unit, and at the University of Montreal. The Sakatchewan project concentrated mainly on a statistical approach to a 'direct translation' English-French system (Ch.12.5), and the CLRU project investigated the possibilities of an interactive approach (Ch.5.2 above).

At Montreal, the Canadian Research Council set up CETADOL (Centre de Traitement Automatisé des Données Linguistiques) in November 1965, under the direction of Guy Rondeau. MT research began slowly; initially the group was concerned mainly with general problems of natural language processing, studies of English morphology, syntax and grammatical classification, but from 1967 access to the Kuno-Woods' parser encouraged serious MT research (Chandioux 1977). In 1970 the group was renamed TAUM (Traduction Automatique de l'Université de Montréal), with an operational English-French system as its goal. The first prototype was constructed under the leadership of Alain Colmerauer (making particulary use of Colmerauer's Q-system software) and first tested in 1971 (TAUM 1971). Research continued until 1977 on the development of the TAUM prototype under Richard Kittredge as director. Since 1973 the TAUM group had been funded by the Translation Bureau of the Canadian Secretary of State; and in May 1975 the group was contracted to develop a system for translating public weather forecasts: TAUM-METEO was delivered the following year and has been in daily operation since 1977. The success of TAUM-METEO led to a contract for a more ambitious project, the translation of aircraft maintenance manuals from English into French. The TAUM-AVIATION project, under the direction of Marcel Paré, was the principal activity of the group from 1977 until 1980. In that year, an independent evaluation of the project (Gervais 1980) concluded that there was little prospect of a cost-effective production system in the near future, and the sponsors brought the TAUM project to an end in 1981. A general retrospective overview of the whole TAUM project has been given by Isabelle (1984); a comprehensive bibliography by Bourbeau (1983); and a detailed description and critical evaluation by Whitelock & Kilby (1983).[1] Reports of the protype system are to be found in TAUM 1971, 1971a, 1973a, 1973b and Kittredge 1972; the principal documentation for METEO is Chevalier et al.1978; and that of TAUM-Aviation is Baudot et al. 1977, Isabelle et al. 1978, Bourbeau 1981, and Isabelle & Bourbeau 1984.

13.1.1: System design

The TAUM team developed its basic design in close cooperation with other research groups, particularly the Grenoble group (Chandioux 1977). The experience of CETA convinced the Montreal group that the best prospect for adequate realistic and practical MT in an operational setting was a system based on the 'transfer' approach. The TAUM system represents a typical example of the 'transfer' approach in perhaps its purest form, having five basic stages: Morphological analysis of English, Syntactic analysis of English, Transfer, Syntactic generation of French, Morphological generation of French.

The other major lesson was the desirability of separating strictly the algorithms from the linguistic data and processing. As its computational metalanguage, TAUM adopted the Q-systems formalism developed by Colmerauer (1971), and which subsequently influenced the design of the Prolog programming language by Colmerauer and others at Marseille and at Edinburgh in the early 1970s (Kowalski 1985). Q-systems (Q=Quebec) are computer programs for the manipulation of tree structures and strings of trees irrespective of the labels attached to the nodes of trees. A tree

[1] See also: P. Isabelle 'Machine translation at the TAUM group', *Machine translation today: the state of the art*, ed. M. King (Edinburgh: Edinburgh University Press, 1987), pp.247-277

may be fully articulated as in a phrase structure representation, e.g. PH(SN(IL), SV(V(MANGE), SN(LA, CHOUCROUTE))), or it may represent a list (items separated by commas), e.g. L(A,B,C,D), where each item may itself be a tree, or it may represent a categorisation, e.g. PREP(TO), or a single node, TODAY. A string of trees is defined as a sequence of trees separated by plus signs, e.g. SN(PAUL) + V(VIENDRA) + DEMAIN + CHEZ + PRON(MOI). A Q-system rule converts strings (of one or more trees) into new strings, it may apply to the whole or to only a part of a string, and may include variables for labels, list or trees. For example, in the rule PREP(A*) + SN(X*) → OBJIND(P(A*), SN(X*)) the A* is a variable for a label (TO, FROM,...) and the X* is a variable for a list (of nouns). Clearly, the Q-system formalism is very powerful, capable of handling morphological and syntactic representations within any formal model, and also suitable for application in dictionary lookup procedures. However, it does have its drawbacks: the formalism does not permit easy copying of features from node to node, or specifying sequences of constituents,and as "the only method of passing control between Q-systems is to chain them" the formalism does not permit conditonal applications of grammars, with consequential wasteful invocations of inapplicable routines (Whitelock & Kilby 1983)

Whereas TAUM-METEO was written entirely in the Q-systems metalanguage, other computational procedures were also developed for TAUM-AVIATION. For certain applications the generality of Q-systems was too powerful; greater efficiency was needed for certain specialized tasks. SISIF was a finite-state automaton for pre- and post-processing of texts; LEXTRA was designed for lexical transfer; and for syntactic analysis Stewart (1975) developed REZO, an adaptation of Wood's ATN parser (Ch.9.13 above).

The first stage of linguistic processing was Morphological analysis. This involved the assignment of category labels (e.g. prepositions: WITHIN → P(WITHIN), including prepositional phrases: IN THE PROCESS OF → P(INTHEPROCESSOF), segmentation of prefixes (e.g. UNDERSTOOD → UNDER + STOOD), regularization of irregular forms (e.g. STOOD → SW(STAND) + ED(PST), restoration of prefixes (e.g. UNDER + SW(STAND) → SW(UNDERSTAND)), identification of suffixes (e.g. TRIED → TRI + ED, PUTTING → PUTT + ING), construction of potential base forms (TRI → TRY, PUTT → PUT). Dictionary lookup searched for both segmented forms (TRY + ED, SERIE + S, FLY + S) and full unsegmented forms (TRIED, SERIES, FLIES), rejecting those not located (SERIE, TRIED). It included the assignment of category labels (ADJ, N, ...), and 'features' (e.g. ANImate, CONCrete, ABSTract, for nouns, features of admissible arguments (subject nouns, objects, etc.) for verbs, and obligatory prepositions, such as TO with *listen*).

Syntactic analysis was in two phases. The first included the recognition of noun phrases and complex verb forms and the rearrangement of constituents as needed, e.g. DET(V*) + N(X*) → NP(N(X*), DET(V*)). The second established the 'canonical forms' of sentences. It incorporated both phrase structure rules and transformational rules: input strings of trees were formed into single complex trees and reordered (or deformed) as 'deep structure'-type representations. Thus, verbs were placed before their argument noun phrases, passive constructiins were made active, extraposed *it* forms were transformed (e.g. It be ADJ that S → S be ADJ) and relative pronouns were replaced by REL and the head noun copied into its argument position in the subordinate clause. An example analysis is shown in Fig.23.

Each arrow line represents a step in the analysis (i.e. the application of a replacement rule) working upwards from the 'surface form' at the bottom to the final form at the top. The example shows the inversion of article and noun in a noun phrase formation: DET(ART(DEF)) + N(COMMITTEE) → NP(N(COMMITTEE), DET(ART(DEF))), the testing of *interview* as noun (N) or verb (ZV), the inversion of the verb and its suffix *-ing* in order to identify the durative tense BE + -ING, and the placing of the verb information before the noun phrase (top line). During syntactic analysis some use was made of semantic features (derived from SL dictionary entries),

e.g. checking for compatibilities of verbs and direct objects. This could also involve the incorporation of semantic features of dependent 'attributes' (adjectives or nouns) into the set of features of a 'governing' noun; e.g. in a *4pound of cake* the properties of *cake* are subsumed in the features of *pound*, and in *defective pump* the selectional features relevant to *defect* are incorporated in the features for *pump*. Syntactic analysis was originally implemented by a Q-systems parser, but later in the AVIATION project the REZO parser was adopted.

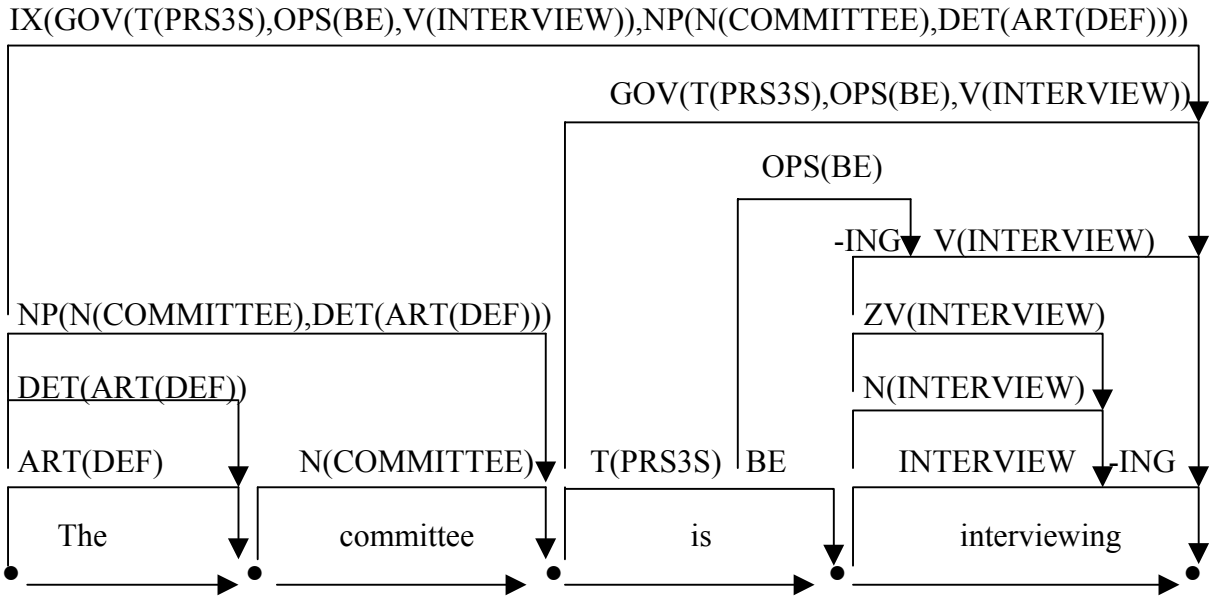


Fig.23 – TAUM syntactic analysis

Transfer had also two phases. First was Lexical transfer: the translation English 'words' (base forms) with their category labels into French equivalents via a bilingual dictionary, which could involve tree transduction, e.g. from *supply x with y* to *fournir y à x*. Although originally implemented in a Q-system formalism, for AVIATION the LEXTRA system was developed, which made it easier to specify the type of tree transducer required for particular lexical items. Unlike Q-systems (and similar formalisms), LEXTRA requires that input trees are stated explicitly. The constraint ensured correct formulation of lexical transfer rules. The second phase of Transfer was Structural transfer: the modification of certain parts of trees to simplify generation. Because of the close similarity of English and French 'deep' structures this phase was simpler than it might have been for other language pairs. Nevertheless, certain difficult problems were encountered, e.g. the handling of tenses and auxiliaries.

In Syntactic generation successive Q-systems broke down the complex tree output from Transfer into strings of trees. For example, the noun phrase:

SN(N(GENS), DET(LES), GP(P(DE), SN(N(VILLAGE), DET(LE))))

became: DET(LES) + N(GENS) + P(DE) + DET(LE) ; N(VILLAGE)

Finally, Morphological generation converted trees and strings into single 'surface' forms, e.g. DET(LES) → *les*, P(DE) + DET(LE) → *du*.

The TAUM system illustrates well the charcteristic features of MT transfer systems: the clear separation of the different stages of analysis and synthesis, the separation of linguistic data from processing algorithms (in this case, Q-systems), and the use of separate dictionaries for analysis, transfer and synthesis.

TAUM was similar to CETA in adopting the predicate-argument structure for 'transfer' representation (and in this respect TAUM 'transfer' representations are equivalent to CETA's 'pivot language' representations). However, analysis went no further than 'deep structure' (and

sometimes not that far, since complete transformational analysis was not pursued as it was not felt to be necesary for English-French translation). The main criterion was that representations facilitated transfer rules. As in CETA, semantic analysis in TAUM was confined to the use of semantic features (such as 'animate', 'mass', 'fluid') during tree conversion in Syntactic analysis and Transfer. No lexical decomposition was performed, and there was no pragmatic or discourse analysis.

