

ETAP-2: The Linguistics of a Machine Translation System

Jurij D. Apresjan, Igor M. Boguslavskij, Leonid L. Iomdin, Alexandre V. Lazurskij, Vladimir Z. Sannikov and Leonid L. Tsinman

Volume 37, Number 1, mars 1992

La traduction en Russie : théorie et pratique / Translation in Russia:
Theory and Practice

URI: <https://id.erudit.org/iderudit/001895ar>

DOI: <https://doi.org/10.7202/001895ar>

[See table of contents](#)

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)

1492-1421 (digital)

[Explore this journal](#)

Cite this article

Apresjan, J. D., Boguslavskij, I. M., Iomdin, L. L., Lazurskij, A. V., Sannikov, V. Z. & Tsinman, L. L. (1992). ETAP-2: The Linguistics of a Machine Translation System. *Meta*, 37(1), 97–112. <https://doi.org/10.7202/001895ar>

Article abstract

On présente ETAP-2, un système de traduction automatique (anglais-russe), en insistant sur ce qui le différencie des autres systèmes du point de vue linguistique. On expose ensuite quelques échantillons de traductions réalisées avec ETAP-2 puis on évalue la qualité des résultats et la vitesse d'exécution. On explique enfin l'idéologie et la structure qui sous-tendent ETAP-2 en comparant le système avec ETAP-1 (français-russe) et en décrivant chacune des étapes de l'analyse linguistique à laquelle ETAP-2 a recours pour traduire.

ETAP-2: THE LINGUISTICS OF A MACHINE TRANSLATION SYSTEM

JURIJ D. APRESJAN, IGOR M. BOGUSLAVSKIJ,
LEONID L. IOMDIN, ALEXANDRE V. LAZURSKIJ, VLADIMIR Z. SANNIKOV,
LEONID L. TSINMAN
*Institute for Information Transmission Problems,
the USSR Academy of Sciences, Moscow, USSR*

Résumé

On présente ETAP-2, un système de traduction automatique (anglais-russe), en insistant sur ce qui le différencie des autres systèmes du point de vue linguistique. On expose ensuite quelques échantillons de traductions réalisées avec ETAP-2 puis on évalue la qualité des résultats et la vitesse d'exécution. On explique enfin l'idéologie et la structure qui sous-tendent ETAP-2 en comparant le système avec ETAP-1 (français-russe) et en décrivant chacune des étapes de l'analyse linguistique à laquelle ETAP-2 a recours pour traduire.

INTRODUCTORY REMARKS

ETAP-2 (which is short for *ElectroTexničeskij Avtomatičeskij Perevod*, second version) is an experimental second-and-a-half generation English-to-Russian machine translation system, which was elaborated over the years 1982-1985 at the Informelectro Research Institute in Moscow. It was designed as a fully automatic system functioning without any human interference at any stage in the translation process. The experiments with ETAP-2 started early in 1984 and were interrupted in 1985, after the transference of our research group to the Institute for Information Transmission Problems of the USSR Academy of Sciences. They were resumed three years later at the said Institute.

At present ETAP-2 makes use of full-fledged morphological and syntactic models of English and Russian, which are considered to be sufficiently complete to provide for the automatic analysis and synthesis of any hard science and technology texts. The combinatory dictionaries of English and Russian come close to 6000 and 7000 entries respectively, which, we must admit, falls far behind the amount deemed necessary for a practicable system of machine translation. Due to these dictionary insufficiencies the set of translation rules cannot be claimed to be complete either. Both these shortcomings, however, are not a matter of principle and can easily be remedied.

From a purely linguistic point of view ETAP-2 differs from the comparable systems of machine translation in the following respects.

1. The source and target languages are described completely independently of each other. The description of each language is a reduced but principled version of the "Meaning — Text" type of model in the sense of Mel'čuk 1974. Each model gives an integrated (unified) description of the morphology, syntax, and lexicon of the language in question. That means that morphology, syntax, and lexicon are ideally coordinated with one another both, in the types of information contained and in the formal languages in which the information is recorded.
2. Linguistic knowledge is represented in a purely declarative way, that is, completely independently of the algorithms. Declarativeness has two important

advantages over the procedural way of representing linguistic information (at least at the experimental stage of work).

