[From: Internat.conf. *Methodology & Techniques of Machine Translation*, British Computer Society, Cranfield, 13-15 February 1984]

AMPAR and NERPA Systems of Machine Translation
Some Specific Features of Software and Technology

B. D. Tikhomirov

The USSR Translation Centre for Scientific and Technical Literature and Documentation, Moscow, USSR

INTRODUCTION

When creating the AMPAR machine translation system (from English into Russian), and the NERPA machine translation system (from German into Russian) at the USSR Translation Centre, unified software oriented to commercial operation has been developed [1],[2]. Much effort has been made to facilitate the system linguists' work in the course of interaction with the machine translation system throughout its development and operation [3].

Main specific features of software and technology are as follows:

- decomposition of the translation process into a number of stages
- use of a specialized programming language for specific algorithms of machine translation
- two-stage organization of information files
- availability of topical information files for various subject fields
- use of a specialized process control language to specify information file input/output instruction or information file processing modes
- possibility of obtaining system operation results from any stage in form convenient for their analysis
- possibility of reorganizing the structure of the system (creation of different version to select the most efficient one)
- possibility of system generation with the prespecified set of functions and topical files.

SOFTWARE STRUCTURE AND LANGUAGES

Since the AMPAR and NERPA systems use inter-editing of intermediate results and post-editing of target texts to increase the translation quality the process is divided into a number of stages: initial processing, inter-editing, automatic translation, post editing and target text printout. In its turn, each stage may be split into a number of steps, each implementing some essential algorithm of the system.

Steps liable to frequent alterations consist of the step subroutines (schemes), each executing some specific analysis and synthesis algorithm.

The step scheme consists of statements, each executing a certain linguistic operation.

To achieve the maximum participation of the linguistic support designers in the creation and modification of the specific algorithms of the system a special programming language oriented to a special data structure has been developed. Information files contain dictionaries, tables, source and target texts and a file of information location to store and modify information on a word or word combination required in the course of translation.

Throughout translation the schemes are called from the external memory into the main memory and the specific algorithms are implemented using an interpreter program.

Subroutines which are not modified during the system operation (the source text input routines, dictionary look-up routines, monitor etc.) are developed by the systems programmer using the assembly language of the EC Computer System to minimize their running time.

Use of a modular principle for creation of the software structure in which an algorithm is divided into sufficiently small algorithms and each specific algorithm is program-implemented as a separate module ensures a significant simplification of the programming process. The modular principle is also applied to the information files (subject field dictionaries, tables).

Due to a relative independence of modules, the software system obtains enough flexibility to allow easy updating of programs and information files by development and inclusion of new modules, by updating new modules (or deleting) or changing their sequence, i.e. it is relatively simple to generate various system versions throughout system debugging and updating when in operation.

The main modes of the system operation are: file maintenance, monitoring and work ones.

The first mode allows you to create, add, update and list files that are used to process the source text in the monitoring and work modes. To maintain files, a utility package independent of work routines has been created. The work mode is used for mass processing of texts by a work version in which the sequence of all system element operations is fixed and the work data files are employed.

Choice of subject field dictionaries is determined by the control information accompanying each text. In addition to the target text language intermediate information concerning some errors which occur when processing the text, words not found in dictionaries of the system, contradictory situations, etc., is provided in the work mode.

In some cases the intermediate information is sufficient to determine the nature of faults. In other cases, to obtain additional information, the system linguist can have the faulty text fragments reprocessed in the monitoring mode. The same system version is used as in the work mode but a selective listing of the system operation at any prespecified section of the text with a high degree of detail (with an accuracy of up to the operation of an individual module statement) is set up using a special directive.

The machine translation system employs files of two types: operational and upgrading. Originally, the upgrade files are fully identical to the operational files. On the basis of information about the nature and location of an error, the linguists correct the specific algorithms and/or information files which constitute part of the upgrade files as well as create new versions of the individual schemes. After changes have been introduced into the upgrade file, an upgrade system version is generated in which the correctness of the introduced modifications is checked out in the monitoring mode.

As a result the system linguists have an opportunity:

- to participate directly in development and debugging of modules implementing specific algorithms of the system
- to obtain information about missing dictionary entries and about typical errors occurring during text processing
- to promptly localize an error and its nature
- to create versions of the system without disrupting the commercial operation of the system and to maintain its operation version intact; each new version may include new and/or modified existing program and information modules as well as changed order of their operation

- to verify operation of created version using textual material of a large volume in order to select the most efficient version and to introduce it into the operating file as the operational version

- to monitor the state of information files and program modules and to ensure their rapid updating.

The specialized process control language (PCL) implemented as a set of directives according to which the program modules of the system perform certain operations is used as a language for interaction with the system and organization of the intermodular links inside the system. The set of directives called an instruction is entered into the computer before solving a task or in the process of its executing. Besides, each program module may issue an instruction that is to be processed by the called module. PCL allows us to alter the standard order of the programs, to specify different printout modes and to correct the information files.

TWO FORMS OF USING MACHINE TRANSLATION SYSTEMS

The AMPAR and NERPA systems have been developed as multi-functional machine translation systems (MMTS) used in large translation organizations and must provide:

- translation of polythematic documentation

- adjustment to any form of input information including information retrieved from data banks or as software descriptions on magnetic tapes

- input and processing of information having a sophisticated structure

- possibility of inter- and post-editing in the interactive mode

- possibility of conducting broad investigations in the field of lexicography

- possibility of prompt correction and upgrading of information files

- generation of MMTSs with prespecified sets of functions and subject field files.