The lack of a discourse component meant, in particular, that the achievement of coherent sequences of sentences during syntheis was virtually ruled out. TAUM translated sentence by sentence; the need for the retention of "global discourse information" during analysis was recognized, but could not be attempted (Isabelle 1984). Another aspect of text structure which TAUM could not tackle was anaphora. At an early stage of the project there had been a proposal to incorporate an intersentential routine, the so-called REF-Bug (Hofmann 1971), which moving left to right, across, into and out of sentences (or rather their 'deep structure' representations) would replace "each pronoun by the most recent noun of the same gender and number which it has met". It was, however, decided that too litle was known at the time about the practicality of incorporating such text processing in large scale systems and so the routine was never implemented (Isabelle 1984).

13.1.2: The METEO project

In 1974 TAUM, under some pressure from its sponsor, was looking for some practical demonstration of its MT programs. It looked for an application within a limited domain. The Canadian government's bilingual policy (Ch.9.1) led to a decision to broadcast weather forecasts in both English and French throughout Canada. As a result, TAUM was commissioned by the Canadian Bureau des Traductions to produce a system for translating meteorological reports from English into French. The development of TAUM-METEO (Chandioux 1976, Chevalier et al. 1978, Chandioux & Guérard 1981) took less than two years. Regional forecasts are written in a 'telegraphic' style with relatively limited vocabulary, i.e. a sublanguage 'dialect' of English. The research on METEO has in fact led to the development of branch of linguistic analysis concerned with the 'sublanguage' concept (Kittredge & Lehrberger 1982), with potential application not only in MT but in many other areas of linguistic research.

The restricted vocabulary and stereotyped syntax of meteorological reports enabled the designers to greatly simplify the basic TAUM system. The most important departure from the TAUM prototype was the elimination of a Transfer component; most of the processing usually done during transfer was incorporated into the analysis component. In effect, the system was no longer a 'transfer' system since analysis was TL dependent, i.e. METEO is a version of a (simplified) 'direct' system, albeit with much more genuine independent SL analysis than most such systems.

Another simplification was the dropping of morphological analysis before dictionary search, because there were so few variant forms of English words in this sublanguage. There was also a simplification of the arrangements for dictionaries. Instead of three separate dictionaries, only one was needed to give the French equivalents of English expressions and French morphological data; this was consulted only during stages of analysis. Changes in the usual TAUM parser were necessary because of the lack of tensed verbs in most meteorological reports. As a consequence METEO implemented a 'semantic grammar' (cf.Ch.9.17 and 15), in which rules operated not on syntatic categories (N, Adj, etc.) but on semantic categories (TIME, CONDition, etc.) Many structures follow a general pattern, such as: atmospheric condition, modification of condition, place, time (Kittredge 1981). Such patterns are the basis for rules producing, for example, the tree in Fig. 24 for the sentence *Snow occasionally mixed with rain tomorrow* (Isabelle 1984):

In this tree, MET1 is the label for the semantic pattern “atmospheric condition”, and COND indicates a “basic condition” defined as a noun phrase whose head bears the feature “weather condition” (e.g. *rain*, *snow*, *cloudy*, *sunny*), and which may optionally have an “accompaying condition”, CMOD.

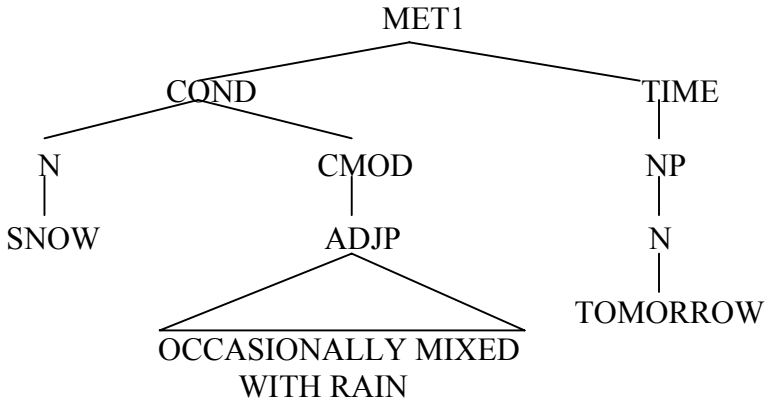


Fig.24 – METEO analysis

The system was completed in 1976 and has been fully operational since May 1977, initially dealing with reports from Halifax and Toronto and extended to the whole country in June 1978 (Thouin 1982). The system runs on a CDC Cyber 176 (previously Cyber 7600), with a Cyber 720 (Cyber 71) as front-end for man-machine interaction), and it translates daily from 1500 to 2000 short reports for the Canadian Meteorological Center in Dorval (a suburb of Montreal), amounting currently to some 8.5 million words per year (Isabelle 1984), Isabelle & Bourbeau 1984). The output is not revised before public broadcasting. An example (from Chandioux 1976):

> VALLEE DU BAS ST JEAN HAUT ST JEAN FIN DE L’AVIS DE VENT POUR LES DEUX REGIONS. CETTE NUIT NEIGE ET POUDRERIE DEVENANT PASSAGERES VENDREDI A L'AUBE. VENDREDI NUAGEUX AVEC FAIBLES CHUTES DE NEIGE PASSAGERES. CETTE NUIT VENTS FORTS DU NORD EST SOUFFLANT EN RAFALES DEVENANT VENTS FORTS DU NORD OUEST VENDREDI APRES-MIDI.

(English original:

> LOWER ST JOHN VALLEY UPPER ST JOHN RIVER WIND WARNING ENDED BOTH REGIONS. SNOW AND BLOWING SNOW TONIGHT BECOMING INTERMITTENT NEAR DAWN FRIDAY. CLOUDY WITH PERIODS OF LIGHT SNOW FRIDAY. STRONG GUSTY NORTHEASTERLY WINDS TONIGHT BECOMING NORTHWESTERLY WINDS FRIDAY AFTERNOON.)

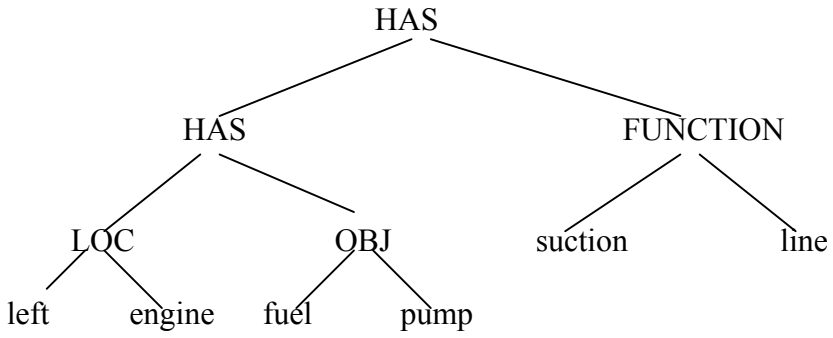
The TAUM researchers were somewhat surprised to find that on initial installation in 1976 the system failed to translate over 50% of reports. Considerable expansion of the dictionaries was required, particularly of regional names, and substantial amendments were made to analysis components to deal with participial verb forms (e.g. *rain developing, becoming cloudy* (Chandioux & Guérard 1981). After eighteen months’ improvements METEO became fully operational in May 1977. During the development stage and for a number of years subsequently the translators of the Meteorological Center made substantial contributions and suggested numerous refinemets (Thouin 1982). It now fails to translate only 20% of reports, largely because input is unedited and contains errors of typing, spelling and punctuation, all outside the control of the system itself. Failures from non-recognition of syntactic patterns are very rare (Kittredge 1981).

If the system is unable to parse it does not attempt a 'best guess' (by some 'fail-safe' strategy) but leaves the sentence to human translators. Translators are said to be pleased to be spared of the boring aspects of their work; their intervention is limited to the difficult and more interesting cases. Productivity of translated bulletins has increased markedly: the average time spent on MT-aided versions is 3.8 minutes, previously manual translations required 30 to 40 minutes. Chandioux & Guérard 1981). Although limited in scope, TAUM-METEO is the first, and so far only, MT system regularly producing translations which are not edited before being made available for public consumption.[2]

13.1.3: The AVIATION project

The TAUM-AVIATION project started in 1977 shortly after the installation of METEO at Dorval. The Translation Bureau of the Canadian Secretary of State commissioned TAUM to develop a system to translate the maintenance manuals of the CP-140 coastal patrol aircraft, amounting unofficially to an estimated 90 million words. TAUM was given a contract to develop a system to be ready for delivery of the aircraft three years later (Isabelle 1984, Macklovitch 1984). It was hoped that TAUM-Aviation would be the first advanced 'transfer' system to become operational.

Considerable organizational problems for such a large scale operation delayed serious detailed research for nearly two years. Nevertheless, there was a successful public demonstration in March 1979, and by the time of the deadline in May 1980, the operational prototype had been assembled (implemented on a Cyber 173), including dictionaries for a trial 70,000 word corpus from hydraulics manuals. The distinctive features of the sublanguage of the manuals have been described by Lehrberger (1982): typically such manuals have no interrogative and exclamatory sentences; commands omit articles, copulas or anaphors (*Remove used filter and discard; Check reservoir full*); and they contain an abundance of complex noun phrases (*hydraulic ground test stand pressure and return line filters*). To deal with the latter it was suggested that AVIATION include semantic categorisation of constituent nouns, in order to produce for *left engine fuel pump suction line* the following analysis (Isabelle et al. 1978), which specifies functional (FUNCTION), locative (LOC), possessive (HAS) and object (OBJ) relations:



However, the idea could not be implemented within the time span of the project.

---

[2] For later developments see: J.Chandioux 'Météo: 100 million words later'*, American Translators Association Conference 1989: Coming of age*, ed. D.L.Hammond (Medford, NJ: Learned Information, 1989), pp.449-453; A. Grimalia & J. Chandioux 'Made to measure solutions', *Computers in translation: a practical appraisal* (London: Routledge, 1992), pp. 33-45; chapter 12 in Hutchins, W.J. and Somers, H.L. *An introduction to machine translation* (London: Academic Press, 1992); and J.Chandioux & A.Grimaila, 'Specialised machine translation'*, Expanding MT horizons: proceedings of the Second Conference of the Association for Machine Translation in the Americas, October 1996, Montréal, Canada* (Washington, DC: AMTA, 1996), pp.206-211.

The basic structure of the AVIATION system followed the one outlined above, with new programs being compiled for certain stages. Non-inflectional morphological analysis was handled by SISIF, inflectional morpholoogy in both analysis and synthesis by separate adhoc PASCAL programs, and dictionary lookup by a new program SYDICAN. Otherwise, the implementation was as described, including REZO for syntactic analysis and LEXTRA for lexical transfer.

Preliminary results seemed to be encouraging. For example (from Isabelle & Bourbeau 1984):

> Les clapets de décharge incorporés sont champignon, sont rappelés par ressort à la position fermée. Une pression de 3450 psi s'exercant sur le clapet-champignon est suffisante pour vaincre la force de rappel du ressort et le clapet-champignon se déplacera de son siège en couteau.
>
> (English original: The in-line relief of valves are poppet-type, spring-loaded to the closed position. A pressure of 3450 psi impinging on the popet is sufficient to overcome the opposing force, and the poppet wil move from its knife-edge seat.)

The only manual revisions required were the replacement of *sont* by *du type* between *incorporés* and *champignon*, and the substitution of *s'écartera* for *se déplacera* in the last clause.

