The METAL System. Status 1989.

Thomas Schneider

Siemens AG
Munich, FRG

1. GENERAL

By now it has become a truism that machine translation is no more (and no less) than a powerful tool for human translators. It is acknowledged that MT will not replace qualified translators - but that it is indispensable in the face of the enormous quantities of texts that need to be translated. Furthermore, it is understood that MT systems cannot, and are not intended to, translate literary texts. This restriction includes any other text which does not simply convey factual information but expresses its meaning by intricate nuances of style, metaphoric descriptions and the use of semantic elements "hidden between the lines". Consequently, any attempt to evaluate MT systems on the basis of such texts is grotesquely inappropriate, naive and irrelevant.

By contrast, MT systems are intended for applications in the area of information transfer, for the translation of texts in which the aesthetic gratification of the reader is secondary and in which the main goal is to convey a precise information content to readers who are usually quite knowledgeable within the subject field. As a rule, the use of an MT system usually yields the best results if the application is limited to relatively narrow subject fields and if there is sufficient volume of similar text types.

The growing internationalization of economies, the interdependence of research and development as well as industrial processes beyond political and linguistic boundaries have already caused an almost exponentially increasing volume of translation. The integration of the European market in 1992 will certainly require an even greater degree of information exchange across several languages.

Needless to say that no economy can afford to employ half of its population as specialized translators. Since it is equally impossible to train whole populations to a near-native competence in several foreign languages, the use of efficient MT systems seems the only solution to the problem of rapid information transfer.

2. HISTORY

A noticeable shortage of qualified technical translators forced Siemens in the late seventies to look for a means to increase the productivity of the in-house translators and reduce the time required for the translation process. The documentation of complex products such as telephone switching systems may run into hundreds of thousands of pages, and a delay in product delivery on account of unavailable documentation in the language of the target country may easily mean the loss of markets.

To solve these problems, the first step was to design a multilingual terminology data base (TEAM) which at present contains more than 2 million records in up to eight languages. The second step was the search for a suitable MT system.

After experiments with some of the systems which were offered commercially, it became obvious that a broader linguistic analysis was required for our intended applications. So in 1978, Siemens entered into a cooperative agreement with the University of Texas at Austin to develop the METAL system . At the time it existed as a rather unwieldy prototype but it was based on more than twenty years of theoretical research at the university. Originally, the cooperative effort was expected to be a long-term research project, but fortunately progress came more rapidly than anticipated. Nevertheless, the complexity of the task of natural language analysis and generation becomes obvious if one considers that the time span between operative prototype and useful product has been a long ten years.

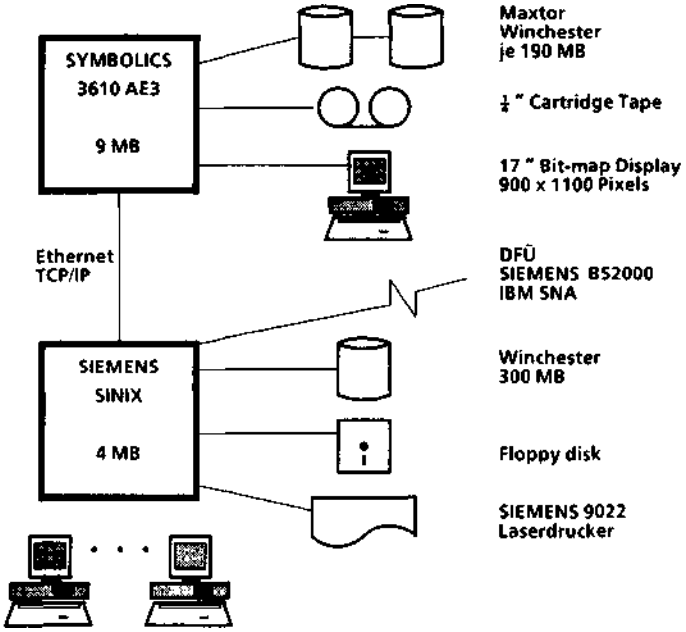
Today, the German to English version of METAL has been implemented in more than a dozen installations; other language pairs with English, French, Dutch and Spanish modules will be available by the end of the year.

3. HARDWARE

METAL runs on a hardware package consisting of several SINIX-based translator work stations and a Symbolics 36-series LISP machine. In accordance with user requirements, the translator work stations can be anything from a Siemens MX-2 or MX-300 to an MX-500.