First, it makes immediately observable the linguistic model underlying the MT system. The model can be scanned and examined in its entirety and can be readily evaluated and compared with similar models. There is no need to piece it together from the fragments extracted from the algorithmic procedures. In this respect a declaratively represented linguistic model is very much like the traditional type of linguistic description. At bottom it is the familiar grammar and dictionary, the only difference being that they are written in a formal language.

Secondly, a declaratively represented linguistic model is easy to correct in the course of machine experiments, provided the system supplies, alongside of every translated sentence, the complete protocol of what the computer has done in order to translate it.

3. The formats of linguistic description of the processed languages are fully standardized. Both operational languages, English and Russian, are described after a unified and sufficiently general pattern. That opens up the prospect of including in the system new languages without having to change anything either in the formats of linguistic knowledge representation or in the algorithms of text analysis, transformation (transfer), and synthesis.

The linguistic components of the system (with the exception of a narrow and purely terminological part of the dictionary) are not oriented to any single object domain. This property of the system is an immediate consequence of the completeness of the linguistic models employed.

As has already been mentioned, the morphological and syntactic models of the operational languages are designed for processing any type of hard science and technology texts, that is, they take into account a very wide range of grammatical forms and syntactic constructions likely to occur in such texts.

As far as the dictionaries are concerned, they contain a large amount of words specific for the chosen object domain. If transferred to a different object domain, they require a considerable expansion of their terminological part. Yet they contain about 2500-3000 common words which are likely to occur in any type of text, and hundreds of rules that go with them. Let it be emphasized that the part of vocabulary common for the majority of various object domains constitutes the lexical core of any language. It is precisely this core that is responsible for a great number of language specific constructions which obstruct smooth translation. Once they are included in the common, that is, transferable part of the dictionary, the basic difficulties of the translation stage proper may be considered to have been taken good care of.

The relative completeness of the linguistic models underlying ETAP-2, besides ensuring the possibility of using the same software in a number of machine translation systems, has interesting theoretical implications as well.

It is obvious that the quality of machine translation is a direct function of the sophistication and reliability of the underlying linguistic model (the bearing of linguistics upon machine translation).

However, there is an inverse dependence of the quality of linguistic models on the actual performance of machine translation systems. Machine translation turns out to be an ideal experimental shooting range for linguistics, a shooting range where it can test its scientific tools and evaluate the soundness of its theories. If a system of machine translation makes use of sufficiently sophisticated language models it is sure to exercise a profound and stimulating influence upon linguistics. In the course of computer experiments "negative" material is accumulated which allows to rectify familiar linguistic

rules and formulate a number of new rules. It may also make obvious for the linguist the necessity to introduce new linguistic concepts and even to formulate new linguistic theories. In short, machine translation creates the prerequisites for an entirely new branch of linguistic science — experimental linguistics, employing experiment as a tool in its search for the scientific truth (the bearing of machine translation upon theoretical linguistics).

SAMPLES, SPEED AND QUALITY OF MACHINE TRANSLATION

To evaluate a system of machine translation it is much more helpful to be able to have a look at its actual output rather than at its technical specifications. ETAP-2 translated original English texts of two types: 1) connected texts (a paper, an abstract), 2) patent headings of the INPADOK database. Translation was carried out sentence by sentence, without any human interference at any stage (except that sentences more than 35 words long were broken, for purely technical reasons, into pairs of shorter sentences). Every sentence was translated in a unique way, *i.e.*, without any variation. In other words, morphological, lexical, and syntactic homonymy and polysemy were resolved to such an extent that the number of alternatives at every ambiguous point in the sentence was reduced to just one possibility. 50% of the total number of sentences were translated quite well at the very first trial. Unsatisfactory first trials in all the other instances are accounted for by the absence of some relevant information in the dictionary or (less often) by certain imprecisions in the rules. In an overwhelming majority of such cases it was possible to introduce non-ad-hoc corrections without endangering the system at large, and thus to enable it to rectify the faulty translation.