Although the quality of the material actually translated was quite good, the prototype system was failing to produce any output for between 20 to 40 per cent of input sentences. The TAUM team took the view that it was better to produce nothing at all than to risk incomprehensible output. This was a reasonable strategy for a system under development, but not for one intended as an operational system. Evaluation in March 1980 (Gervais 1980, summarized in Macklovitch 1984 and in Isabelle & Bourbeau 1984) revealed that "raw machine output" had "a degree of intelligibility, fidelity and style which reaches 80% of unrevised human translations (HT)", but that "revision costs are twice as high for MT" and thus "revised MT turns out to be more expensive than revised HT." (18.3 cents per word compared to 14.5 cents per word). Costs could be reduced by suitable man-machine interfacing, but the system could not be cost-effective until a volume of processing had been reached (5 to 6 millions words a year) which depended on the extension of the system beyond the protype's restricted domain of hydraulic system maintenance manuals. The final rejection of the system (Gervais 1980) was, consequently, not so much on the grounds of quality of output as on the amount of time and cost which would obviously still be needed to extend the system, in particular the dictionaries, for a production environment. The TAUM researchers naturally considered that the evaluation had been premature, that the system was still in its early developmental stages and rapid progress would have followed shortly; internal technical evaluations had concluded that 70% of failures had known solutions which could be corrected within 12 person/months of work (Isabelle & Bourbeau 1984). In brief, TAUM-AVIATION had not had a chance to show its potential (Isabelle 1984) – it is a story which has been repeated on numerous occasions in the history of MT.

A further feasibility study was conducted in May 1981, which compared TAUM-AVIATION with the Systran English-French system developed for the Commission of the European Communities (Ch.14.1) and with the interactive systems ALPS and Weidner (Ch.17.11-12). Although both operating costs and costs of dictionary updating were lower in the three systems than those in TAUM, the quality of output was considered to be much inferior. Indeed, revisers refused to rank them in terms of technical accuracy, "saying that they were all... arduous to revise", often finding it easier to retranslate directly from originals. The authors of the feasibility study were, consequently, unable to recommend any of the systems for purchase and regular use by the Canadian Translation Bureau (Macklovitch 1984).

The most important and telling objection to TAUM was, therefore, the high costs of dictionary compilation. The sublanguage approach aiming for good quality translation within a restricted domain demands careful semantic classifications of vocabulary and time-consuming analysis of lexical items in semantic and syntactic contexts. Without such laborious analysis, the

ambiguity of, e.g., *Remove fitting and drain plug* (is there an object called a *drain plug*?) could not be resolved. It can of course be argued that we should not expect any automatic system to be capable of resolving such ambiguities, but it was TAUM's aim to produce a high level of MT output.

Between 1976 and 1980 the Translation Bureau had invested over 2.7 million dollars in a MT system which proved not to be cost-effective. The Bureau has not lost interest in MT (it is still supporting MT-related research), but, as Macklovitch says (1984), TAUM had "made the unfortunate error of putting all its eggs in the same basket. When the AVIATION contract with the Bureau ended, it found itself with no other source of funding". In September 1981 TAUM was disbanded.

13.2: University of the Saar, SUSY (1967-)

Research on MT started in Saarbrücken in the mid-1960s with an experimental parser for German (Eggers et al. 1969) and an experimental Latin-German system (SALADIN) by Hubert Martin (Maas 1978). A more substantial project between 1967 and 1970 was set up to explore the possible development an early version of the Systran Russian-English system (Ch.12.1) for translation from Russian into German (Janda et al. 1970) On the failure of this attempt, the Saarbrücken group decided to develop its own ideas for a protoype Russian-German system. In 1972 (stimulated in part by a change of computer at Saarbrücken), this project was combined with other activities on language data processing and mathematical linguistics to form the 'Sonderforschungsbereich 100 Elektronische Sprachforschung' subsidized primarily by the Deutsche Forschungsgemeinschaft. The Russian-German prototype MT system was the starting point for research on the multilingual MT system known as SUSY (Saarbrücker Übersetzungssystem).

The SUSY system is an experimental 'transfer' system: programs for SL analysis and TL synthesis are independent, and there is the familiar separation of algorithmic procedures and linguistic data. It is designed as a multilingual system: programs have been written for a number of languages: Russian, German, French, English and Esperanto (the latter being a 'private initiative' according to Maas 1977); at one point, Danish was also considered (Maas 1978). There have been various changes in SUSY operations over the years. All versions of the system have been written in FORTRAN, initially for a CD3300, later for a TR440 computer (Maas 1981).Development of the basic SUSY system was essentially brought to an end in 1980. Since 1981, under the direction of Heinz-Dieter Maas, research has been on two fronts. On the one hand there is the project to develop a new version, SUSY II, incorporating insights from earlier research, and on the other hand there are investigations into possible practical applications of the basic SUSY system (SUSY I).

A detailed description of the current situation has been given by Maas (1981) and by Luckhardt et al. (1984); details of the basic SUSY I system are given by Maas (1978), and a series of reports entitled *Linguistische Arbeiten*, particularly Luckhardt (1976), Luckhardt & Maas (1976), and Maas (1977a).[3]

The system incorporates a number of dictionaries: SL morpho-syntactic dictionaries containing stems plus grammatical information; SL lexico-semantic dictionaries containing semantic markers, routines for disambiguation, for idioms and for creating (interlingual) transfer representations; SL-TL bilingual dictionaries containing SL-TL lexical equivalences and routines for aiding syntactic transfer; TL lexico-semantic dictionaries containing routines for deriving TL lexical entries from transfer representations; and TL morpho-syntactic dictionaries containing TL syntactic and inflectional information. Dictionaries have been developed for the following as SLs:

[3] For full descriptions with further references see: H.D.Maas 'The MT system SUSY', *Machine translation today: the state of the art,* ed. M. King (Edinburgh: Edinburgh University Press, 1987), pp. 209-246; and chapter 11 in Hutchins, W.J. and Somers, H.L. *An introduction to machine translation* (London: Academic Press, 1992)

German, English, French, Russian, and Esperanto; and for German, French and English as TLs. Bilingual (SL-TL transfer) dictionaries exist for English-German, French-German, Russian-German, Esperanto-German and for German-English, German-French, German-Esperanto; these pairs therefore represent the present SL-TL translation facilities of SUSY (Luckhardt et al. 1984) The largest dictionaries are those for German (140,000 entries in the SL morpho-syntactic dictionary, 75,600 in the lexico-semantic dictionaries, and 17,000 in the TL morpho-syntactic dictionary); the dictionaries for other languages are substantially smaller (the next largest is the Russian SL morpho-syntactic dictionary with just 15,000 entries), and only the English-German transfer dictionary has more than 10,000 entries. These sizes are indicative of the experimental nature of the system.

The basic SUSY I translation program has three main processes: Analysis, Transfer and Synthesis, divided into a number of stages. The first stage of Analysis (LESEN) inputs text from the terminal or file, and identifies word and sentence boundaries. Morphological analysis (WOBUSU) consults the SL morpho-syntactic dictionary, identifying stems and inflections, and providing tentative information for any words not found. The next stage (DIHOM) attempts the resolution of homographs by: examination of the compatiblities of adjacent word classes (the LEMMAT routine); tests for irregularities and inconsistencies, e.g. a German preposition, as opposed to a separable prefix, cannot occur at the end of a sentence (the TABHOM routine); and statistical evaluation of probabilities, preferences and weightings of word-class pairings (the GEWICHTE routine). The homograph resolution program is followed by SEGMENT, which divides the sentence into subclauses (on the basis of punctuation marks, conjunctions, etc.) and from such clues as relative pronouns provides a tentative dependency structure to indicate relations between the subclauses. The next two stages identify noun groups (NOMA) and verb groups (VERA), by using SL-specific information on possible combinations and structures. The following stage (KOMA, earlier: SYNAN) builds the phrase structures for transfer: noun groups are attached to governing verbs (using valency information), relative pronouns and reflexives are replaced by their antecedents, and uniform descriptions are provided for subordinate constructions (adjectival, participial clauses). The final stage of Analysis (SEDAM) operates with SL information from the lexico-semantic dictionary to resolve problems of lexical ambiguity; it also provides (interlingual) case markers for relevant prepositions or inflections, and identifies the valency relationships of verbal nouns functioning as nominalised governors, etc.; i.e. the result is a syntactic ‘deep structure’ representation with some (interlingual) semantic elements, akin to those found in GETA (Ch.13.3 below)

Transfer is a single stage process (TRANSFER) using the bilingual dictionary to replace SL lexical forms by TL lexical forms., and attempting to translate unknown words by reference to the morphological analysis (WOBUSU) or. Some entries may indicate changes in syntactic structures during transfer, e.g. in valency relationships; in the absence of explicit instructions, TRANSFER applies standard SL-TL syntactic transfer routines. The first stage of Synthesis (SEMSYN) is the counterpart of SEDAM, using the lexico-semantic dictionary to produce the TL forms for modal verb and idiomatic (fixed phrase) constructions, and TL lexical forms for (interlingual) case markers. Syntactic synthesis (SYNSYN) converts phrase structure trees into TL constructions using the TL morpho-syntactic dictionary. Then it produces sequences of strings (TL words) or stems with grammatical information, which can be handled by the final Synthesis stage (MORSYN) for the derivation of correct morphological forms and the production of the TL text.

A characteristic feature of SUSY I is its RESCUE mechanism. The main modules of analysis (SEGMENT, NOMA, VERA, KOMA) include checks for consistency of outputs. The identification of an inconsistency triggers RESCUE which sets less stringent checks and initiates a repeat analysis. In this way total failure of the program is prevented.

The basic framework of the SUSY programs was designed to handle inflected languages like German and Russian. The introduction of languages such as English, where information on

syntactic relationships is carried mainly by word order, was an important test of the modularity and multilingual capabilities of SUSY (Freigang 1981) Integration of a new language entails both changes in existing programs to cope with new structures and adaptation of new linguistic data to existing practices. An example of the latter was the redefinition of the English gerund (normally considered a verbal form) as a substantive, in order that existing rules of phrase analysis did not need to be altered. An example of the former concerned the (German SL) rule that only one noun phrase may precede a finite verb. If applied to English the sequence *commission measures* in: *In the commission measures have been considered concerning aid for developing countries* would be analysed as a noun phrase. In order to ensure the correct segmentation, for English the rule had to state that finite verbs must be preceded by one independent noun phrase (i.e. not within a prepositional phrase). Since such rules are probably not unique to German and English, they are considered part of the 'interlingual body' of the SUSY multilingual system, i.e. procedures which can be called upon to augment the 'interlingual core' of an absolutely language-independent routine. Each language specifies (through its 'linguistic characteristics') which of the procedures in the 'interlingual body' are to be employed. As a consequence, therefore, of the inclusion of English as a SL, algorithms were developed which could be directly applied in the analysis programs for other languages.

The new version SUSY II incorporates certain differences in stages of analysis and synthesis. In analysis, one of the main differences is that there is no longer a separate routine for homograph resolution; this is handled along with other disambiguation processes during syntactic analysis. After input (LESEN) and morphological analysis (WOBUSU), SUSY II has three analysis processes: the first deals with "linguistically simple cases", such as simple noun and verb groups which are unlikely to be ambiguous; the second deals with the construction of noun groups (including coordination, attribution, modification), computing 'scores' for semantic compatibilities and passing on only those constructions with the highest scores (i.e. those judged most likely to be correct); the third stage of analysis deals mainly with clause structures and relationships. Transfer in SUSY II remains as in the basic SUSY system. The Synthesis programs in SUSY II are deterministic tree transducers deriving structured TL representations for handling by the same Morphological synthesis (MORSYN) program as in the basic system. An important feature of SUSY II is its ability to produce TL output even with incomplete analyses, primarily because it can linearise any sequence of trees, subtrees or individual elements (cf. Ch.9.14 above). There is thus no longer any need for the special RESCUE mechanism in SUSY II.

