METHODOLOGY AND SOFTWARE DEVELOPMENT FOR APPLICATION IN INFORMATION SCIENCE

S. Perschke, H. Fangmeyer

Introduction

The present activity of CETIS in the field of scientific information and documentation is directed towards using experience in information science to create operational systems.

By 1970 three such systems appeared to be well enough defined to warrant starting intensive implementation:

1. A generalized software package and programming language for natural-language processing (SLC-II).
2. A fully automatic information storage and retrieval system (FAIRS).
3. A Russian-to-English machine translation system.

A valuable contribution to the realization of the projects was given by the group's staff, in particular Mr. Fassone and Mr. Geoffrion. Some of the computer programs were prepared under contract by a computer software company (ITALSIEL, Rome). At present, these projects are at various stages of completion, and naturally every effort was made to accelerate the first of them, which should become the software basis for the others.

SLC-II

It was possible to conceive a generalized software for applications which appear to be widely different (such as translation and information retrieval) by treating all the processes involved as problems of translation and communication of natural and artificial languages. Further, it was possible to break down all these processes into a series of recurrent basic functions. Each one of these basic functions employs an algorithm, a dictionary and a grammar. For some of the functions, the algorithm could be defined as invariant with respect to the different source data, grammars and dictionaries, and in this case the SLC programming language is non-procedural and is used for coding the grammars and dictionaries, while in other phases the algorithms vary depending on the application and the methodology employed, and in this case the SLC programming language is algorithmic.

If one considers a few characteristic applications in automatic information and documentation, the similarity of functions becomes striking: automatic language translation; automatic indexing, information retrieval for SDI and updating of the data base; automatic query formulation and information retrieval.
In all applications, the first part of the operation is basically the same.

A1) Source text input: The source text — some foreign-language text to be translated for MT, a new document for indexing or a query for retrieval — is scanned and those character sequences which are elements of the source language ("word item") according to the associated dictionary and grammar (D_1 , G_1) constitute the arguments for the subsequent phase (A2). The algorithm is invariant.

A2) The word items are looked up in the source language morphological dictionary (D_2). In parallel, if the source language has inflections, a morphological analysis of the words and, optionally, the segmentation of compound words are performed. Each word item is represented by a lexical code and a morphological classification. Using the lexical code, the source language dictionary entry containing all information about the word, relevant for a given task, is located. Phase A2 is invariable.

A3) In principle, the object of this phase is to give a formalized description of the contents of the source text according to the grammar and the dictionary entries associated with the words. Largely, it covers syntactic and semantic analysis and the resolution of lexical and structural homographs. The objectives, the methods applied and the degree of sophistication are quite different from application to application. Therefore, no invariant algorithms for A3 (or for A4 and A5) could be defined and the programs are written in SLC (optionally, one also can use PL/1 for these phases).

A4) The transfer, in principle, has the function of substituting target language features for source language features in all instances where a metalinguistic description is not achieved and one has to take the equivalence between two languages as a basis.

A5) The target text generation is the inversion of step A3, i.e., one starts from a metalinguistic description of the target text and produces a string of items consisting of a lexical code and the definition of the inflectional form.