Sample translation of a connected text

- (1) The Future of Electronics in instrumentation
Buduúčee elektroniki v oborudovanii
- (2) Advances due in spectroscopy, chromatography, spectrometry and electrochemistry
Dostiženija, ožidaemye v spektroskopii, xromatografii, spektrometrii i electroximii.
- (3) In recent years, government regulations, technical demands, microprocessor development and a host of other influences have forced a revolution in chemical, biochemical and industrial instrumentation.
V nedavnie gody pravitel'stvennye postanovlenija, texničeskije trebovanija, razvitie mikroprocessorov i množestvo drugix faktorov vyzvali perevorot v ximičeskom, bioximičeskom i promýšlennom oborudovanii.
- (4) Thirty five years ago, significant breakthroughs were achieved in chemical analysis instrumentation, and UV and IR spectrometers, mass spectrometers, and PH meters became available.
Tridcat' pjat' let nazad značitel'nye uspechi byli dostignuty v oborudovanii ximičeskogo analiza i ul'trafioljetovye i infrakrasnye spektrometry, mass-spektrometry i PH ščetčiki stali dostupnymi.
- (5) The search for a synthetic substitute for natural rubber, for example, accelerated during World War II and spurred the development of instrumentation that could analyze molecular structure in the laboratory.
Poisk sintetičeskogo zamenitelja natural' nogo kaučuka, naprimer, uskorilsja v tečenie vtoroj mirovoj vojny i stimuliroval razvitie oborudovanija, kotoroe moglo analizirovat' molekularnuju strukturu v laboratorii.
- (6) In the early sixties, chromatography became an analytical tool (now the most highly used instrumental technique in the chemical laboratory).
V načale šestidesjatyx godov xromatografija stala analitičeskim instrumentom (v nastoja — ee vremja naibolee široko ispol'zovannyj metod v ximičeskoj laboratorii).

- (7) Nuclear instrumentation, which appeared in the late fifties, dramatically changed medical diagnosis.
Jadernoe oborudovanie, kotoroe pojavilos' v knonce pjatidesjatyx godov, rezko izmenilo medicinskuju diagnostiku.
- (8) Improved detectors and sampling technology have led to more exotic instrumentation methods, such as gas chromatography, mass spectrometry (GCMS), high performance liquid chromatography (HPLC), X-ray fluorescence and plasma emission spectroscopy.
Usoveršenstvovannyje detektory i texnika vyborki priveli k bolee èkzotièskim instrumental'nyx metodam, takim, kak mass-spektrometrija gazovoj xromatografii (GC-MS), vysokoèffektivnaja zidkostnaja xromatografija (HPLC), rentgenovskaja fluorescencija i plazmennaja èmissionnaja spektroskopija.

Sample translations of patent headings

- (9) Pendulously suspended bucket with a steering curve for a bucket conveyor.
Majatnikoobrazno podvešennyj kovšs napravljajuščej krivoj dija kovšovogo konveera.
- (10) Diffusion bonding of aluminium alloys.
Diffuznoe soedinenie alljuminievyx splavov.
- (11) Hydrocarbon oil based silicone antifoams.
Kremnieve protivopennye sredstva na osnove uglevodorodnogo masla.
- (12) Static changing device for drive-brake operation of variable speed asynchronous motors fed by a current convertor.
Statièeskoe pereključajuščee ustrojstvo dlja operacii puska ostanova nad-asinxronnymi dvigateljami peremennoj skorosti, pitaemymi tokovym preobrazovatelem.
- (13) Stripless electrical wire terminal for distributors of telecommunication installations, especially telephone installations.
Bespolosnaja električeskaja klemma dlja raspredelitelej ustanovok telekommunikacii, osobenno telefonnyx ustanovok.
- (14) Conveying apparatus for sheet- or foil-like material.
Peredajuščee ustrojstvo dlja listovogo ili plenočnogo materiala.
- (15) Planetary gear train for automotive transmission or the like.
Planetarnaja zubčataja peredač dlja samodvizuščejsja peredači ili tomu podobnogo.
- (16) Gas-insulated electrical apparatus.
Izolirovannoje gazom električeskoe ustrojstvo.
- (17) Flat-card-shaped semiconductor device with electric contacts on both faces and process for its manufacture.
Poluprovodnikovoe v forme ploskoj karty ustrojstvo s električeskimi kontaktami na obeix storonax i process dlja ego izgotovlenija.
- (18) Optical phase grid arrangement and coupling device having such arrangement.
Ustrojstvo optičeskoj fazovoj setki i soedinjajuščee ustrojstvo s takim ustrojstvom.

To evaluate properly the speed of machine translation it is necessary to bear in mind that ETAP-2 is implemented on an outdated ES-computer with the nominal internal performance of 1,000,000 operations per second (the actual performance is considerably lower). The set of programs amounting to 15,000-20,000 operators was originally written in PL-1. In the course of subsequent work several programs were rewritten in the Assembler to speed up translation.