The SUSY II processes include, therefore, a heterogenity of linguistic approaches and techniques. The system includes phrase structure rules (e.g. in SEGMENT) and transformational rules (in proceeding from surface structures to deep structures and vice versa) Operations may be rule-driven, lexicon-driven, or table-driven and may refer to features, categories or structures. The system includes valency frames, but not as obligatory conditions for acceptable parsings. In SUSY valencies represent preferred interpretations, thus allowing for incomplete or unusual grammtical relations in actual texts (cf. Ch.15.1 below). Optional valencies permit both the parsing of almost any input sentence and the elimination of false parsings; obligatory valencies are, however, necessary for lexical disambiguation.

SUSY II includes other techniques to overcome problems of rigidity in analysis sequences, e.g. the lack of backtracking, irrecoverability of earlier structures (as encountered in SUSY I, CETA and the LRC system, Ch.10 above) During the analysis of noun phrases, for example, parsing does not test for acceptability of input structures but applies all routines it can until there is a criterion which instructs it to stop. A number of stages include 'scored interpretations', preferential procedures and rule ordering; all methods for indicating that one procedure or analysis is better than others.

Much of this flexibiltiy has been made possible by the main innovation in SUSY II, the development of the chart data structure (as in Q-systems), for the representation of complex

structural ambiguities (Ch.9.15) Nodes of a chart representation can be individual elements (e.g. words) or subtrees (e.g. noun groups), and analyses can be made of any segments. Different types of analysis (subgrammars) can be attempted. Some will lead to 'dead ends', others will suggest alternative partitions of the string (sentence, clause) The analysis procedure, therefore, consists of repeated searches for the 'best' (most complete, least complex) combination of the partial analyses. The chart approach permits the easy incorporation of new analysis modules, it allows the definition of any sequence of subgrammars, it provides a uniform data structure, and it ensures that there is always a result of some kind even if some subgrammars fail. It is admitted, however, that SUSY II analysis procedures are likely to be much slower than those of SUSY I: where SUSY I fixes upon a single solution (which might turn out to be wrong), SUSY II pursues all the possibilities (Maas 1981).

Against these undoubted improvements in computational flexibility and multilingual capability must be set SUSY's sentence orientation. Analysis is explicity restricted to sentence structures, and it is predominantly syntax-based. Over the years more semantic processing has been incorporated, first by the insertion of some semantic analysis routines after syntactic analysis (in the basic SUSY I system since November 1976, when work began on the semantic disambiguation algorithm SEDAM), and later by the integration of some semantic procedures into syntactic analysis (in SUSY II) Nevertheless, semantic procedures remain relatively weak. But the principal consequence of SUSY's sentence orientation is the lack, both in SUSY I and (so far) in SUSY II, of procedures for discourse analysis and dealing with intersentential anaphoric reference; although admittedly, in this respect SUSY is little different from most other contemporary systems.

During the mid-1970s the SUSY group maintained close contacts with the GETA group. Between 1974 and 1978 much of the work on the French analysis program was carried out in Grenoble (Weissenborn 1977) There are a number of similarities in the techniques of SUSY and GETA (Ch.13.3 below): tree transducer algorithms, semantic dependency structures, the use of valency information, and the development of chart representations. In 1974 the SUSY group joined with GETA in the Leibniz group which had been set up for international cooperation in MT research (Ch.14.2); the formulation of a standard 'transfer' language appears to have owed as much to the Saarbrücken achievements as it did to GETA's influence. The later involvement of the Saarbrücken group in the Eurotra project (Ch.14.2) has been highly influential, both in terms of overall MT philosophy and with respect to computational developments.

While the emphasis has continued to be on experimental and theoretical development, the Saarbrücken group has not neglected practical implementations of the SUSY system. Various possibilities are discussed by Thiel (1981) and Luckhardt (1982), such as the compilation of glossaries, translation of titles and abstracts, information retrieval and question-answering systems. Some have already been tried. One of the first was SALEM (Saarbrücker Lemmatisierungssystem), which, using the SUSY analysis programs, automatically compiled concordances of German texts and provided a rich source of information on the syntactic and semantic contexts of German vocabulary (Eggers 1981) Shortly afterwards, there was the use of the SUSY-parser in an information retrieval system for the automatic indexing of legal documents (TRANSIT, under development by Harald Zimmermann since 1981). So far the system has analysed some 7 million words of patents texts and automatically compiled a "lexicon of 110,000 automatically decomposed German compounds" (Luckhardt et al. 1984)

Later, in 1981, came an investigation of the possible integration of SUSY I in the translation service of the Bundessprachenamt. The requirements for an operational SUSY translation system were outlined by Wilms (1981): sophisticated facilities for text processing before and after translation, access to specialised dictionaries (in this particular instance, integration with the service's own LEXIS databank (cf. Ch.17.6), improved procedures for dictionary updating, and creation of text-specific *ad hoc* glossaries. According to the designers the project was terminated in 1983 after a test of SUSY output which revealed that the Bundessprachenamt was "not willing to

accept any translation below the usual standard level of translations produced by its own translators" (Luckhardt et al. 1984). More pertinent perhaps was the lack of text-handling facilities in the SUSY configuration. A recent project SUSANNAH (SUSY ANwenderNAH, or 'SUSY user-oriented') plans to rectify this lack by developing a translator workstation which will support interfaces to term banks and glossaries, office communication systems, as well as access to SUSY itself (cf.Ch.17 below).[4]

The most recent application of SUSY is in the cooperative project since 1983 with Kyoto University (Ammon & Wessoly 1984-5). The goal is the automatic translation of German document titles into Japanese and of Japanese titles into German. The unusual feature of the project is the use of English as an intermediary (or 'switching') language: translation from German to English and vice versa will be done by SUSY, translation from English to Japanese and vice versa by the Kyoto system TITRAN (Ch.18.13). Because translation is to be 'indirect' the differences in MT design between the 'deep analysis' approach of SUSY and the restricted 'surface analysis' approach of TITRAN are not considered to be as significant as they would obviously be if the two systems were to be directly coupled. For this project the two systems can be developed independently; all they need to have in common is an orientation to the same subject field (computer science). Nevertheless, both have their limitations. TITRAN is not able to insert English articles, it lacks routines for relative clauses, and it relies considerably on 'lexicographic' solutions to compound forms. On the other hand, the centrality of the verb in SUSY's methods of analysis (valency relations, clause structures) may inhibit satisfactory treatment of titles, which generally contain few finite verb forms. The project is strictly practical; the evaluation criteria are comprehensibility and readability by technical experts, and fidelity to information content, but not stylistic adequacy.

13. 3: Groupe d'Etudes pour la Traduction Automatique (GETA), University of Grenoble (1971-)

A change in computer facilities in 1971 encouraged the research team at Grenoble to rethink the design of their MT system. Now renamed Groupe d'Etudes pour la Traduction Automatique (GETA), the team decided on a transfer approach. Experience with the basically 'interlingual' CETA system (Ch.10.1 above) had revealed disadvantages in reducing texts to semantic representations and destroying in the process a good deal of 'surface' information useful for TL synthesis. (There is no point, for example, in converting a SL passive form into an active representation if it has only to be reconverted into a similar TL passive.) The aim of the GETA team is to design a MT system which is highly flexible both in its programming and in its linguistic aspects, a system which will encourage cooperative activities with other MT research groups, a 'multilingual' system capable of translating from and into any European language, and which is as 'portable', i.e. machine independent, as possible. As in most contemporay advanced MT systems, GETA maintains strict separation of linguistic data and programming procedures (a feature also of CETA), thus enabling linguists to work with familiar linguistic concepts (grammatical categories, dictionaries, grammar rules, semantic features, etc.) The earlier stages in the development of GETA are described in Boitet 1978, Chauché 1975, Vauquois 1977. The most recent version of the GETA system, developed since 1978, has been named ARIANE-78; the most comprehensive descriptions are to be found in Boitet & Nedobejkine (1981) and Whitelock & Kilby (1983); and updating on the most recent versions is to be found in Boitet (1983), Boitet & Nedobejkine (1983), and Boitet et al. (1985).[5]

---

[4] For later applications of SUSY see: H.D.Luckhardt & H.Zimmermann, *Computergestützte und maschinelle Übersetzung: praktische Anwendung und angewandte Forschung* (Augsburg: AQ-Verlag, 1991)

[5] For later developments see chapter 13 in Hutchins, W.J. and Somers, H.L. *An introduction to machine translation* (London: Academic Press, 1992)

GETA-ARIANE is basically a 'transfer' system with morphological and syntactic analysis, transfer, and syntactic and morphological synthesis. The results of the analysis programs are dependency-tree type 'deep structure' representations in essence rather like those in CETA, but including a certain amount of 'surface' information (to assist in TL synthesis). It is no longer the objective to establish 'universal' pivot languages, rather each SL has its own 'pivot' or 'transfer' interface. The Transfer program has two stages: the conversion of SL 'lexical' elements into equivalent TL 'lexical' elements (involving reformation of tree structures as necesary), and the conversion or transformation of SL 'pivot' structures into equivalent and appropriate TL 'pivot' structures.

GETA has three main algorithmic components, one for the conversion of string representations into tree structures, ATEF, one for the transformation of trees into trees, ROBRA, and one for the conversion of trees into strings, SYGMOR. In addition there is an algorithm for consultation of the transfer dictionary, TRANSF; (recently modified as EXPANS, Boitet et al. 1985). Each algorithm is suited to particular stages of the translation process. There are basically six phases (Boitet & Nedobejkine 1981): 'Morphological analysis', using a battery of dictionaries to produce all possible category assignments and preliminary identification of some noun and verb groups, i.e. converting strings into partial tree structures by the ATEF algorithm; 'Multilevel analysis' (i.e. syntactic analysis), producing dependency-tree representations which combine both 'surface syntactic' information and the kind of 'deep syntactic' information found in the CETA trees (Ch.10.2. above) via the tree-tree conversion algorithm ROBRA; 'Lexical transfer', converting TL representations by SL-TL dictionary substitutions (via TRANSF); 'Structural transfer', transforming SL trees into TL 'pivot' trees (via ROBRA) – the two transfer phases working closely together; 'Syntactic generation', producing TL 'surface' trees; and 'Morphological generation', converting these trees into TL strings by the SYGMOR algorithm. The diagram illustrating the general system of ARIANE-78 (Fig. 25) is adapted from Boitet & Nedobejkine (1981).

A major premiss of the GETA team has been that the algorithms employed at any particular stage should be no more complex and no more powerful than necessary for handling the linguistic data in question. On this argument it rejects the use of such powerful algorithms as the Q-systems (found in TAUM, cf.Ch.13.1 above) and ATN parsers (Ch.9.13) for the simple manipulation of strings in, for example, morphological analysis and synthesis. In GETA-ARIANE algorithms of various levels of generality are applied according to the relative simplicity of the processes; thus the algorithms for morphological analysis (ATEF), lexical transfer (TRANSF) and morphological generation (SYGMOR) are less general than the powerful ROBRA algorithm for syntactic analysis, structural transfer and syntactic generation.

The ARIANE-78 system itself does not, in principle, define the 'depth' to which a text is analysed; the level for translation of particular structures may be determined by the linguist or analyst himself; and it could in theory range from surface syntax to an abstract interlingual representation (Whitelock & Kilby 1983). In practice, however, GETA researchers agree upon transfer structures combining both deep and surface syntactic information. It is indeed a major advantage of GETA's data representation that different levels of analysis can be incorporated simultaneously on a single labelled tree structure. It allows for interaction between levels and provides a fail-safe mechanism, i.e. if further analysis at one level is unsuccessful, analysis at another level can be attempted – complete reanalysis of the whole text segment is not necesary. This is possible because representations may include information derived from a variety of levels of interpretation.