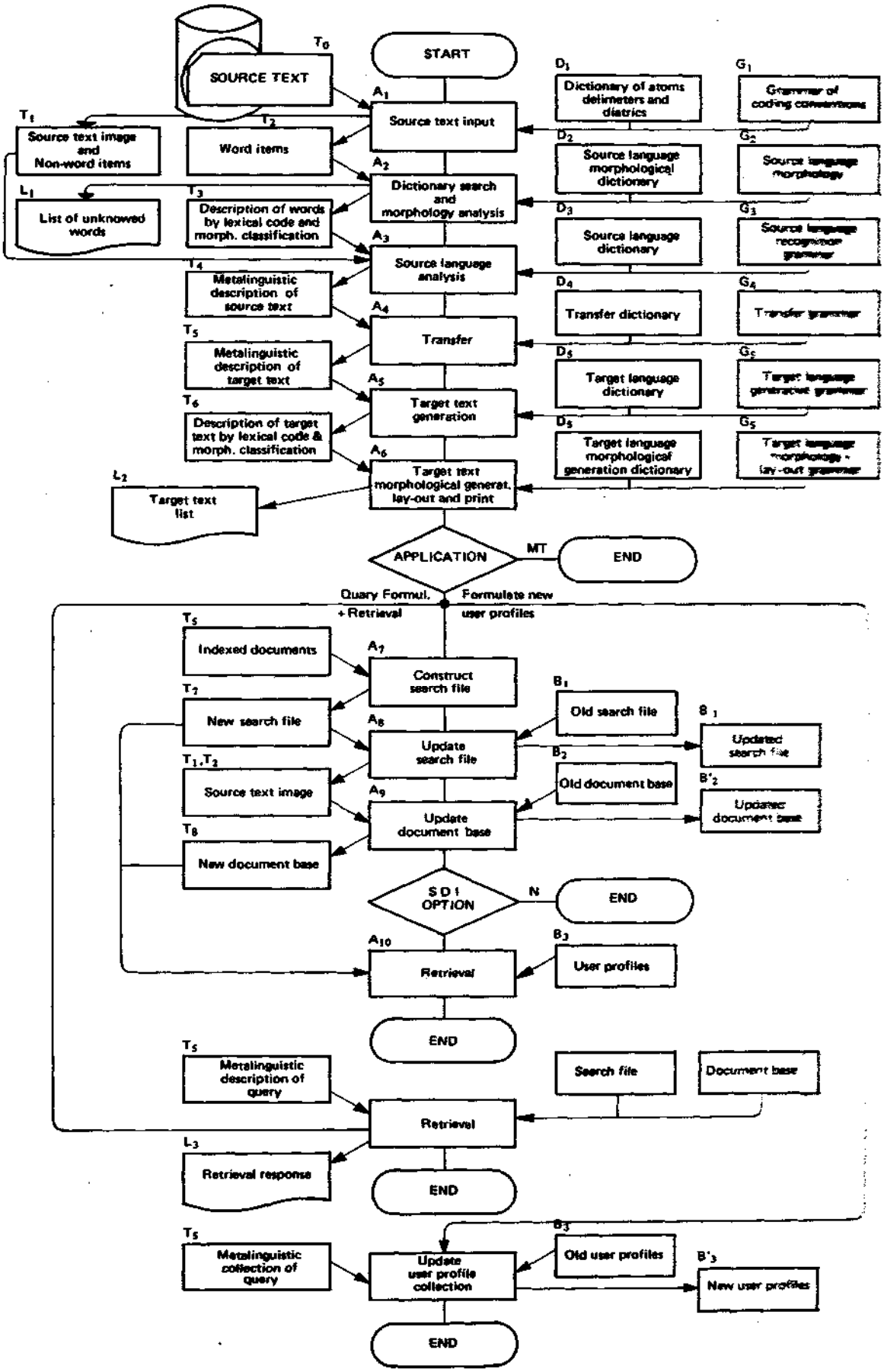

Diagram of the logical flow of data in SLCII.

A6) This phase is the inversion of phases Al and A2 and is applied only if the target text is to be used by man (always in the case of translation, and for the visualization of the indexing and query formulation results).

The subsequent phases are specific in an IR environment and include, basically, the IR data base management and retrieval functions.

A7) The indexing output is converted into the form required by the search algorithms. It consists either in file inversion, if an inverted file strategy is used, or in the construction of a search directory (by means of automatic classification) if a direct file strategy is chosen.

A8) Updating the search file is a merge procedure between the new search file (built in phase A7) and an old search file.

A9) The new document units are added to the document base and linked to the search file.

A10) The retrieval, in principle, consists in the comparison of a query (formulated by the phases A1 through A4) and the document collection (indexed in the same way). The strategy itself depends on the information retrieval language and the data base organization (e.g., direct or inverted file).

A11) This is an auxiliary module for the maintenance of the collection of user profiles (processed by phases A1 through A6) for SDI.

Further, the system is equipped with a series of modules for creating and maintaining the various dictionaries and grammars involved in the process.

In summer 1971 a set of modules (covering principally the phases A1 and A2 and the interface to PL/1 for programming phases A3 through A6) became operational with all the housekeeping modules and has been largely used for automatic indexing. The SLC programming language which is to replace PL/1 with the relative compiler and control modules had been defined and implemented by the end of 1971 and at present is undergoing the final system test. The modules involving IR are in an advanced state of completion and should all become operational in mid-1972.

