Syntactico-Semantic Analysis in the AMPAR Machine Translation System

YURI N. MARCHUK

All-Union Centre for Translation of Scientific and Technical Literature and Documentation USSR State Committee for Science and Technology, Moscow, USSR

The model of machine translation by correspondences allows an MT system to be designed in a way that makes possible the introduction into it of modifications and improvements suggested by translations of new texts. The syntactico-semantic module of the AMPAR system is based on initial parameters which can be supplemented with subtler sub-parameters in the course of further development

The idea of syntactic analysis by configurations, which was put forward in the early period of machine translation (MT), reflected, to a certain extent, the process of sentence comprehension as it is interpreted by modern psycholinguistics [1]. As far as translation proper is concerned, determining syntagma boundaries invariably leads to establishing translation correspondences, provided it is admitted that language comprehension proceeds by syntagmas and translation is immediately linked to identification of meaningful segments. As is known, however, configuration analysis and synthesis have never been implemented on a practical scale due to the complications arising in the process of identifying syntagmas and establishing their internal and external relations by formalized algorithmic means. Machine translation parsing procedures developed instead along the lines of the complete syntactic analysis of a sentence, have also failed, save for a few exceptions, to produce practically valid results. By 'practically valid' results we imply those used in operational MT systems. A survey of such systems demonstrates that they retain the syntagma as a basic unit of analysis, though its boundaries are determined in each case by what might be called successive approximation methods.

The theoretical foundation of the operational MT systems is provided by specific applied working models called 'reproducing engineering-linguistic models' [2]. These models are distinguished by prompt feedback as well as by the proximity of their basic parameters to those traditionally utilized by man in language analysis and synthesis. Thus, word connections within a syntagma and relations between syntagmas in a sentence can be fairly well represented in terms of categories close to traditionally employed parts of speech and sentence parts. The AMPAR system (the Russian abbreviation for 'Automated English-Russian Machine Translation' system), developed at the All-Union Centre for Translation of Scientific and Technical Literature and Documentation, relies on a model of translation correspondences (MTC model) which, in its turn, incorporates a system of lexico-grammatical word classes and syntactic word functions sufficiently close to the traditional grammatical system, common in Russian and many European languages. The MTC model consists of two components: material and dynamic ones. The material component contains a description of what must be translated. It is composed of two principal elements — material proper and translation ones. The material component comprises the following constituents: vocabulary, grammar, and semantics (see Figure 1). Each constituent

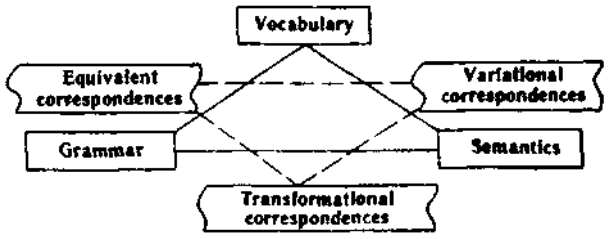

Fig. 1. Structure of the material component in the MTC model

embraces a set of interconnected elements: vocabulary is a stock (list) of words belonging to a given sub-language; grammar is a set of grammatical features; and semantics is a set of semantic features. The translation element of the material component includes correspondence types, subdivided into equivalent, variational, and transformational parts. The correspondence types cover a sufficiently large number of cases where adequate translation can be attained through purely linguistic means, as the 'transformational correspondences' envisage, for the most part, linguistic transformations.

The dynamic component of the MTC model responds to the question of how the translation is to be effected. This component contains a mechanism for establishing correspondences, which realizes the dynamics of detecting correspondences in the input text as well as that of constructing an equivalent output text. The dynamic component includes two elements: algorithmic and programming ones. The algorithmic element comprises an algorithm proper, a translation grammar, and a machine dictionary. The algorithm controls application of the translation grammar and dictionary at the stages of establishing correspondence and constructing an equivalent text. The translation grammar is a specific binary grammar that yields an optimal arrangement of requi-

red grammatical features to establish correspondences in translation. The dictionary holds a collection of lexical units accompanied with essential data and ordered according to the following three grounds: source/target language, ambiguity/unambiguity of words, idiomatic combinations/single lexemes. The programming element is a standard program package designed to accomplish linguistic operations. A prominent role is played by a language of standard operators, which were developed for this particular system [3] and serve to implement linguistic analysis and synthesis. The programming element comprises lists as well as lexical analysis, translation and parsing schemata (see Figure 2).

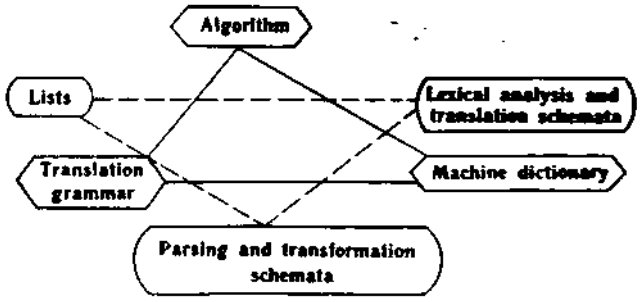

Fig. 2. Structure of the dynamic component in the MTC model

In order to establish correspondences, the following operations are carried out:

— the limits of correspondences are determined by rules that eliminate words irrelevant for this task;

— algorithmic transformations are fulfilled at the lexical and syntactico-semantic levels;

— linguistic categories relevant to translation are deduced;

— data obtained in the course of analysis is verified and corrected. This process also involves the systemic organisation of the model. (For a detailed description of the AMPAR system set-up see [4]).