It is obvious that MMTS must be maintained by the system linguists and programmers who are fully aware of the particularities of the multi-functional machine translation systems. To hand the multi-functional machine translation system over to any other translating organizations it is expedient to generate a simplified version of the multi-functional system, specialized MMTS.

Any specialized MMTS must have fewer functions than any general-purpose MMTS; its information files must be oriented to specific subject fields. Provisions must be made to maintain the system by personnel of not very high qualification in any computer installation.

TWO-STAGE ORGANIZATION OF INFORMATION FILES

Such factors as speed of translation and ease of the linguists' interaction with the system during its development and operation are of great importance for the commercial MMTSs.

Obviously it is impossible to select the information representation form that might be equally suitable for a human being and a computer. As a means of settling this discrepancy, provision is made to ensure a two-stage organization of the information files in the AMPAR and NERPA systems. This implies that two files, a linguistic information file (LIF) and machine information file (MIF), as well as converters of information from one form to another are created. Both files are stored in the computer memory, amended and updated using a mini-computer. Each array is divided into a number of subfiles by functional and technological indices.

Information is stored in LIF in the form of words and word combinations in the source and target languages with the aid of the convenient mnemonic codes and a language for programming the specific algorithms. LIF consists of subfiles of separate words and word combinations, unambiguous words and grammar.

In MIF the information coding technique and information arrangement is selected with due regard for the most efficient data processing by the system programs. MIF contains subfiles of the source and target dictionaries, homography tables, word combinations, subprograms for translation of any compound word combination and ambiguous words as well grammatical subprograms.

Each element in LIF (a dictionary entry or scheme) is accompanied by Keys indicating that the entry belongs to a specific subject field. The said keys also indicate a creation date or entry update date, etc. A service routine periodically selects an entry matching against the prespecified key from LIF and forms an update file to be further handled by the converter, mapping each entry of LIF onto one or several entries of MIF and writes it into the MIF subfile.

Use of the two-stage organization of information files allows us:

- to simplify considerably the linguists' work with the system due to elimination of information having a non-linguistic nature, and also due to the fact that the linguists are relieved from taking into account all links that arise when information on each word is embedded into several subfiles of the system
- to decrease the number of errors in MIF which are associated with coding mistakes and data transfer to the machine-readable medium
- to form several subject field MIFs using a single LIF
- to exchange information with automated dictionaries.

SUBJECT FIELD INFORMATION FILES

Each element of a linguistic information file in an MMTS may consist of several fields related to different subject fields and containing various information. Each field like this has a subject field code.

While MIF is formed, MIF entries that have the identical subject codes are grouped into separate field blocks (dictionaries, tables) in each MIF subfile. The aggregate of blocks related to one subject field is called the subject field information file (SFF) . A special role is played by the base information file (BIF) of the common vocabulary file.

When translating a text related to a certain subject field, BIF and one or several SFFs are used. The order of their operation is specified when post-editing. As a rule, BIF has the lowest priority. At each stage, the specified subject field blocks associated with the appropriate MIF subfile are selected, and information from these blocks is used in accordance with the prespecified order.

For instance, the words from the source text are first matched against the subject field dictionary having the highest priority at the stages. Then the words missing in the subject field dictionary are matched against the base dictionary (common vocabulary dictionary).

The main advantage of the modular generation and use of information as compared to utilization of a single file, from our point of view, lies in the complete independence of all subject field files. This, allows us to use several independent groups of linguists who work on different subject fields when the system is upgraded. Besides, making the subject field more narrow simplifies creation of files due to decrease of lexico-grammatical homography and lexical ambiguity, which enables us to enlist specialists unfamiliar with the particularities of the system linguistic support for the creation of files.

An important consideration is the fact that - despite the enlargement of the entire information file (due to duplication of information in various SFFs) when translation takes place - only a part of the common file is read into main memory of the computer, decreasing, thereby, the translation time.

CONCLUSION

The program complex originally developed for machine translation from English into Russian due to the general-purpose features used in it appeared also applicable, on the whole, for machine translation from German into Russian. The specific features of German required insignificant modifications of the program complex, development of some new program modules, alteration of source dictionary organization and introduction of new stages.

Exploitation of the AMPAR and NERPA MT systems confirmed correctness of the adopted software and technological decisions.

REFERENCES

1. Yu. N. Marchuk, B. D. Tikhomirov & V. I. Shcherbinin.
   Ein System zur maschinellen Uebersetzung aus dem Englischen ins Russische.
   Sonderdruck Darmstadt. 1982, pp 319-336.

2. Yu. N. Marchuk, A. N. Vlasov
   Nekotorye printsipy avtomatizatsii perevoda s nemetskogo yazika na russkii
   Zeitschrift "Fremdsprachen", 1980, No 2, pp91-99.

3. I. I. Oubin, B. D. Tikhomirov,

"Machine Translation Systems and Computer Dictionaries in the Information Service. Ways of Their Development and Operation".
Proceedings of the ninth International Conference on Computational Linguistics, Prague, 1982, pp 289-294.