The first version implemented operates in a complex overlay structure with static storage management in batch mode. An interactive version of SLC-II, to provide access from a network of terminals, is planned for 1973, and as a first step the transformation of all programs into re-entrant recursive modules with dynamic storage and task management has started and will be terminated in summer 1972. These modules will constitute the basis of the conversational SLC without major modifications.

FAIRS

Historically, the interest of CETIS in automatic documentation has been focussed on the EURATOM Nuclear Documentation system (ENDS) developed and used at Luxemburg by CID. ENDS is the classical solution of a semi-automatic information retrieval system in that the analysis and indexing of the source documents, the construction of the Information Retrieval Language (IRL) vocabulary (thesaurus) and the formulation of queries are performed intellectually, while the computer only does the matching of queries and document descriptions (i.e. its function is limited to the mechanical portion of the retrieval process). Furthermore, both indexing and retrieval are performed on a purely binary base so that the answer is given in such a form that a system specialist must scan it and select those documents which he considers to be most pertinent to the query (screening).

The objectives of FAIRS can be summarized as the elimination of human intervention from the process, and are to be achieved gradually.

In the first stage, the most costly and time-consuming operation — indexing — was mechanized, maintaining unchanged the thesaurus, the query formulation and retrieval strategy. In 1968 an experimental indexing program was introduced, and a collection of some 100 documents were indexed. With such a small collection, the only acceptable method of evaluation was by checking the human and computer consistency. As a term of comparison an inter-indexer consistency investigation was used, which had been performed with a collection indexed by two distinct indexer teams. The results showed that the values are more or less equivalent (Ref.[1)]). However, as the indexing quality can only be judged from the retrieval results, in 1969 a new collection containing some 500 documents was indexed and some 20 queries were processed both with the CETIS and the ENDS system. Again the quality of the automatic indexing turned out to be equivalent or even somewhat better than manual indexing (Ref.[2)]). These results encouraged the implementation of a fully automatic operational indexing program on the basis SLC-II, which was finished by the end of 1971 (Ref.[3)]). As the new program has a very high performance and enables large document collections to be indexed, a new evaluation series on a semi-operational basis was

prepared in cooperation with CID. It is planned to process documentation queries against the two systems in parallel, and to obtain in this way a practical assessment of the indexing quality in terms of Recall and Precision. This investigation is to be conducted through 1972.

The results described above were obtained in spite of the fact that the IRL has been maintained unchanged (i.e. presented all of the disadvantages of an empirically compiled thesaurus for manual usage), and the queries had been formulated manually. It is evident that the maximum possible consistency between the indexing on one hand and the query formulation on the other must be aimed at in order to obtain optimum employment of the system. This objective can only be achieved through query formulation with the same parameters as used for the indexing of documents. The algorithms of query formulation are basically those of indexing with two additional features:

a) a more complex syntax (with the logical operators AND, OR, NOT)
b) the assignment of weighting factors at the start.

With automatic query formulation it is also possible to abandon the binary logic in favour of a probabilistic evaluation of the similarity of queries and documents and, thus, to present the answers in order of decreasing pertinence.

A last point is the replacement of the Euratom thesaurus, which was compiled empirically and tries to reduce the number of descriptor terms for more efficient use by man. The method of automatic thesaurus construction is based chiefly on the research and experimental work of Salton (Ref.[5]) the principles being as follows:

—the vocabulary of the source language is divided into significant and non-significant words;
—non-significant words are defined on one hand through a given list which basically contains the "function words"; on the other hand they are defined statistically, and comprehend all the words whose frequency of usage in the overall collection exceeds a certain threshold.
— the remaining words represent, in principle, the IRL vocabulary. It presents, however, two sorts of error: 1. synonyms, 2. homographs;
—homographs are detected automatically during dictionary search and are resolved either intellectually or statistically;
— synonyms must be recognized intellectually or statistically;
— since the IRL envisaged is of the syntax-free class, for major precision, it is advisable to include in the IRL vocabulary not only single words but also compound expressions. Here again, the methods to be applied are intellectual (with a given list of expressions), or statistical.