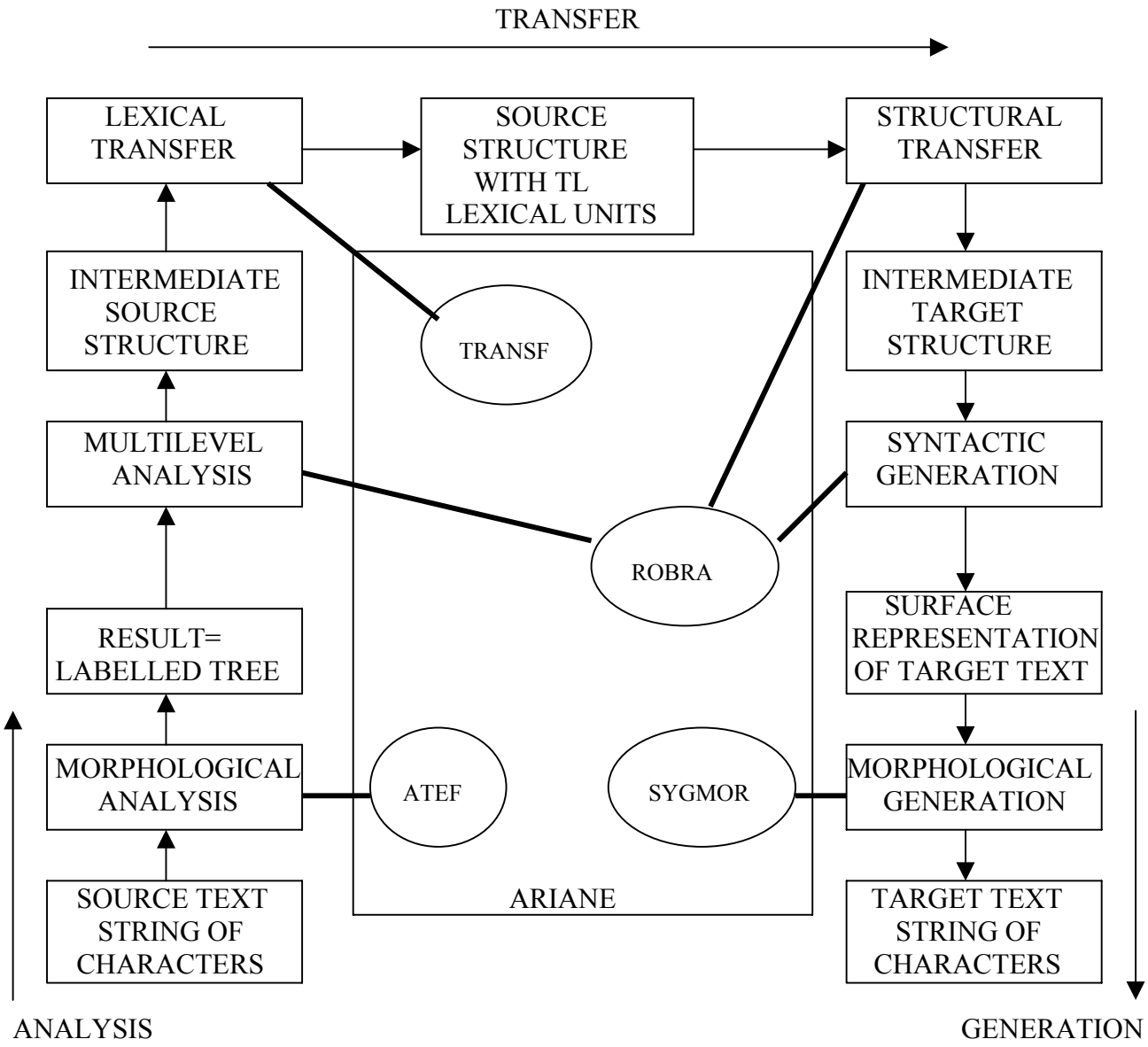


Fig.25 – The GETA system

Experience with CETA had shown the value of including different levels of grammatical and semantic information in SL representations. In GETA 'deep structure' or SL transfer representations may include a mixture of: syntactic classes (adj, noun, NP, VP), grammatical functions (subject, object, etc.), and logico-semantic relations (predicates and arguments). In other words, they combine information about phrase structure relations, dependency relations and semantic or logical relations. For example,the sentence: *Cette musique plaît aux jeunes gens* would have the tree shown in Fig.26, where

(a) UL indicates lexical items (MUSIQUE), (GENS) etc.
(b) CAT indicates grammatical categories such as noun phrase (GN) and adjective (ADJ)
(c) FS indicates dependency relations such governing node (GOV), subject (SUJ) and attribute (ATR)
(d) RL indicates logico-semantic relations such as ARG1 and ARG2.

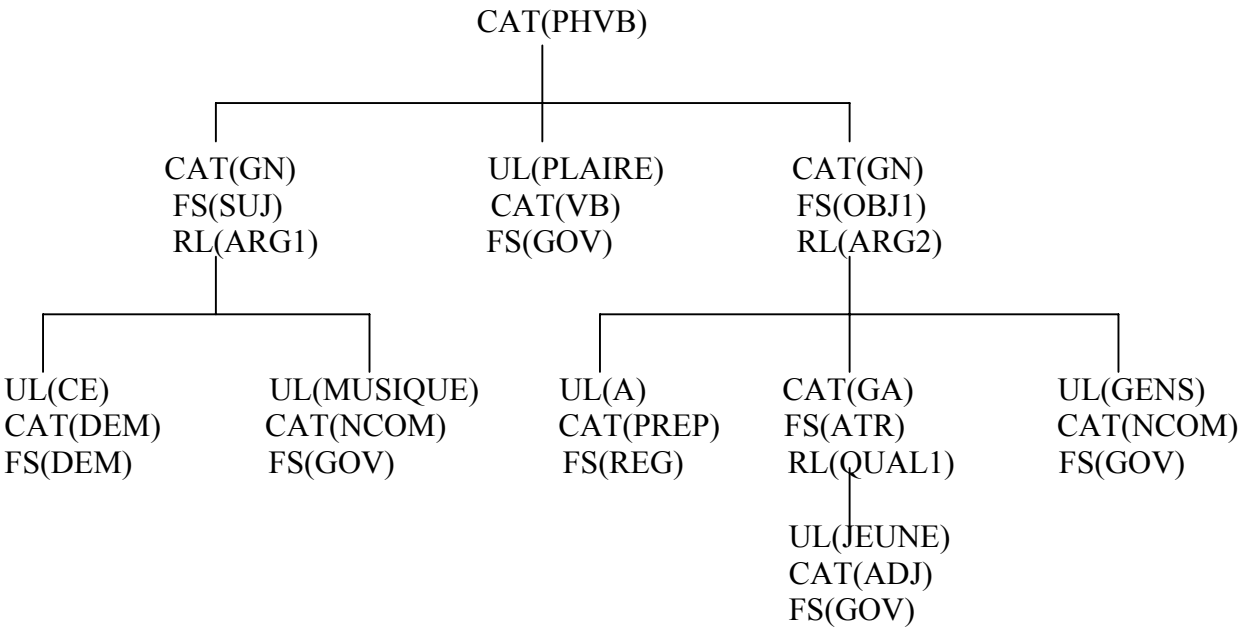


Fig.26 – GETA transfer representation

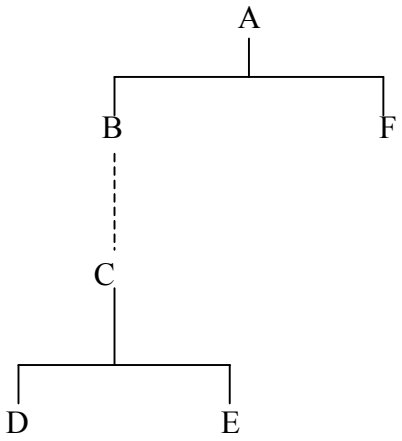
The first stage of analysis is performed by the non-deterministic ATEF algorithm (Analyse des Textes en Etats Finis). ATEF is basically a program for segmenting input SL text into stems and affixes (morphological analysis) and assigning tentative grammatical categories (preliminary stage of syntactic analysis). The basic information is provided by series of dictionaries including dictionaries of idioms, lexical items (stems) and affixes. Essentially the procedure is as in TAUM (cf.Ch.13.1 above) where all possible segmentations are tested against stem and endings dictionaries. However, in addition ATEF performs preliminary syntactic analysis by testing adjacent forms for morphological compatibilities, e.g. the endings of putative adjective and noun forms. In essence it is a finite state non-deterministic algorithm (but not as powerful as an ATN parser) transforming strings into sequences of labelled trees. (The argument for the adequacy of finite-state parsers for this purpose was made a number of years earlier by Vauquois et al. 1965, Vauquois 1971)

The basic ATEF procedure examines all the possible segmentations of a text form (i.e. usually a word). For example, German *AUTOMATEN* might be segmented as AUTOMAT+EN, AUTOMAT+E+N, AUTO+MATE+N; of these only the first will be retained by ATEF after consultation of a stem dictionary. Unless constrained, ATEF examines every possible segmentation of a word; the combinatorial implications can however be anticipated by linguists who may introduce conditions on application and intervene in procedures. The result is a highly complex algorithm; ATEF is undoubtedly highly flexible, allowing analysts to make use of whatever techniques seem most appropriate in each instance, but this flexibility would appear to be at the cost of a complexity of procedural structure which makes the linguistic bases of the algorithm less accessible to the researcher compiling and debugging programs.

For syntactic and semantic analysis the team has developed an algorithm for the transformation of one abstract tree or subtree into another. This is the extremely flexible ROBRA algorithm, deriving from translation formalisms of the CETA system (Ch.10.1 above), cf. Veillon et al. (1967). ROBRA is a tree-transducing mechanism (ch.9.14) taking the output of ATEF to produce the kind of multi-level syntactico-semantic representation shown above. In effect, it operates by applying cycles of subgrammars (each consisting of ordered sets of 'transformational rules', i.e. tree transducing rules), where each subgrammar is concerned with a specific range of linguistic phenomena, e.g. relative clause structures, infinitive complementation, comparitive

structures, etc. A typical sequence (Boitet & Nedobejkine 1981) in Russian analysis might be: division into clauses (using punctuation, conjunctions, etc.), identification of noun phrases, searching for elements dependent on noun phrases, finding the nominal antecedents of relative clauses, looking for infinitives, identifying subjects, objects, etc., specifying case ('actant') relationships, and so forth.

In general transformational rules operate on sub-trees and are constrained by the context in which they apply, i.e. the rest of the tree. Thus, as an abstract example, the sub-tree C(D,E) in the following (sub)tree:



may be converted into C(E,D) only if B is to the immediate left of F in a subtree A(B,F). Other rules might permit B and F in any order; or might allow optional intervening elements or subtrees; or might allow intervening structures only of a specific type; and so forth. In addition, of course, conditions may specify information of any kind (e.g. grammatical category, constituency structure, dependency) already generated, whether by ATEF or by preceding phases of ROBRA.

The considerable variety of possible sub-trees, the need to permit any number of optional structures within sub-trees and the abstractness of the transduction procedures make ROBRA both a highly flexible and a highly complex algorithm. This abstractness and this flexibility allow the linguist to apply any methods of analysis which appear most appropriate. The linguist decides what transformations are to be used in particular instances and what conditions are to be attached to their use. The linguist can therefore construct subgrammars to be applied in any order and under any conditions he may specify. He might, for example, construct a set of different subgrammars for the treatment of noun groups, one for simple cases, another for complex cases. He might apply a strategy using dependency relations in one subgrammar and a strategy using phrase structures im another. Each subgrammar can function in a number of distinct execution modes (at the discretion of the linguist), e.g. (i) in mode 'U' (unitary) where rules are applied once only, or alternatively in mode 'E' (exhaustive) where rules may be applied to trees on which they have already operated, or (ii) in mode 'H' (high) where priority of application is given to subtrees at the highest level of the tree, or in mode 'B' (bottom) where rules are applied first to the lowest subtrees. The system provides the linguist with a vast choice of possible approaches to analysis and assures him that, whatever the strategy or 'grammar' used, there will always be a result at the end of a finite application of rules. This is because tree-transducing algorithms do not test for the *acceptability* of structures (i.e. they do not filter out ill-formed structures) but test for the *applicability* of transduction rules (Vauquois 1977) The subgrammars work on sub-tree specifications. If a rule does not apply the tree remains unchanged; if no rule of a subgrammar can be applied there will always be a tree as output on which other subgrammars may operate. However this very flexibility may perhaps have its disadvantages. The interactions of component procedures and the interrelationships of transduction rules reduce the modularity of the system as a whole. It is

therefore difficult to maintain consistency in linguistic procedures, particularly when transfer rules may apply at any level (or any combination of levels) for any linguistic structure. The danger is an *ad hoc*-ness in grammatical analysis almost as great as in the early MT systems which worked virtually without any conception of grammatical theory.