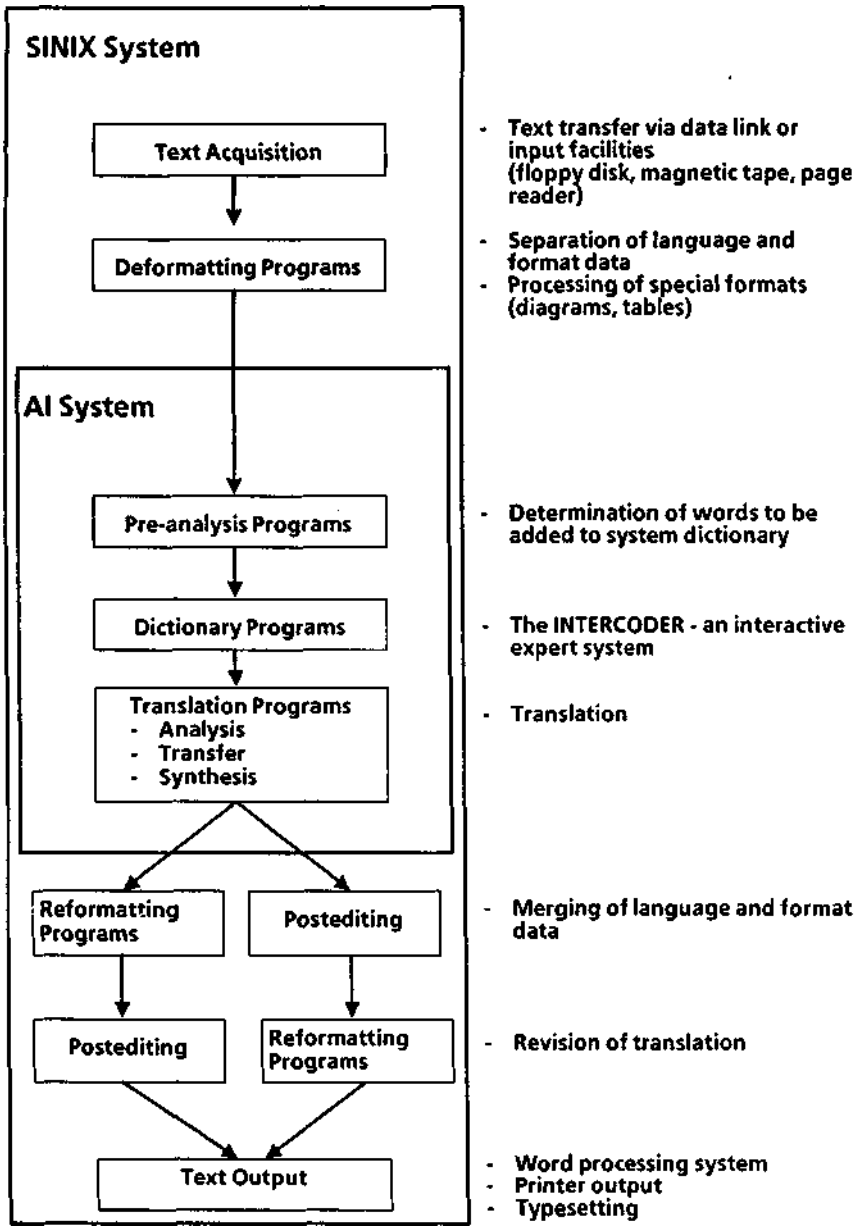
Since the METAL throughput is more than 200 pages a day, several translators can be supplied with sufficient text for post-editing. The post-editing functions do not interfere in any way with the actual translation process so that batch translations can run in the background while previously translated files are being corrected.

The hardware configuration looks as follows:



An operative productive system needs to do more than simply translate individual sentences entered from the keyboard. Most of the texts which have to be translated quickly and are of great volume such as e.g. technical documentation are heavily formatted. In some texts, more than half of the characters on a page may be non-translatable material, notably flow charts, diagrams, tables and various control characters for format and layout. It would be highly uneconomical to manually extract the text portions to be translated and afterwards manually reinput them. That would not only be expensive but it would also invite errors in the additional reformatting tasks. Therefore, METAL was integrated into a chain of processes, from text acquisition via automatic deformatting and translation to automatic reformatting procedures. A text is usually received in machine-readable form, by file transfer, floppy disk exchange or from a font reader. Several programs running on the SINIX system check the pages for tables, graphs etc and mark them. They identify the text portions to be translated and generate a mask of the page. The individual translation units, usually sentences but in the case of headlines or table entries also single words, are automatically recognized, numbered consecutively and extracted from the page mask. They are written into a file and transferred to the LISP-machine for translation.