After these improvements the time of translation of a middle-sized sentence of about 25-30 words and of medium complexity amounted to some 30 seconds. This speed may be evaluated as quite satisfactory considering an exceptionally high degree of homonymy in English texts and a relatively high quality of translation. It is much higher than the speed of human translation, as was exemplified in an experiment with 10 fifth year students of the translators' department of the Moscow Foreign Languages Institute

who translated the same text. Their performance was 5-6 times worse than that of the computer.

As far as the quality of translation is concerned, its objective evaluation presents a more difficult task.

To begin with, the quality of translation should be evaluated not for separate sentences but for a big chunk of text, with the stipulation that in the process of its translation the system does not undergo any radical modifications. Otherwise there would be no way to judge how stable the results of translation are. In this respect the performance of ETAP-2 was sufficiently stable. Although we did introduce some minor modifications in the system, we tested its stability by sample translations of the fragments that have previously been translated. That is why it may be claimed that on the whole the translation of all the texts has been carried out by the same system.

Secondly, to form the right idea of the quality of translations they should undergo expert evaluation on the basis of a certain set of criteria. O.S. Kulagina (1979), to quote but one authority, cites three such criteria: the adequacy of translation (the degree of its semantic identity to the original), its comprehensibility, and its grammatical well-formedness. Unfortunately, we have not yet been able to carry out such work and will have to confine ourselves to a more or less impressionistic evaluation of the translation. We shall compare it with the original and discuss some obvious violations of adequacy, comprehensibility, and grammatical well-formedness.

All the translations listed above are comprehensible, with the exception of sentence (1). But it is no more understandable in the original.

Most of the translations are adequate, *i.e.* they render the meaning of the original correctly. The only semantic mistake is the perfective aspect of the verb *ispol'zovat'* in sentence (6). Instead of the required translation *naibolee siroko ispol'zovannyjmetod...* the computer yielded *naibolee siroko ispol'zovannyj metod...* In this way the general statement about what is going on in any chemical laboratory was turned into a description of a single event. This mistake is unavoidable. The point is that in generating aspectual verbal forms in Russian we cannot make use of their direct prototypes in the corresponding English verbal forms since there are none: English aspect, if there is any, is utterly different from Russian. It follows then, that to generate the necessary aspectual form the system should process the more implicit information contained either in the context of the original verbal form, or in the lexico-semantic properties of the English verb, or in the context of the respective Russian verbal form, or in the latter's lexico-semantic properties. In the case at issue there is absolutely nothing to go by: the contexts of the original and the translation, as well as the lexico-semantic properties of the verbs *USE* and *ISPOL'ZOVAT'*, give no grounds for choosing the imperfective aspect.

All the other mistakes may be considered to be deviations from grammatical or stylistic norms of the Russian language.

In sentence (4) a comma is missing in front of the conjunction *I*, although the latter connects two independent sentences. The mistake does not hamper understanding and can easily be rectified.

In sentence (12) there is also a punctuation error — a hyphen in the group *nad-asinxronnymi (dvigateljami)* instead of a hyphen in the group *(operacija) puska-ostanova*. This error is so funny that it cannot in the least hamper understanding either. It goes back to a slight and easily rectifiable program mistake.

In sentence (16) the translation *električeskoe ustrojstvo gazovoj izoljaciej* is preferable to the translation *izolirovannoe gazom električeskoe ustrojstvo* that was actually put out by the system. Again there is no difficulty at all in suppressing this error.

In sentence (17) the word order *poluprovodnikovoe ustrojstvo v forme ploskoj kary* should be preferred to the one actually yielded by the system: *poluprovodnikovoe v forme ploskojk karty ustrojstvo*. The error is rectifiable.

In sentence (18) there is a negligible stylistic error — the repetition at a close distance of the same meaningful word. The error does not hamper understanding, and it is doubtful that one should try to do away with it.

On the whole it seems possible to estimate Russian translations yielded by ETAP-2 as machine translations of sufficiently high quality comparable to human translations.

IDEOLOGY AND STRUCTURE OF ETAP-2

In both these respects ETAP-2 is a continuation of an earlier French-to-Russian experimental machine translation system that functioned for a year and a half and yielded translations of sufficiently high quality. It has been reported in a series of publications; cf. in particular Apresjan et al. (1985). We shall compare the two systems, starting with their differences and then going over to their similarities since it was continuity more than anything else that played a decisive role in developing ETAP-2.