Transfer operates in two phases: Lexical transfer and Structural transfer. The former utilizes the TRANSF dictionary relating 'canonical' stems of two languages (Russian and French, usually) and specifying in transfer rules any necessary subtree constructions, e.g. in the case of idioms or where the equivalent for a SL word is a TL phrase (Russian *nuzhno*: French *il faut*; English *tend*: French *avoir tendence*) or a SL phrase is a single TL word (English *let...know*: French *informer*). In general, compound SL forms (e.g. English *carry out*) will have already been identified as units during an analysis phase. It is during Lexical transfer that most problems of multiple equivalence are resolved by taking into account SL structural relationships, e.g. English *look* becomes French *regarder*, if in SL transfer representation it is linked with *at*, or *chercher*,if linked with *for*, or *ressembler*, if linked with *like*, etc. Nevertheless, some problems remain unresolved and are left to post-editors for the final decision, e.g. English *process*: French *processus* or *procédé* (as in the example translation below)

Structural transfer is concerned with syntactic differences, e.g. where a Russian verb may govern the instrumental (*pol'zovat'sya*) the French equivalent may require a preposition (*utiliser de*). Other obvious changes are the transposition of nouns and adjectives and the treatment of negation. Again the ROBRA algorithm is utilized to transform subtrees. The stage is followed by Syntactic generation (of French), also using ROBRA. As in Syntactic analysis, it consists of cycles of subgrammars, handling such linguistic operations as the generation of TL articles, the specification of auxiliary verbs (in the determination of correct tenses and moods), the distribution of agreement information (e.g. adjectives modifying nouns), and the treatment of TL coordination.

The final stage is Morphological generation, converting labelled trees output by Syntactic synthesis into strings of characters. For this, the SYGMOR algorithm is applied; like ATEF, SYGMOR is a realisation of a finite-state mechanism, reflecting the lesser complexity of this process. There are two phases, tree-string convertion and string-character conversion, during each of which TL dictionaries are consulted (e.g. to establish correct morphological forms, such as: doubling of consonants, JET+E $\rightarrow$ JETTE; insertion of 'e', MANG+ONS $\rightarrow$ MANGEONS; contractions, DE+LE $\rightarrow$ DU, etc.). There would seem little doubt that SYGMOR is sufficiently flexible to deal with the linguistic operations appropriate at this level, e.g. contraction, elision, inflection (Whitelock & Kilby 1983)

In principle, the units of translation treated in GETA-ARIANE procedures are "not sentences but rather one or several paragraphs, so that the context usable, for instance to resolve anaphors" is more extensive than earlier systems (Boitet & Nedobejkine 1981). In practice, however, "the bulk of analysis and generation operates essentially at the sentence level" (Vauquois & Boitet 1984). During Structural analysis it is possible to use some inter-sentence context to resolve anaphors (if no suitable candidate can be found in the sentence being analysed), but in general ARIANE does not incorporate systematic discourse analysis; in particular no attempt has been made to derive 'hypersyntactic' representations of paragraph units.

The Russian-French version of ARIANE-78 has been developed on a corpus of scientific and technical texts in the fields of aeronautics, space science, thermodynamics and nuclear physics. By June 1984 (Vauquois & Boitet 1984) the dictionaries contained 7000 Russian lexical units (about 17000 words) and 5600 French lexical units (about 16000 words). A number of example translations are to be found in Boitet & Nedobejlkine (1981); an example of a 'raw' translation is the following extract:

> LE SYMPOSIUM EST CONSACRE A LA SPECTROSCOPIE NUCLEAIRE ET A LA STRUCTURE DU NOYAU ATOMIQUE. DANS LE MOT D'ENTREE ON SOULIGNE LE ROLE IMPORTANT QUE LE SYMPOSIUM A JOUE DANS LE DEVELOPPEMENT

DE LA PHYSIQUE NUCLEAIRE DES ENERGIES FAIBLES EN UNION SOVIETIQUE. PENDANT LE SYMPOSIUM ON A EXAMINE LA SERIE DES ETUDES IMPORTANTES REALISEES PAR LES SAVANTS SOVIETIQUES. EN PARTICULIER, ON A ETUDIE LA NON-CONSERVATION DE LA PARTIE DANS LES PROCESSUS?PROCEDES? NUCLEAIRES, CREATION DU MODELE DU NOYAU NON-AXIAL, DIVISION SPONTANEES DES ISOTOPES DES ELEMENTS SUPERLOURDS ET DECOUVERTE DE L'EFFET DES OMBRES POUR LA DISPERSION DES PARTICULES...

As the authors note, the only awkward rendition is that of *mot d'entrée* instead of *introduction,* which could easily be solved by an addition to the idiom dictionary. The quality is impressive, but what must be kept in mind with ARIANE (as with most experimental systems) is that these are translations from prepared texts with a relatively limited range of vocabulary.

In many respects GETA is the most advanced of current MT systems, representing close to what is considered by most researchers at the present time to be the 'state-of-the-art'; nevertheless it has its shortcomings, as Whitelock & Kilby (1983) have pointed out. The GETA team are aware of these and are considering improvements: e.g. a unified formalism for the metalanguage of all components, a single data structure (the chart, cf. Ch.9.15 above) to replace string and tree forms, and the improvement of system portability (by implementations in Pascal or LISP). The cost of 'portability' could be high, however, as Vauquois & Boitet (1984) have discovered. A comparison was made with Kyoto University's system, whose design is similar to that of ARIANE-78 (cf. Ch.18.6 below) and whose GRADE tree-transducer is comparable to ROBRA, but which is implemented in a dialect of LISP rather than PL360 assembler. The results showed that "the LISP implementation is 40 times more voracious in computer time and space".

The strength of the GETA system is its linguistic and computational techniques, particularly in the areas of morphological and syntactic analysis and transformation. Its weakness remains, as Boitet (1982) readily admits, in its dictionaries; unlike other MT groups, such as TAUM, the Grenoble team has not been able to call upon terminological assistance.

Related to this weakness is what must be regarded as the main 'deficiency' of GETA from the viewpoint of current interest in the implementation of AI techniques (cf. Ch.15 below): the lack of semantic processing beyond the traditional lexical approaches. Part of the reason for this is, of course, the experience with the 'pivot language' of CETA. However, researchers at GETA have outlined various ideas in this area. Boitet (1976) suggested the incorporation of a 'semantic parser' and the expansion of its semantic information to include 'preference' and 'inference' semantics on the lines indicated by Wilks (cf. Ch.9.17 and 15.1). More recently in Boitet & Gerber (1984) and Gerber & Boitet (1985) plans have been outlined to incorporate 'real world' or 'metalinguistic' knowledge of the type found in AI expert systems. The basic idea is that two 'expert corrector systems' could be inserted between analysis and transfer and between transfer and generation. The first, for example, might resolve problems of adjective and noun coordination: in *The mixture gives off dangerous cyanide and chlorine fumes* the analysis required is not *(dangerous cyanide) and (chlorine fumes)* but *dangerous (cyanide and chlorine) fumes.* An expert 'knowledge base' for chemistry could help in such cases; but problems of practical implementation are yet to be tackled.

Development of the Russian-French system on ARIANE was the dominant activity at GETA from 1973 to 1980. By 1979 it became evident that GETA had to pay more attention to the construction of a practical working MT system. To this end, GETA has initiated an "experimental translation unit" using a "production-oriented subset of ARIANE-78, PROTRA" translating monthly about 50-60 abstracts from *Referativnyi Zhurnal* in the fields of space science and metallurgy (Boitet & Nedobejkine 1983, Vauquois & Boitet 1984). The dictionary contained 7000 Russian lexical units (about 20,000 words) in 1983 (Boitet 1983). With no OCR equipment and no access to magnetic tapes of abstracts, texts have to be entered manually. There is no expectation that 'raw' translations will be adequate; post-editing is envisaged as an essential component. In

practice, revisers have available the original text, the 'raw' translation and access to dictionaries, from which they produce translations in an interactive environment. An example of output is from Boitet & Nedobejkine (1983):

> Ce satellite est destiné à l'étude des ressources naturelles de la terre à l'aide de l'appareillage télévisé et des radios mètres de micro onde élaborées et fabriquées par les spécialistes de 1'lnde.

The main change on revision was to: *appareillage de télévision et de radiomètres à microondes.* Evidently, PROTRA would seem to have some potential as an interactive system (cf. Ch.17.7ff.). In connection with this project, the Grenoble team has developed various software facilities: ATLAS for updating dictionaries, VISULEX for revising MT dictionaries, and THAM for interactive computer-aided translation (Bachut & Verastegui 1984, Boitet *et al.* 1985); the latter is constructed on similar lines to Melby's system described later (Ch. 17.10).

Although the Russian-French system has been the principal activity at GETA, the facilities provided at Grenoble by the ARIANE configuration have supported a number of other MT projects (Boitet 1982, Vauquois & Boitet 1984). From the creation of the 'Leibniz' group (Ch. 14.2 below) in 1974 until 1978, researchers of the Saarbrücken MT group (Ch. 13.2 above) made use of the GETA facilities in the French analysis projects of Stegentritt and Weissenborn. A somewhat similar arrangement was made between 1977 and 1980 with the Nancy MT group, which lacked sufficient computer facilities at the time. Using GETA-ARIANE programming facilities, members of the group were able to experiment on an English-French system, which in general conception apparently differed markedly from that of the GETA group — an illustration of the flexibility of the computational facilities at Grenoble.

Among the small-scale MT models developed on GETA facilities are the Portuguese-English MT project pursued intermittently since 1973 by P. Daun Fraga, a researcher from the University of Campinas (Brazil), a multilingual system from Chinese into Japanese, French, English, and German (Feng Zhi Wei), an English-Chinese transfer system (Yang Ping), an English-Thai system, an English-Japanese system in collaboration with the University of Kyoto (Tsujii 1982), a French-English system (J. P. Guilbaud and M. Dymetman), and a German-French system under development since 1979 using the same generation programs as in the Russian-French project (the analysis programs are being written by J. P. Guilbaud and M. Stahl)[6]. More substantial is the English-Malay project, begun in 1979 in cooperation with Universiti Sains Malaysia, Penang, with the intention of producing teaching materials in technical fields. It is now approaching the stage of systematic testing of the prototype — although as yet the dictionaries are still small (1800 English and 1800 Malay lexical units). Most significant of all for the long-term future of the Grenoble MT group is the involvement since 1983 with a number of industrial companies in a French national project (ESOPE). The project's target is a system for the translation of aircraft manuals from French into English, and perhaps later for the translation of computer manuals from English into French.[7]

Although, as remarked earlier, the quality of ARIANE translations appears to be quite high, a detailed evaluation of the system has not been undertaken, nor has any attempt been made yet to measure real costs (Vauquois & Boitet 1984). The team remain convinced of the overall validity of their 'transfer' approach; of the need for contrastive information about the languages concerned and its retention in 'multilevel' representations; of the use of mixtures of syntactic and semantic

---

[6] See: J.P. Guilbaud 'Principles and results of a German to french MT system at Grenoble University (GETA)', *Machine translation today: the state of the art*, ed. M. King (Edinburgh: Edinburgh University Press, 1987), pp.278-318

[7] See: C.Boitet, 'The French national MT project: technical organization and translation results of CALLIOPE-AERO', *Computers and Translation* 1 (1986), 239-267; C.Boitet, 'Current machine translation systems developed with GETA's methodology and software tools', *Translating and the computer 8: a profession on the move. Proceedings... 13-14 November 1986*, ed. C. Picken (London: Aslib, 1987), pp. 114-132.

information during disambiguation; of the implementation of 'transducers' rather than 'analysers' given the non-specificity of linguistic 'knowledge' in MT systems (akin to the probabilistic nature of AI 'expert system' knowledge bases); of the use of heuristic techniques in programming grammatical analysis rather than complete searches of complex grammars; and of their insistence that the system produces some output, however full of errors, rather than stopping completely when ill-formed structures are encountered.

