[From: *Information retrieval and machine translation*, ed. Allen Kent. International conference… September 6-12, 1959, Cleveland, Ohio; part 2 (New York: Interscience Publishers, 1961)]

CHAPTER 23

Basic Directions of Work at the Experimental Laboratory of Machine Translation

N. D. ANDREYEV

Experimental Laboratory of Machine Translation,
Leningrad, U.S.S.R.

The Experimental Laboratory of Machine Translation, established in the beginning of 1958 at the Leningrad University by a pilot group of linguists and mathematicians, led by this writer, has launched its work in six directions.

The main direction of work of the Laboratory leads to the construction, programming, and testing of machine translation algorithms. At the present time (July, 1959) the Laboratory is developing 21 algorithms of independent analysis and one algorithm of independent synthesis.

The basic characteristic, setting the Laboratory apart from many other teams working in the field of machine translation, is the absence of binary algorithms; from the very beginning of the work it has been decided to proceed towards the development of algorithms of independent analysis and independent synthesis. The input text is here translated into an intermediate language (abbreviated IL, see below) without an a priori knowledge which output language was intended for the subsequent translation; indeed, no translation might be intended at all, and informational language might be used instead. The translation from IL into the output language is performed without knowing which input language yielded the IL text, or if there was any input language to begin with, the text having come from the information machine. The independence of analysis and synthesis assures the fulfillment of the following three objectives:

1. Reduction in the number of algorithms of translation (2p instead of p^2-p).

2. Simplification of the structure of the algorithms, since the translation into, and from the intermediate language is much simpler than a translation into or from an extra-machine* language.

---

*The input and output languages have, in recent works, been sometimes called real or concrete languages. It should be admitted, however, that the intermediate language also possesses the attributes of reality and concreteness. Therefore, it would be more correct to use

3. Feasibility of conversion from IL to other machine languages (such as, the informational language).

Another characteristic of the work of the Laboratory consists in the widest possible extension of the field of languages. This was based on the consideration that all the problems of algorithmic transformation of linguistic information become most acute in the case of structures far removed from one another. As a result, truly general methods of solution can be found, and can be applicable to many languages. At the same time, a considerable economy of resources can be achieved, since the working methods and approaches developed by one team can be relayed to other teams, boosting the efficiency of each solution.

A third characteristic of the Laboratory's work is the particular treatment of the quality of translation. The motto of our Laboratory holds that "the best translation is the timely translation." Therefore, the requirement of a good style of the output text is rejected in principle; the only limitation imposed upon the transformation of the message in our algorithms resides in the rigid requirement of invariance of the scientific sense in the message. If the translation is somewhat "clumsy," but fully understandable, it is considered quite adequate. Such an approach substantially simplifies the algorithms and accelerates the processing of information involved.

The methodology of constructing algorithms is based on working hypotheses and standard analysis. Statistical analysis of texts is used to define the sequence of working hypotheses which approximate the language event with an increasing accuracy, giving rise to a series of commands and blocks whose execution manifests a progression from the treatment of more frequent events to that of less frequent events. This defines the area of standard analysis. Events, whose probability is below a given level are treated either through commands and blocks of a non-standard analysis, or are not treated at all, depending upon the desired speed of operation of the algorithm (i.e., the algorithm can operate in two modes: full and abbreviated, slow and fast). Thus, the abbreviated treatment of text deliberately allows for "skipping," which does not affect the transformation of the main flow of information : several rare syntactic forms will remain untranslated, while the remaining body of the translated text will pass through the algorithm with a sufficiently high speed.

The Laboratory is developing the following algorithms of independent analysis: Russian (L. Zasorina, leader), Chinese (S. Yakhontov), Czech (E. Andreyeva), German I (A. Belopolskaya), German II (N. Paronyan), Rumanian (R. Piotrovskiy), Vietnamese (I. Bystrov), Serbo-Croatian (P. Dmitriyev), English I (B. Leykina), English II (V. Burlakova), French I (V. Tarkhova), French II (R. Piotrovskiy), Spanish

---

terms capable of reflecting the essence of the difference: languages used in the machine are machine languages, as opposed to the input and output languages under the common name, extra-machine languages.

(E. Golubeva), Norwegian (V. Berkov), Arabic (O. Frolova), Hindustani (E. Katenina), Japanese (A. Babintsev), Indonesian (N. Andreyev), Burmese (N. Andreyev), Turkish (S. Ivanov), Swahili (N. Okhotina). Independent synthesis is, as yet, being developed only for the Russian language (L. Zasorina and N. Andreyev, leaders).

The presence of two algorithms of English, German and French analysis is due to the availability of sufficient resources in each of these languages for work in more than one semantic area. In principle, any algorithm is characterized by three parameters: extra-machine language, type of conversion (analysis or synthesis), and type of texts (semantic area). For example, B. Leykina directs the construction of analytic algorithm of English medical texts, while V. Burlakova is in charge of English chemical texts. A universal algorithm of any extra-machine language requires an excessively large volume of passive memory; furthermore, such a universal algorithm renders the solution of the multiple-meaning problem much more complex. Finally, parallel branch terminologies of different languages prove to be closer, in some aspects, to each other than the terminologies of various branches of science in the same language. Consequently, the third parameter of the algorithm plays, in our practice, a fairly important role, and especially so in the light of our work on the intermediate and informational languages.