ETAP-2 DIFFERS FROM ETAP-1 IN THE FOLLOWING FOUR RESPECTS.

1. First of all, in developing ETAP-2 we set out to considerably expand its potential with the view to transforming an experimental system into a really operational system yielding commercially valid translations. To meet this requirement, machine translation had to be fast and the whole system had to be shaped as a subsystem readily joinable to any information system processing information in foreign languages.
2. This new objective determined the new structure of the system's linguistic software. To speed up the work of the system all rules were broken into three types — general, specific, and dictionary rules. General rules making up no more than a quarter of the total number of rules participate in the processing of every sentence although that does not necessarily mean that every rule applies to every sentence. Specific (or small-scope) and dictionary rules which make up about three quarters of the total number of rules participate in the processing of only those sentences whose lexical composition may require their application. They are activated by the respective dictionary entries in which they are either mentioned by name (specific rules) or included bodily (dictionary rules). In this way ETAP-2 was equipped with the means of ideally tuning itself up to the processing of the current sentence. As a result the time of translation of almost every sentence was considerably reduced due to the reduction of the number of rules that had to be resorted to in processing the sentence.

Since it became necessary to enter in the combinatory dictionaries of ETAP-2, alongside classificatory information (part of speech, syntactic and semantic features, etc.), operational, or rule information as well, the entries of the combinatory dictionaries acquired a much more complicated and logically ramified structure.

Let it be emphasized that this rearrangement of the rule files has profound theoretical foundations besides being profitable from the purely pragmatic point of view. The basic idea of rearrangement is rooted in a fundamental property of any natural language. Every natural language has a small number of very general laws, each of which has a wide scope in language (holds for a great many words) and a great probability of application in texts. It is precisely such laws that are formulated as general rules. Apart from them every natural language has a great many specific regularities, each of which

has a small scope in language (holds true for a small number of words) and a modest probability of cropping up in texts. Such regularities are formulated as specific and dictionary rules.

3. ETAP-2 works with English. The main difficulty inherent in the switch from French to English consists in the following.

Owing to very insignificant formal marking of the semantic and syntactic relations between words in a sentence English texts do not render themselves so readily to formal analysis as do French texts.

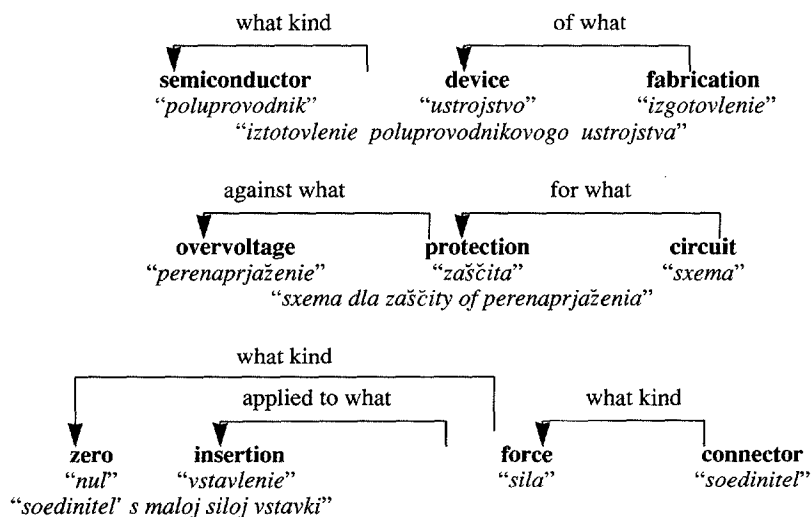
Indeed, English parts of speech have very few specific paradigms. One orthographic word may stand for a noun, an adjective, and a verb, and quite often it has the functions of an adverb and a preposition as well; cf. **ROUND** — “krug” (substantive), “kruglyj” (adjective), “okrugljat’” (verb), “krugom” (adverb), and “vokrug” (preposition).

The endings of different parts of speech are highly homonymous, cf. **rounds** — “krugi” or “okrugljaet’”. Grammatical suffixes are also homonymous, cf. **rounding** — “okrugljaja” or “okrugljajuščij”, **rounded** “okruglii” or “okruglennyj.” Grammatical suffixes may also be homonymous to derivational ones, cf. **rounding** — “ijrygkhaščij” (present participle) and “okruglenie” (verbal noun).

Adjectives and participles have no flexions, and their grammatical agreement does not manifest itself formally, cf. **electric device** “električeskoe ustrojstvo” and **electric devices** “električeskije ustrojstva.”