13.4 Linguistics Research Center, University of Texas, METAL (1978-

At the end of the German-English 'interlingual' MT project in 1975 (Ch.10.3) there was a further hiatus in research at the Linguistics Research Center (LRC). The complexities of transformational parsing and the inadequacies of the syntactic interlingua led LRC to a redesign of their system on different principles. In 1978 the USAF Rome Air Development Center provided funds for researching an operational MT system, generally referred to as METAL. In 1979 this funding was augmented by support from Siemens AG, West Germany, who became sole sponsors of the project in 1980. The main objective has been the development of a German-English system for translating telecommunication and data processing texts, though there are reported to be other applications (e.g. English analysis) under investigation (Slocum 1984). The aim is not a fully automatic system, it is recognised that current techniques do not make such a goal realistic; therefore, post-editing facilities are built into the total configuration. The principal researcher for the LRC-Siemens project is Jonathan Slocum, while overall direction of the Linguistic Research Center remains in the hands of Winfred Lehmann.

Like the previous system, METAL is based on methods of theoretical and computational linguistics: context-free phrase-structure algorithms, transformational rules, and case-frame analyses. However, whereas previously 'deep structure' analyses were intended to be language-independent (interlingual), the METAL 'deep' analyses are representations specifically for bilingual transfer. The LRC group has concluded that a MT interlingua or 'pivot language' is "probably impossible" and have adopted the alternative, "a transfer component which maps 'shallow analyses of sentences' in the SL into 'shallow analyses of equivalent sentences' in the TL" (Slocum 1984).

In basic form then, METAL is a standard 'transfer' system with stages of Analysis, Transfer, and Synthesis. However, the stages of transfer and synthesis are not as clearly separated as in other 'purer' examples of the 'transfer' approach, since Transfer includes operations which might normally be considered part of syntactic synthesis. The programs are written in LISP, making this LRC system more accessible to other MT researchers than previous LRC systems which were written in lower-level codes. Since LISP functions may be driven by linguistic data, there is no longer in METAL a strict separation of algorithmic and linguistic data. Details of METAL are given by Slocum (1980, 1982, 1984), by White (1985), and by Whitelock & Kilby (1983).[8]

METAL dictionaries are either monolingual (SL or TL) or bilingual (for SL-TL transfer). In 1984 there were about 10,000 lexical entries for each of the languages: German SL and English TL (Slocum 1984). Entries in both SL and TL monolingual dictionaries are formatted in the same way, namely a series of 'features': a 'canonical form' (i.e. base stem), a set of allomorphs or morphological variants (i.e. surface forms), grammatical information for each variant (adjective, noun, etc.), lexical collocation (e.g. indicating a discontinuous element associated with the stem, as in *look ... up),* a concept number (used to relate semantically related forms, e.g. *compute, computer, computation, computational, computable),* and subject field (e.g. to assist in the disambiguation of polysemes). In addition, preferences (weightings) may be assigned to particular interpretations (e.g.

[8] See also: J. Slocum 'METAL: the LRC machine translation system', *Machine translation today: the state of the art*, ed. M. King (Edinburgh: Edinburgh University Press, 1987), pp.319-350

a compound or fixed idiom may be investigated before simpler forms), thus enabling output to be ranked and highest-scoring interpretations to be selected. Lastly, there are sets of codes specific to particular grammatical categories. Nouns are given semantic markers (e.g. 'entity', 'material', 'temporal'); and the arguments of verbs are defined in terms of their semantic roles, i.e. case relations (e.g. 'agent', 'benefactive', 'instrument'), their grammatical types (e.g. noun phrase, prepositional phrase), and their syntactic roles (e.g. subject, object, complement).

The processes of Analysis and Transfer call at various stages upon a corpus of 'grammar rules' — about 550 in the case of German (Slocum 1984). Each grammar rule specifies a rule for phrase structure constituency, some tests for well-formedness, 'construction' rules, and transfer functions. In addition Analysis and Transfer may refer to a corpus of transformational rules which may apply to phrase structures.

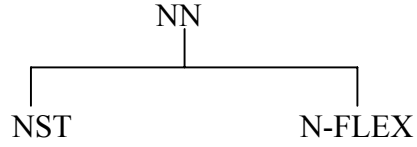
Dictionary lookup locates surface forms in dictionary entries and retrieves the canonical forms with their grammatical and semantic data. The next stage refers to the grammar rules and identifies those whose phrase-structure components apply to the sentence in question. The result is a structure which can then be tested for well-formedness by the relevant part of the same grammar rule. If the tests are satisfied, the relevant 'construction' rules are then applied, giving a more abstract structure which may be converted to a 'deeper' (SL-TL intermediary) structure by transformation rules. This structure is then passed to the Transfer stage where the transfer rules indicated in the grammar rules are applied to derive a surface syntactic structure. The final Generate stage produces TL output.

An example of a simple grammar rule for building nouns is in Fig. 27 (from Slocum 1984).

```
NN          NST          N-FLEX
0           1            2
(LVL 0)     (REQ WI)     (REQ WF)
TEST        (INT1CL2CL)
CONSTR      (CPX 1 ALO CL)
            (CPY 2 NU CA)
            (CPY 1 WI)
ENGLISH     (XFR1)
            (ADF 1 ON)
            (CPY 1 MC DR)
```

Fig. 27 — METAL noun building rule.

The first line indicates that a noun (NN) is to be formed from a noun stem (NST) and an inflected nominal ending (N-FLEX). The code (LVL 0) indicates this is a 'top-priority' rule. Each constituent (numbered 0 to 2) has individual conditions: the first element (NST) must be word-initial (WI) and the second (N-FLEX) must be word-final (WF). The TEST part of the rule states agreement restrictions: that when the values for the feature CL coded on NST are compared (INTersected) with those for CL on N-FLEX, there is at least one value in common. The appropriate tree can now be CONSTRucted:



where CPX will copy onto NN all the features associated with NST except for the allomorph (ALO) and inflectional class (CL), the first CPY will copy the grammatical number (NU) and case (CA) features from N-FLEX, and the second CPY will copy the WI feature from NST (i.e. stem form). The final part of the rule is invoked during Transfer, deriving the equivalent English noun from top to bottom by copying and adding appropriate features.

Other examples of rule applications are given by Whitelock and Kilby (1983). Tests of structures may, for example, check for noun and adjective agreement or the presence of a past participle *(gegangen)* with a finite verb form *(ist),* or the occurrence of a preposition followed by a noun phrase. A 'construction' rule might, for example, derive a predicative representation from the conjunction of a passive auxiliary (i.e. WERDEN) and an infinitive (e.g. *gehen),* by attaching the indicators of tense, voice, and mood carried by the surface form of WERDEN (e.g. *wird)* to the 'abstract' verb form GEHEN. The conditions specified for the arguments of verbs may be referred to at any stage during tests or during 'constructions'; any structure which does not satisfy the specifications, whether in terms of grammatical roles or case relations, may therefore be rejected. For example, the analysis part of a 'grammatical rule' for verbs such as *gehen,* is formulated as in Fig. 28 (Whitelock and Kilby 1983).

```
(DEXPR 12AL (VC MD)
  (COND ((SYNTAX)
          (COND ((AND (ACTIVE)
                      (NON-COMMAND)
                      (FRAME N NPAGT)
                      (FRAME NIL NIL LOC)
                 T)
                ((AND (ACTIVE) (COMMAND) (FRAME NIL NIL LOC))
                 T)))
```

Fig. 28 — METAL verb analysis role

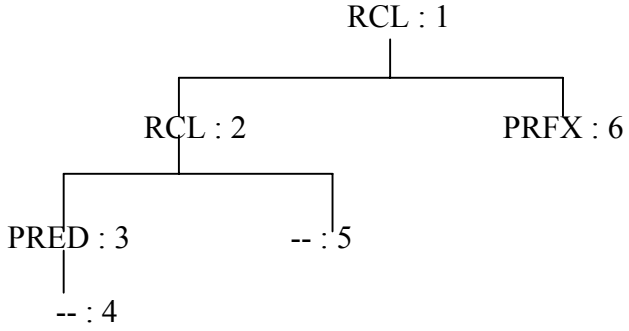
It may be interpreted as stating that if the verb is active and declarative (NON-COMMAND) then it has two arguments (AGT and LOC); the 'agent' must be a noun phrase (NP) in the nominative case, and the form of the locative depends on the particular verb. If, however, the verb is active and imperative (COMMAND) then there is just a 'locative' whose form is specific to a particular verb.

Transformational rules may be invoked in a grammar rule by a function XFM, which may be a combination of operations copying, adding, and deleting tree structures. For example, a 'construction' rule may include the following (Whitelock and Kilby 1983):

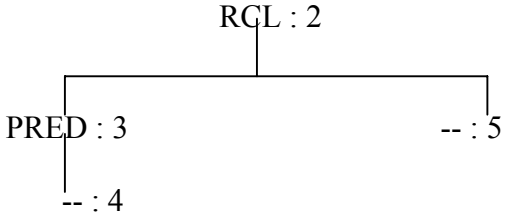
```
XFM (RCL:1 ((RCL:2 ((PRED:3 (—:4)) —5)) PRF: 6))
    (RCL:2 ((PRED:3 (—:4) (CPY 6 CAN) (ADD VC A)) —: 5)
    (ADD CLF) (ADD SPX)))
```

This has the effect of transforming the subtree 29(a) into the subtree 29(b):

(a)



(b)



and, in addition, copying the feature CAN ('canonical form') from PRFX ('prefix') to PRED, and adding the feature VC ('voice') with value A ('active') to PRED, and adding the features CLF ('clause final') and SPX ('separable prefix') to RCL:2.

After the analysis of SL sentences into 'deep structure' representations, processing passes to the Transfer component. The purpose of this phase is to convert representations into TL surface syntactic tree structures. As such, Transfer assumes most of the operations of TL synthesis leaving relatively little for the final Generation phase. An example of a transfer rule is another part of the grammar rule cited already (Fig. 28) which dealt with verbs of the *gehen*-type. It continues as follows:

((AND (ACTIVE) (NON-COMMAND) (PRED AGT LOO) (POL-ORDER (AGT) (PRED) (LOC)
((AND (ACTIVE) (COMMAND) (PRED LOC) (ROL-ORDER (PRED)(LOC)))

It has the effect of reordering the predicate-argument structure PRED AGT LOC into a form appropriate for TL (English) output, viz. AGT (subject) PRED (verb) LOC (location).

Lexical transfer involves access to the bilingual dictionary to identify the TL canonical form (e.g. *go)* corresponding to the SL canonical form *(gehen).* In cases where there is more than one TL form for a given SL form, the entry defines tests of local structure. For example the Transfer entries for German *in* would be:

(INTO (IN) PREP (GC A))
(IN (IN) PREP (GC D))

indicating that the German PREPosition '(IN)' may translate as the English PREPosition INTO if the G(rammatical) C(ase) of the German prepositional phrase is Accusative, and as IN if the GC is Dative. Finally, Lexical transfer accesses the TL monolingual dictionary in order to obtain the appropriate allomorph (morphological variant) for the English output, e.g. *went.* All that is left for the final stage, Generate, is to strip off the TL word string from the tree produced by Transfer.