Mathematicians participate side by side with linguists in the construction of algorithms; algorithmic groups, the basic working cells of the Laboratory, are usually composed as follows: the algorithm leader is a specialist in an extra-machine language, his assistant is a mathematician, and there are two or three linguists. For example, the Arabic algorithm group consists of an Arabic specialist, O. Frolova, mathematician S. Fitialov, and Arabic specialists A. Smirnov, I. Rozenbaum, and F. Isayeva. Some mathematicians and linguists work in more than one group; the staff of all the algorithmic groups numbers about 80 persons.*

The leaders of algorithmic groups form the algorithmic section, whose weekly meetings (known as "machine Tuesdays") are devoted to the discussion of concrete results of group work, theoretical developments of other sections, and publishing and organizational problems. In other words, the algorithmic section is the working center of our Laboratory.

The second direction of work of the Laboratory is closely connected with the above: parallel with the construction of particular algorithms, we are simultaneously developing a special symbolic meta-language of machine translation. The symbolic language (SL) comprises three levels: para-language (symbols of the treated extra-machine language), meta-language (symbols to record information on the elements of the extra-machine language), and ortho-language (symbols to record in-

*The total personnel of four sections of the Laboratory, algorithmic, mathematical, machine languages, and design, numbers over 100 workers, the overwhelming majority of whom are not on payroll.

formation on the algorithm). Symbolic language has determined symbols of the working and type variables, operators, commands, and simple and complex addresses; and has formulated rules of combining symbols, i.e., forming meta-phrases in symbolic language, where these rules are connected with the methods of transforming information adopted in our algorithms. The basic purpose of the symbolic language is the maintenance of accuracy and brevity in the formulation of machine translation rules. The programs recorded in symbolic language contain not a single word of any extra-machine language (which, incidentally, ensures the invariance of algorithmic record with respect to its representation in the texts written in extra-machine languages). There is also the problem of automatic programming of algorithms recorded in the symbolic language. The latter may well be the most important problem in the use of symbolic language: the functional relationship with the intermediate language of machine translation and with the recording of various types of algorithms of language modeling.

With respect to the third direction of activity of the Laboratory, the modeling of language, one should note that it is the subject of, as yet, a small part of the personnel: S. Fitialov, N. Andreyev, G. Tseytin, M. Otkupshikova, L. Andreyeva, I. Bystrov, and B. Palek. The work includes the examination of general problems of building language models, the development of a statistical-combination model of grammar, a model of a semantic structure of language, and a model of intermediate-language construction by means of a translation system; the above models are being subject to experimental testing.† The work on symbolic language and language modeling, as well as the general problems of logical structure of algorithms and information processing, are discussed in the mathematical section whose personnel includes mathematicians and a small number of linguists.

Fairly intensive is the work in the fourth direction: theoretical development and experimental testing of the intermediate language. IL is regarded here as a machine language with its own structure. Any input message is transformed by the analytic algorithm in a way that the form of the resulting IL message does not reveal the language origin of the input message. The structure of grammar and terminology of IL is determined as a result of statistical generalization of data of the totality of input languages, in conjunction with a logical analysis of generalized data. The development of the intermediate language is accomplished in the form of a sequence of models of increasing complexity; the transition from one model to the next follows a series of experiments carried out to evaluate the characteristics of the given model and to introduce suitable corrections. This work is carried out by about one-third of the personnel of the Laboratory; permanent participants in the development of the intermediate

† See: "Theses of the Mathematical Linguistics Conference," L., 1959, "Materials on ML and MP," vol. 2, L., 1960, and "Reports and Communications on ML and MP," vol. 1, L., 1960.

language are N. Andreyev, V. Berkov, S. Fitialov, L. Zasorina, N. Gurov, B. Leykina, A. Belopol'skaya, A. Ogloblin, V. Kormushin, T. Zubkova, N. Kremneva, S. Pestova, R. Pazukhin, G. Pak.

After the outlines of the intermediate language have been more or less determined, the Laboratory team commenced work in the fifth direction: In the field of informational languages. As a special area, it has been decided to select informational languages for cybernetics (computer mathematics), jurisprudence and medicine.

The development of problems of the construction of machine languages is carried out within the Section of Machine Languages.

The sixth and last direction of work of the Laboratory is connected with the design of an experimental cybernetic device for testing linguistic algorithms. The design section includes engineers, mathematicians, and linguists, not limited to members of the Laboratory; the design of the device is carried out in cooperation with the Leningrad Computing Center of the Academy of Sciences of the USSR. The team of scientists and engineers working on the solution of this problem is led by N. Posnov, while the leading group also includes N. Andreyev, V. Varshavskiy, S. Fitialov, G. Bondarenko, V. Strelkova.

It is readily apparent that all six directions of the work of the Laboratory: the construction of algorithms, development of a symbolic language, modeling of language structures, construction of an intermediate language, creation of informational languages, and design of a linguistic machine, are closely connected with one another and represent an integral complex. The harmonious development of all the aspects of this complex will provide an effective solution to the intricate problem of machine transformation and processing of linguistic information.