Subordinative conjunctions and conjunctive words may be omitted, cf. **He claimed he was ill** → “On utverždal, čto byl bolen,” or **The man I saw yesterday** → “Čelovek, kotorogo ja videl včera.” Thus subordinative relations may also be left unexpressed.

In substantive strings of the **cannon ball** type which are so typical of English, syntactic relations between the members are not formally marked, thus giving no clue to their right semantic interpretation. Such constructions, however, are notorious for the multiplicity of ways in which they may be construed and, consequently, translated; cf.



Due to these and similar peculiarities of English the task of automatic parsing of English sentences turns out to be more difficult by an order of magnitude than that of automatic parsing of French sentences. As is clear from the examples quoted above, the choice of the syntactically (and, consequently, semantically) correct structure for a given

phrase cannot be carried out on the basis of syntactic considerations alone. If a specialist finds the right interpretation easily and unmistakably, he does so at the expense of his knowledge of the object domain. But systems of machine translation, at least in their present stage, have no recourse to external knowledge. The only way to obviate the difficulty is to devise some sort of its linguistic substitute. In ETAP-2 the function of such a substitute is fulfilled by the so-called rules of syntactic structure reinterpretation that make wide use of the lexical composition of the sentence.

4. ETAP-2 has been devised to translate not only abstracts and papers but also an entirely new genre of texts — patent headings. At the time of our work on the system this was the only type of information available on magnetic tape. This proved to be an unexpectedly serious difficulty.

It goes without saying that the translations yielded by the system should be not only semantically and stylistically acceptable, but also technically valid. This means that they should be tuned to the terminological stereotypes accepted in electric engineering literature and patent science. We sought for such stereotypes in a number of specialized editions, but with little success. There was little terminological unification there and a lot of mistakes that could be detected with the naked eye. Consider the following two examples.

High voltage direct current transmission apparatus was translated in an authorized source as “*vysokovol'noe ustrojstvo dla peredači po postojannomu toku.*” The translation is next to nonsensical, something like “high voltage apparatus to be transmitted through direct current.” The right translation should be “*ustrojstvo dla peredači postojannogo toka vysokogo naprjaženija.*” **Ringling signal supply** was translated in the same source as “*ustrojstvo dla posylki vyzivnogo signala*” which also falls a long way short of the right “*pitanie dla sistemy zvukovoj signalizacii.*”

In all such cases we had to steer a middle course between the fidelity to the authorized sources and common sense considerations. The reader may judge for himself to what extent we have succeeded.

The main asset of ETAP-2 is, to our mind, the quality of translation. It is a direct function of the linguistic theory underlying the system. The theory we chose is sufficiently well known. It was first proposed in Mel'čuk (1974). We have also made full use of the results of I.A. Mel'čuk and a number of his colleagues reported in a series of publications on the syntactic models of natural languages (Mel'čuk, Pertsov 1975; Mel'čuk, Pertsov 1987; Iomdin, Mel'čuk, Pertsov 1975; Iomdin, Pertsov 1975; Savvina 1976; Apresjan, Iomdin, Pertsov 1978; Iomdin 1979; Uryson 1981; Uryson 1982; Sannikov 1980). In particular, from Mel'čuk, Pertsov (1987) we borrowed the notion and the set of syntactic features, syntactic relations, syntactic rules (syntagms) and the very approach to a description of natural language syntax by a system of translating rules mapping the morphological structure of a sentence upon its syntactic structure (dependency tree), rather than by any generation procedure. We have naturally had to revise many solutions, introduce new features, new relations, new rules and even whole sets of rules, such as presyntactic and preference rules. As a result our model of English syntax has considerably deviated from the prototype, yet basically it remained a component of the Meaning-Text type of model in the sense of I.A. Mel'čuk.

In the algorithmic software we have taken into account certain results reported by O.S. Kulagina (Kulagina 1979).

As we have already stated, the linguistic software of ETAP-2 is logically independent of the algorithms. Yet to give the reader an overall idea of the main components of the system it is convenient to follow the algorithmic order of their inclusion in the process of translation. From this point of view one can single out in the

software of ETAP-2 the following blocks (with the output of the preceding block forming the input of the subsequent block):

- (1) morphological analysis of the input English sentence;
- (2) syntactic analysis (parsing) of the English sentence;
- (3) normalization of the English syntactic structure;
- (4) transformation of the normalized English structure into the corresponding normalized Russian structure;
- (5) expansion of the normalized Russian structure into the full-fledged syntactic structure of the future Russian sentence;
- (6) syntactic synthesis of the Russian sentence;
- (7) morphological synthesis of the Russian sentence.