There are clear advantages in employing the LISP formalism: ease of adding new rules, perspicuity of the grammar, and efficient parsing — it is estimated that a total of only 1000 rules may be adequate for sentences in technical texts. The METAL parser is described as "a variation on the Left-Corner parallel, bottom-up parsing algorithm", with resemblances to the Earley parser (for details see Slocum 1984). An important feature is the incorporation of the "some-paths" technique, in which the parser neither stops at the first 'successful' parse nor exhaustively explores all possible parses, but selectively pursues the 'most likely' ones. The parser is guided by linguists' judgments of preferred analyses, indicated by partial orderings of rules (i.e. LeVeLs, as in the example above). Nevertheless, on occasions, the parser will still fail to produce a complete analysis; METAL includes a "fail-soft" technique ("phrasal analysis") which looks for the longest constituent phrases recognised during analysis of a sentence and treats these as its parsing.

The efficiency of the parser is augmented by tying closely together phrase structure, construction, transformation and transfer rules in single 'grammar rules'. In this way, the METAL system seeks to ensure that a specific procedure is carried out at a later stage for a particular construct (e.g. tree representation). The objective is to avoid the difficulties of systems (like CETA and the earlier LRC system) which do not tightly control rule applications and which have to search a large corpus of grammatical routines for one which matches the structure under analysis. Tying rules of different stages together should eliminate the danger of inadvertently applying inappropriate or 'wrong' routines. On the other hand, the close linking of analysis rules and transfer rules would seem to reduce substantially the independence of the stages. Whitelock & Kilby (1983) argue that the inextricable coupling of lexical transfer and structural transfer reduces the modularity of individual stages of Transfer and increases the complications of writing and extending transfer grammars. Whatever the programming implications, these features do not seem to permit the characterisation of METAL as a pure 'transfer' system.

Since METAL is intended to be an operational system much attention has been paid to practical MT management (Slocum 1984). The system includes, therefore, facilities for validating

input (correcting misspellings and syntax errors), updating dictionaries, text processing before translation (distinguishing what has to be translated from diagrams, acronyms, formulae, etc., which do not, identifying sentence boundaries, detecting unknown words for dictionary updating, etc.), and text processing after translation (post-editing facilities, reconstitution of original format).

The METAL prototype German-English system was tested between 1980 and 1984 on over 1000 pages of texts in the field of telecommunications and data processing. Quality control was measured in terms of the amount of text which, after revision by professional post-editors at Siemens, did not entail changes in morphological, syntactic, or semantic procedures. Figures for "correctness" varied between 45% and 85%. An evaluation in 1980 (Slocum 1980) revealed that 83% of sentences of a 50-page text were analysed and translated correctly, 7% were analysed but not correctly, 9% were not analysed but provided with tentative 'phrasal analyses', and only 1% could not be analysed at all. Improvements in computer speeds on Lisp machines and improvements in rates of post-editing (by 1984 around 29 pages a day) suggest that METAL may have reached the goal of cost-effectiveness for an operational MT system (Slocum 1984).

The METAL German-English system is to be marketed as LITRAS by Computer Gesellschaft, Konstanz (West Germany), in "an office work-station package" (White 1985). METAL will therefore be the first of the advanced 'linguistics-oriented' transfer systems to be brought into commercial production.[9] In due course, it is anticipated that other language pairs will be added which are at present at various stages of development: German-Spanish, English-German, English-Spanish, and German-Chinese. Given the sophisticated nature of many features in METAL, its success in the market place will undoubtedly be watched with great interest.

13.5 Charles University, Prague (1976-)

After its earlier MT research in the 1960s (Ch. 6.10), the group at Charles University continued development of programs for automatic synthesis of Czech texts, largely within the framework of Petr Sgall's stratificational model. Particularly notable research was conducted in the area of text cohesion, topic-comment structures and anaphoric reference (cf. Sgall *et al.* 1973). MT research as such was revived in 1976 with a series of fairly modest experimental systems for English-Czech translation (Hajičova & Kirschner 1981, Kirschner 1984).

The linguistic group of the Department of Applied Mathematics at Charles University began its first experiment in close cooperation with the TAUM group in Montreal (Ch. 13.1). A relatively limited English parser was designed which resembled the TAUM analyser in its use of the Q-system formalism but which differed in that it produced a dependency structure analysis. Output from the parser was converted by a transfer program for input to the already developed system for random generation of Czech sentences. The system was tested on a small sample of journalistic texts on economics.

The second experiment introduced greater lexical and semantic complexity. Abstracts selected from the INSPEC database in the highly specialised field of microelectronics were taken for the corpus; the aim was to extend and refine the grammar, to provide more extensive morphological analysis, to tackle the complex problems of English noun compounds, and to develop a device for directly converting 'international' vocabulary into Czech forms without recourse to dictionaries, e.g. *application* into *aplikace, philosophy* into *filozofie,* etc. (Hajičova & Kirschner 1981). In addition it was hoped that the system might act as the front end of a natural language understanding system, TIBAQ (Text- and Inference-Based Answering of Questions), an AI project under the direction of Petr Sgall (Hajičova & Sgall 1981). In general, the experiment was seen as the first stage of an exercise in adapting the theoretical research of the Prague group to a practical application. Strict stratification of analysis and synthesis was compromised by

[9] For a description of the commercial system see chapter 15 in Hutchins, W.J. and Somers, H.L. *An introduction to machine translation* (London: Academic Press, 1992)

orientation of procedures to particular source and target languages and to the specialised subject field.

The third experiment, beginning in 1982, represented a further step towards practicality (Kirschner 1984). Analysis does not go as far as in the TIBAQ system, remaining at a relatively 'surface structure' level; the 'random generation' program for Czech has now been abandoned in favour of a synthesis program linked more closely to the output, and the former single bilingual dictionary has been split into separate ones for analysis and synthesis. Finally the program for converting 'international' vocabulary has been moved from transfer to the stage of morphological synthesis. Starting from a fairly orthodox 'transfer' strategy, the system has gradually incorporated transfer components into analysis stages, which thereby have become more TL-specific; the Charles University group have concluded that "universal analysis ... is not available at the present time". The group has also adopted an orthodox approach to semantic analysis: semantic features partially organised in verb 'frames', features specific to the subject domain and features specific to text structures of abstracts. Some ambiguities are left unresolved, e.g. English prepositional phrases where an equivalent Czech structure would be equally ambiguous. As in most other systems, problems of inter-sentence coherence and ambiguity have not been tackled yet.

The group has continued with the Q-language formalism, considering that the advantages of its simplicity and transparency outweigh its limitations. It has continued also with the TAUM system of English morphological analysis. The emphasis in the syntactic analysis program has been on dealing with the complexities of English nominal compounds, which are of course prominent in abstracts. Analysis results in a dependency tree representation (of the predicate and argument type found in TAUM). In synthesis this is restructured for Czech word order, Czech lexical items substituted for English, and Czech morphological rules are applied. The group recognises that the system is still at a very early stage and "much remains to be done". It is, however, a good example of the greater assurance and realism with which MT projects are now being undertaken.

13.6 Logos Corporation (1982-)

The earlier involvement of the Logos company in MT had been the development of systems for governmental and military applications (Ch. 12.2). In 1982 the Logos Corporation introduced their 'Logos Intelligent Translation System', a commercial product aimed at translation agencies. The first version demonstrated was the German-English system; an English-German version has also been announced (a prototype has been installed at Wang in the United States), and work has been reported on a French version (Lawson 1984). Logos are developing their systems on a mainframe and design programs to be machine-independent, running on small computers as well as large ones[10]. However, at present, the German-English system is available only on the Wang OIS (Office Information System) word processor and the Wang VS minicomputer.

The system integrates machine translation into a word processing environment; translators are able to adapt the dictionaries to their special needs, run the Logos translation program, and revise the results, all at the same computer (or word processor) terminal (Hawes 1985). It is claimed that in the larger 256K Wang configuration the system is capable of translating more than 20,000 words in 24 hours, "enough to meet the needs of three translators engaged in post-editing" (Staples 1983). In 1984 the German-English system cost about £6500 ($10,000) to install, plus a charge of £200 a month for up to 10,000 words of 'usable output', and then £16 for every additional 1,000 words (Lawson 1984).

The German-English system comes with a basic dictionary of over 100,000 entries; to these can be added the terminology specific to translators' needs, using an interactive dictionary

[10] For later information see: B.E.Scott, 'The Logos system', *MT Summit II, August 16-18, 1989, Munich. Second and complete edition, final programme, exhibition, papers* (Frankfurt a.M.: Deutsche Gesellschaft für Dokumentation, 1989), pp.174-179.

compilation system (ALEX), which asks questions concerning the syntactic and semantic properties of words being entered and ensures that their coding is compatible with entries already in the dictionary. In addition, users can specify subject fields for particular entries, thus ensuring that polysemes are translated appropriately (as far as possible) in accordance with the subject context.

In basic strategy the German-English system retains the characteristics of the earlier LOGOS, although now it is a true 'transfer' system, with language-independent programming and separation of SL and TL data files. Much is made by the manufacturers of the way in which "semantic information (is) integrated into the translation algorithm." During the 1970s Logos developed and refined a Semantic Abstraction Language (SAL), "a hierarchical tree-structure language into which the Logos system translates every natural language string before it begins its parse." As a result "semantic information is ... present at every point of the analysis, available, as needed, for resolving ambiguity at whatever level — lexical, syntactic, or semantic." (Staples 1983). It is not clear whether this approach differs substantially from the practice in many MT systems of assigning semantic features to dictionary entries and using this information to resolve lexical ambiguity and syntactic ambiguities during analysis routines. It would appear, nevertheless, that Logos has paid greater attention to the establishment of semantic classifications than previous MT systems. It is possible, therefore, that SAL has elements of a genuine interlingual character (since the Logos Corporation has researched a wide variety of languages). Interlingual elements are to be expected in transfer systems, the line of demarcation between them is often blurred and for this reason there are grounds for Staples' (1983) claim that the "Logos algorithm entails an integration of both transfer grammar techniques and the interlingua approach".

Practical experience of Logos in a working environment has so far been relatively brief. One agency (Tschira 1985), which installed Logos in December 1982, estimated in 1983 that the quality of output was good enough "for preliminary information purposes, without post-editing, for about sixty to eighty per cent of the text". Tschira stressed however, that output can be deceptive; very often, considerable post-editing is still required to produce high-quality texts from MT output. Nevertheless, translation agencies in Germany report favourably about the capacity of the German-English system to increase their productivity (by up to 100% and at least by 30%), while also showing improvements in the quality and consistency of their translations (Lawson 1984). Logos was found to work best with highly specialised texts for which the agency had input its own terminology; it was least satisfactory for general correspondence and texts with comparatively few technical terms and for interdisciplinary texts.

An example of an unedited Logos translation from German is the following (extract from Tschira 1985):

> A field-by-field control of the occurrence of the different fields is luckily not necessary. There are groups of fields which occur always commonly, or occur not and which we sucked therefore too. Field-groups are able to combine. In fact, it suffices to relate the control mentioned above (depending on account and posting key) to field-groups instead of on fields what a considerable simplification means.

(German original:

> Eine feldweise Steuerung des Auftretens der verschiedenen Felder ist glücklicherweise nicht nötig. Es gibt Gruppen von Feldern, die stets gemeinsam auftreten oder nicht auftreten und die wir deshalb zu sog. "Feldgruppen" zusammenfassen können. Tatsächlich genügt es, die oben erwähnte Steuerung (in Abhängigkeit von Konto und Buchschlüssel) auf Feldgruppen anstatt auf Felder zu beziehen, was eine erhebliche Vereinfachung bedeutet.)

As this example shows there are obviously still problems with certain linguistic structures (the *was* construction in the last sentence), as well as familiar problems in distinguishing between abbreviations (e.g. *sog.* for *sogenannt(en)* in this passage) and the sentence marker, which can result in nonsense. Further such problems are: "the splitting of some phrases in parentheses,

mixtures of letters and numerals" and some paragraph numbering interpreted as dates. These annoying errors are surprisingly difficult to circumvent. As for the strictly linguistic deficiencies, the Logos Corporation is itself aware of many of them (it admits, for example, that it has "only just begun to attack such problems as ellipsis and anaphora"), but Logos promises to continue research on improvements which it can pass on to its customers. As always, time will tell how successful the Logos system will be in the future.