After translation, the file of the target language text units is returned to the SINIX system for post-editing. Here, the translators can choose whether they want to postedit an interlinear version which groups single source language/target language units sentence by sentence, or work on two windows with source and target text, or whether they prefer a target language output that has already been reformatted. In the former cases, the posteditors would start the reformatting program after having made their corrections. At the end, the target language text is available with all the formatting information and with the same layout as the original. The whole process of a METAL translation can be visualized as follows:



In designing METAL, great care has been taken to keep the system modular so as to permit the addition of new elements or the modification of existing elements without major ill effect on the other components. Unfortunately we are still faced with the dilemma that there is not a single linguistic theory available that would describe even a single language unambiguously and completely. Even though various new formalisms have been proposed it seems that none of them have been tested on large enough segments of a language to prove their viability. Therefore grammatical approaches will by necessity have to be eclectic.

METAL employs a modified transfer approach rather than an interlingua. If a meta-language were to be used for translation purposes it would need to incorporate all possible features of many languages. That would not only be an endless task but probably a fruitless one as well. Such a system would soon become unmanageable and perhaps collapse under its own weight. If, on the other hand, the meta-language were reduced to a manageable degree of abstraction then too much information necessary for a faithful translation would be lost. In METAL, it has turned out that - with minor adaptations - the identical source language analysis module could be used for the translation into several target languages.

Basically, METAL uses phrase-structure rules which are augmented by tests on individual constituents and their interaction. The rules are recursively applied; that means that their number can be kept low in regard to the degree of coverage of possible surface structures. The rules are indexed to make processing more efficient. The most commonly applied rules, e.g. those for word level morphology and for frequently occurring basic structures, are defined as the most basic level. Higher level rules deal with more complex or even ungrammatical structures. If a surface structure can be interpreted using lower level rules then the more complex and less likely rules are disregarded. If no interpretation is possible with the lower-level rules then incrementally higher levels of rules are added to the lower level rules to attempt an interpretation.

METAL uses a prioritized chart parser. Unlikely paths are eliminated via preferential weightings calculated from lexical and grammatical data. Based on heuristics supplied by the linguists, this approach not only increases processing speed but usually yields the best interpretation for transfer to the target language. If no sufficiently plausible interpretation of the complete sentence can be reached, the system will go into a fail-soft routine and output a translation of the individual phrases it had been able to interpret. In many cases, the output is still a grammatically correct translation of the original sentence; otherwise the posteditor is called on to correct the output.

5. LEXICON

In accordance with the principle of modularity, METAL operates on both monolingual lexicons and one transfer lexicon for each language pair. The monolingual lexicons contain morphological, syntactic and semantic information needed for the analysis and/or the generation of a language - irrespective of the target language. The transfer lexicon provides a link from the source language to the target language, indicating in which contextual environment and in which subject field a given source language entry should point to a certain target entry.

The advantages of such a structure are obvious. The extensive grammatical information contained in the monolingual lexicons needs to be stored only once even if a given lexicon entry has, depending on context, many different translations in the target language. This separation of monolingual and transfer information reduces storage space and saves coding time. Moreover, the monolingual lexicons can be used in various language pairs.

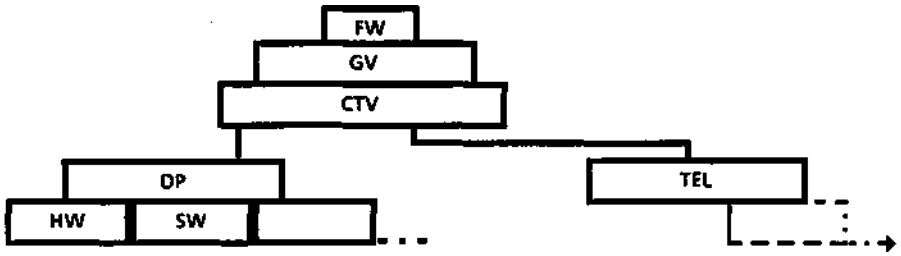
A translator working with the system needs to be able to extend and modify his lexicon. While it is not recommended to alter function word entries (they are too closely linked to the grammar) a translator may code all other word classes including verbs. Even though the grammar rules are not accessible to an end-user, the transfer lexicon permits significant syntactic transformations. On top of being able to specify transfers on the basis of the instantiation of frames, the presence of arguments of a certain semantic type or of a specific canonical form, the user can influence the target structure considerably. Surface structure active phrases may be turned into impersonal constructions, roles of arguments can be changed, complements can be converted, elements can be added or deleted etc.