All the seven blocks of sentence processing require the use of dictionaries. At the stages (1) and (7) use is made of the morphological dictionaries of the system — English and Russian respectively. Before the activation of block (2) the system extracts, in a single scan, the relevant information from the English combinatory dictionary. Every lexeme in the sentence (or, rather, the set of homonymic lexemes in every syntactic position) is supplied with the information necessary for syntactic analysis, normalization, and translation of the English sentence. At the stages (4)-(6) the system falls back upon the Russian combinatory dictionary, extracting the lexicographic information necessary for polishing up the Russian syntactic structure and for its subsequent morphologization.

The English combinatory dictionary contains the following types of information about a lexeme: 1) its ordinal number in the dictionary; 2) part of speech; 3) trivial Russian translation; 4) syntactic features; 5) semantic features, or descriptors; 6) pattern of government, namely, information on the number of the lexeme's semantic valencies, the ways of their manifestation, and the requirements, syntactic and semantic, that the potential actant dependents of the lexeme should meet; 7) names of specific rules, with the information on the values for parametric variables, if there are any; 8) dictionary rules.

The Russian combinatory dictionary includes similar types of information, with the exception that in ETAP-2 there was no translation zone. However, we are working at present on a new version of ETAP designed for machine translation from Russian into English, and in this version the entry's format in both combinatory dictionaries is exactly the same.

From the purely formal point of view the stages (3)-(6) are indivisible because they are carried out by a single program of dependency tree transformation. However, in the present discussion we shall proceed from informal considerations, and from the latter point of view each of the stages (3)-(6) is sufficiently autonomous to be dealt with separately.

(1) The input of morphological analysis is an English sentence in conventional orthography. As a matter of principle, the stage of morphological analysis should have been broken into three substages: a) premorphological analysis (the splitting up of such fused forms as *it'll*, *it's*, etc.); b) identification of the so-called absolute set expressions (see below); c) morphological analysis proper.

For a number of reasons premorphological analysis in ETAP-2 was not algorithmically implemented, so that at present the stage of morphological analysis is composed of stages (2) and (3).

By an absolute set expression we mean, following Kulagina (1979), a string of words having fixed forms, following one another in strict order and expressing a single notion: BY MEANS OF "*posredstvom*," WITH REGARD TO "*otnositel'no*," etc. Such expressions are fused together by means of a special dictionary of absolute expressions

and are assigned an ordinal number by which they are subsequently searched in the combinatory dictionary.

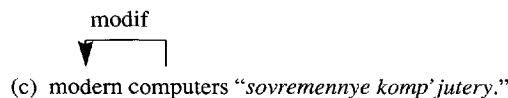
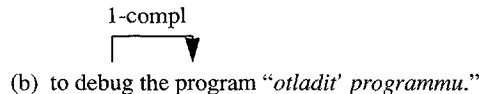
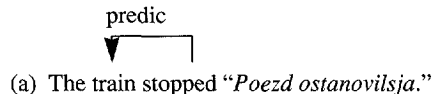
The input of morphological analysis proper [substage (3)] is an English sentence with fused absolute set expressions; its output is an object which we shall conventionally call the morphological structure of a sentence. Strictly speaking, the morphological structure of a sentence is a string of the names of lexemes occurring in the sentence, with a set of morphological characteristics (number, person, case, tense, etc.) assigned to every lexeme, each lexeme representing a certain wordform. Practically, however, what we have to deal with is not a string of lexemes but a string of actual wordforms occurring in the sentence, each wordform representing not a single lexeme but a set of lexico-grammatical homonyms possible within our morphological dictionary. Such a set of homonyms will be conventionally called an elementary morphological structure.

The transition “sentence —> morphological structure” is effected by means of the stem dictionary, lists of standard paradigms and some other devices of compact representation of morphological information.

(2) The stage of syntactic analysis is divided into two substages. The first is presyntactic analysis whose objective is the resolution of lexico-syntactic and morphological homonymy by the nearest linear context. Thus, each of the wordforms **need, needs** represents two homonyms: a substantive (“*potrebnost*”) in the singular and plural and a verb (“*nuždat'sja*”) in the third person singular of the present indefinite tense. If there is an article (THE or A) immediately to the left of these and similar wordforms, they are identified as substantives, and the verbal homonyms are deleted, since verbs do not occur in such contextual conditions.