Original syntactic and semantic information is stored in the system in the form of syntactico-semantic word codes which reflect the input dictionary division into semantic subclasses organized on a distributional-statistical principle. The problem of transition from individual semantic features of conjoinable words to the characteristics that unite word groups and, further, subclasses and classes of words has been discussed in a special work on resolving lexical ambiguity with the help of a contextological dictionary [5]. It was demonstrated that this inductive way is the only possible one leading to a resolution of lexical ambiguity.

However, as is shown by the example below, it is also possible and even convenient to build up syntactico-semantic analysis in an inductive manner utilizing, in the first place, certain basic categories, similar to the 'parts of speech — sentence parts' system, which are essential for translation and can be expanded when necessary. Consider, for instance, an English sentence *Participants are expected to cover their* own *expenses,* which the AMPAR system will translate into Russian as *участники ожидаются охватить их собственные расходы* To provide for this translation, it is necessary: (a) to resolve lexical and grammatical homonymy of words *cover, own, to;* (b) to adequately translate the polysemantic words *cover and expense,* and (c) to correctly assign the syntactic functions of subject, predicate, attribute and object to the respective sentence parts. It should be noted that these functions are quite sufficient for correctly translating a sentence, in fact, general translation theory almost never has to resort to any means of syntactic and semantic analysis other than traditional sentence parts [6—8] However, the problem of idiomatic collocation with the verb ожидать remains unsolved — it should be said instead *как ожидаеся участники должны оплатиь их собственные расходы*.* The initial inventory of syntactico-semantic features is insufficient to provide for such a transformation. What is required here is a system of semantic features that would help to introduce more idiomatic translations than the one above. It would also be quite feasible to localize the corresponding features in the dictionary or in the appropriate parts of the parsing system relying on particular verbs that express a given semantic content. Generally speaking, there is actually no complex translational task that cannot be tackled with a finite inventory of sufficiently formalized operations. However, when analysis constitutes only one of the system functions, the principal problem is that of whether an integral system can emerge from all these *ad hoc* solutions and translation devices, to what other verbs these transformations can be applied, what classes of verbs (or other words) would need similar (or different) transformations, etc. An a priori constructed system of such transformations cannot be but a very limited one. Conversely, errors of this kind, if collected and sorted out, might help to expand the original system, provided, of course, an expansion is envisaged in its design

The syntactico-semantic analysis involving syntactic functions and an algorithmic component relies not merely on simple or complex sentences used as analysis units, but on the entire input text in the computer's internal storage This allows the discovery of acceptable solutions for a number of problems related to the quality of translation. Thus, for instance, the problem of pronoun antecedents can be solved in a practically satisfactory way within the limits of a text.

Elaborating formalized operations on fuzzy sets of linguistic objects calls for a procedure leading to provisional solutions subject to further improvement. Organisation of syntactico-semantic analysis along the above described lines allows the accumulation of data relevant to translation in an optimal way. It is obvious that a judgement concerning the practical value of the machine translation system in question must be based on the quality of its output. This proved to be quite satisfactory in all the experiments.

REFERENCES

1. Reimold, P. A formal psycholinguistic model of sentence comprehension. *American Journal* of *Computational Linguistics,* 1975, 12, No. 5, 3—46.

* The translation оплатить instead of оживатиь is not quite accurate either, but this problem must be solved at the lexical ambiguity level.

2. Piotrovsky. R. G. Engineering linguistics and language theory. Leningrad: Nauka Publishers, 1979, 111 p. (in Russian).

3, Motorin, Yu. A.; Marchuk, Yu. N. Implementing automatic translation on modern serial general-purpose computers. *Voprosy radioelektroniky*, *ser. EVT*, 1970, No. 7, 20—29 (in Russian).

4. Marchuk, Yu. N.; Tikhomirov, V. D,; Scherbinin, V. I. An English-Russian machine translation system. — In: *Machine Translation and Computerisation of Information Processes*. Moscow, 1975, 18—33 (in Russian).

5. *Contextological dictionary for machine translation of words with multiple meaning from English into Russian. In 2 parts*. Comp. by Yu. N. Marchuk. Moscow: All-Union Centre for Translation, 1976 (in Russian).

6. Barkhudarov, L. S. *Language and translation*. Moscow: International Relations Publishers, 1975, 239 p. (in Russian).

7. Retsker, Ya. I. *Translation theory and practice of translation*. Moscow: International Relations Publishers, 1974. 215 p. (in Russian).

8. Shveitser, A. D. *Translation and linguistics*. Moscow: Voyenizdat Publishers, 1973, 279 p. (in Russian).