All these options are available to the translator via a menu-driven INTERCODER which has proven its usefulness in reducing coding time by a factor of ten. The INTERCODER guesses at the morphological and syntactic behavior of new lexicon entries and proposes the necessary coding; the missing pieces of information are inferred from a set of rules and partial information already contained in the lexicon.

For productive application, the lexicon of an MT system has to be sizeable. However, an unstructured voluminous lexicon can cause more problems by introducing unwanted ambiguities than would be gained by having extended text coverage. It should also be kept in mind that MT systems are intended to translate specialized texts and not general language or literary texts. In technical texts, for example, the percentage of general vocabulary is quite limited while subject-specific terminology abounds.

This makes it possible to modularize the lexicon and assign preferences for specific transfers based on the subject area of the text to be translated.

The METAL lexicon is organized as follows:



There are modules for function words (FW) like prepositions, determiners and conjunctions, for general vocabulary (GV) and for common technical vocabulary (CTV) organized in a tiered hierarchy. From the next level down, each end-user can define and structure his own modules and tailor them to his specific application. For in-house applications in Siemens, there are for example modules like Data Processing (DP) with submodules software (SW), hardware (HW) etc. Furthermore, it is possible to define transfers on the basis of a specific customer, a specific product or a target country. The German noun "Lastwagen" would automatically appear as "truck" in a text for the USA and as "lorry" if intended for Great Britain.

Before a translation run is started, the modules appropriate to the subject area of the text are defined. If the syntactic and semantic criteria for the selection of a lexicon entry are met and there are several candidates for transfer, then the one tagged for the subject area of the text or tagged for a hierarchically closer module is chosen. This assures that highest priority is assigned to subject-specific transfer.

6. USER EXPERIENCES

Machine translation is a recently evolved technology and as such vulnerable in its status. A new technology can easily be proven inadequate or even useless if the intended recipient refuses to accept such a system or insists on applying it in unsuitable ways.

Therefore, the introduction of a machine translation system into an existing organization, be it a large industrial company or a translation bureau, requires several steps. First of all, end-users must have a clear picture of what can be expected from an MT system and what is beyond the scope of today's technology. Inappropriate use will only lead to frustration.

Once the conditions for the installation of a system have been assessed, i.e. translation volume, suitable types of text, hardware environment, and a positive decision has been reached, the organizational setup needs to be discussed. From which sources does the translator receive the original texts? Is there a possibility to influence the style of the original, i.e. can certain guidelines be imposed in regard to complexity of verbal expressions? And can the customers be persuaded to use standardized formatting and layout routines so that the tasks of deformatting and reformatting can be simplified?

Machine translated output cannot be expected to improve the original text. Input with grammatical errors or semantic absurdities will not be turned into poetry. Sometimes it may be advantageous to correct an input text first - especially if the same text is to be translated into different languages - rather than relying on a post-editor to correct the mistakes.

Translators using MT systems need an introductory training. It should focus on a general introduction to the system's structure and the tools it provides. Equally important is a first training in the different work techniques that an MT system requires. Provided that the reader of a target document is not concerned with intricacies of style, the post-editing phase of a machine output can focus on changing this output to an acceptable version with the least effort. Certainly, a given version could be rewritten in various ways, sometimes with a gain in quality but sometimes also with simply an idiosyncratic change of style without improvement of quality.

Postediting machine output is different from revising a "human" translation. While the machine will make "severe" errors in syntax (e.g. PP attachment) or semantics (ambiguous structures) a human translator will make fewer but random and less predictable errors. Usually, it takes a translator several weeks of practical work with an MT system to be able to anticipate the common errors and look for them.

Experiences with the METAL installations have been positive and can be summarized as follows:

- Translators as well as management have to understand that an MT system is not a substitute for a highly qualified translator but a powerful tool.

- For the use of METAL, an initial training period of one week has been sufficient. A second week of training after a few months answers questions which arose during the actual productive application. After that, consultation on a case by case basis seems adequate.

- During the first few months of operation, the translators' productivity will actually decrease. There is the initial overhead of bringing the lexicon up to a level where it covers most of the specific texts to be handled. Also, translators have to get used to the different work technique and acquire skills in lexicon building and system administration.

- After this initial period, which may vary from a few months to more than a year, users have reported considerable gains in productivity and a decrease in turn-around time. It appears that under favorable conditions a productivity gain of a factor of 2 to 3 is a realistic goal.

- The consistency of terminology throughout all documents has been viewed as a qualitative improvement in the target version which could not have been achieved with human translation.

While the present METAL system has stood the test of practical application, development will continue. Additional language pairs will be developed, the integration into various office environments will be stream-lined, and further research will center on add-on semantic components and linking METAL to data bases, expert systems and teaching/learning systems.