The statistical methods to be applied for the resolution of all of these problems are those of automatic classification based on the co-occurrence and the context of words in the source texts. In principle, an expression is a group of words which occur together with a certain frequency. Synonyms are two words which are rarely used together in the same text, but have a very similar context. Homographs are resolved by the classification of the contexts of each occurrence. If the meanings of a homograph are not too close to each other, the contexts of the word through automatic classification are grouped into highly distinctive clusters. All known methods of automatic classification require a heavy load of data processing and it is therefore advisable to resolve the problems by combined automatic and intellectual work and to avoid using statistics if the only aim is to confirm well-known facts.

All methods and necessary software will be defined and tested during 1972 in connection with a large-scale IR project of an Italian public institution. This project is also to use SLC-II as the bask software.

Automatic language translation

Machine translation was originally one of the principle interests of CETIS. At the beginning of the sixties the in-house research work on linguistics turned out not to coincide with the CETIS objective, which was to produce an operational translation system. As a consequence, external relations were reinforced and by 1963 CETIS acquired, through a research contract, the Russian-to-English translation system developed by Georgetown University. In spite of the evident inadequacies and the unsatisfactory linguistic baas of the system, it was possible to adapt it to the needs of the scientific community of the JRC and to organize a valuable translation service all over Europe, producing over 100 translations a year. This translation service for the Community institutions and the member countries is still expanding (Ref.[4]).

As the hardware (IBM 7090) used has meanwhile grown obsolete, it was decided not to reprogram the system in a conventional way for a third generation computer, but to design a new system which should if possible integrate all the progress achieved in linguistics in the last decade, and use the new SLC-II system as basic software, so as to raise considerably both the quality and the economy of translation.

The development of the system started with the transformation of the existent data bases into the form required by SLC-II. Work was particularly slowed down owing to the impossibility of obtaining

personnel qualified for this task. Meanwhile, all the phases of the system were defined, and a strategy for the resolution of the problems involved was developed. In particular, the translation process was subdivided into three principal phases, source text analysis, transfer and target text generation, of which two, basically, are self-sufficient, i.e., can be developed and checked independently of the others.

In order to characterize the difference with respect to the old translation system, the following points must be emphasized:

— the source text analysis is syntax and semantics oriented;
— the transfer function is no longer the principle task of the translation, but is considered as an auxiliary function in those cases where the analysis does not permit a full metalinguistic formalization:
— the text generation in the target language is performed independently of the source language, being based on the metalinguistic description of the text.

It is accepted that the final objective, the FAHQMT (Fully Automatic High Quality Machine Translation) dreamt of by Bar-Hillel, will still not be achieved with this approach. But as far as one can estimate the future translation quality, it should more or less be equivalent to that of a man-made translation where the translator knows the source and target languages, but does not know and understand the subject matter of the text he is translating.

This level of translation quality should satisfy to a considerable extent the information needs of the scientific community, but it should also introduce machine translation into a far wider field of international communication. Further, the system has been explicitly designed to allow the addition of new source and target languages with the final objective of a multilingual reversible translation system.

The development of the translation system has been seriously impeded by staff shortage over the past years. Nevertheless, thanks to the personal sacrifices of the staff, some results were achieved. However it is evident that the schedule proposed in 1967 for the 3rd research programme 1968-72, which included the realization of the project, fell very short of its target.

References

1) H. Fangmeyer, G. Lustig, The EURATOM Automatic Indexing Project I.F.I.P. Conference 1968, Edinburgh.

2) H. Fangmeyer, G. Lustig, Experiments with the CETIS automatic indexing system. Symposium on the Handling of Nuclear Information, Vienna, 1970

3) H. Fangmeyer, Stand der Entwicklungsarbeiten für die automatische Indexierung bei der europäischen Forschungsanstalt Ispra. Jahrestagung der Deutschen Gesellschaft für Dokumentation, Bad Herrenalb, Oktober 1971

4) S. Capobianchi, G. Lustig, S. Perschke, A. Petrucci, W. Rittberger, Vernimb, Proceedings 5th Euratom Sponsored Meeting of Librarians Working in the Nuclear Field. - The Use of Machine Translation in Documentation. EUR 4256.e (1968)

5) G. Salton, Information Storage and Retrieval. Scientific Reports to the National Science Foundation.