Presyntactic analysis is the only stage in the whole procedure of translation devoted specifically and exclusively to the resolution of homonymy. The resolution of homonymy is carried on at the stage of syntactic analysis proper as well, but there it turns out to be a by-product of procedures expressly designed for other purposes.

The morphological structure of a sentence, with lexico-syntactic and morphological homonymy considerably reduced, forms the input of syntactic analysis. Its output is the syntactic structure of the processed sentence — a marked and linearly ordered dependency tree. The nodes of the tree are in a one-to-one correspondence with the wordforms of the sentence, while the arcs (or, rather, arrows) binding them represent language-specific syntactic relations of subordination, such as the relation between a predicate and its subject, as in (a), or between a verb and its complement, as in (b), or between a substantive and its adjectival attribute, as in (c):



The main device for turning the morphological structure of a sentence into its syntactic structure are syntagms, *i.e.*, rules that transform two elementary morphological structures representing different wordforms into a hypothetical binary subtree. The application of all the relevant syntagms to the processed sentence yields a set of admissible hypotheses that exceeds by two or three times the number of correct hypotheses. By the correct hypotheses we mean the ones that confront the sentence with the correct syntactic structure. The correct syntactic structure for a given sentence is a dependency tree meeting certain requirements which are imposed upon paired combinations of hypotheses and upon the relative positions of the elements of such pairs within the sentence.

As has been pointed out above, apart from syntagms the syntactic parser makes use of preference rules that choose out of the set of admissible hypotheses the ones that have the highest probability to be correct in the concrete conditions of the processed sentence. They are activated if the application of syntagms has not yielded any dependency tree. Consider the sentence **The bureau bought a computer for 3,000 dollars** "*Bjuro kupilo kompjuter za 3,000 dollarov.*" The syntactic parser will generate for it two hypotheses about the possible heads for the nominal group **for 3,000 dollars: bought for 3,000 dollars** and **a computer for 3,000 dollars**. In such conditions the actant relation (**bought for 3,000 dollars**) is preferred, so the second hypothesis will be filtered out.

Generating the syntactic structure for the processed sentence is the central and most difficult stage of the whole translation procedure. The quality of translation depends precisely on how exhaustively, delicately, and strictly the syntactic subordination relations appearing in the structure mirror the semantic relations between words in the sentence. ETAP-2 makes use of some 50 subordination relations, which, together with highly ramified lexical and morphological information, guarantees a sufficiently full control of the sentence meaning.

(3) The syntactic structure of a sentence makes the input of the normalization stage. It tackles two main tasks.

First, all the morphological and some of the syntactic idiosyncrasies of English are done away with: elements of an analytical form are fused into a single word; such wordforms as **make**, **tell**, **inform**, are supplied with certain morphological characteristics (for example, the characteristic INF in the context of the particle TO2); wordforms with the suffix -ING are interpreted; certain words carrying little or no information, such as the infinitival particle TO2, are deleted; certain new nodes are introduced, for example the node THAT1 in conjunctionless sentences like **He claimed he was ill** (see above); word order is changed; and some other changes are carried out.

Secondly, certain unmistakably incorrect syntactic structures are improved by means of a special block of reinterpretation rules.

To give the reader an idea of the function of reinterpretation rules we shall have to return to the collocations

semiconductor device fabrication

"izgotovlenie poluprovodnikovogo ustrojstva,"

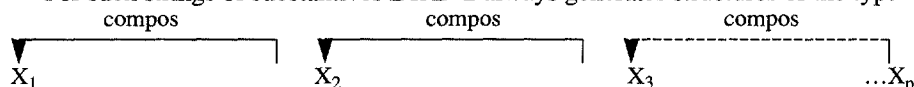
overvoltage protection circuit

"sxema dlia zashchity ot perenaprjazhenij,"

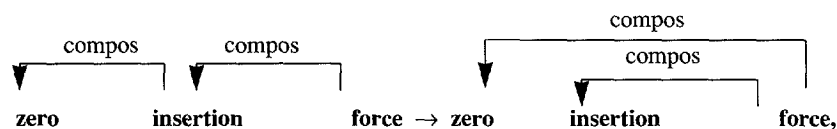
zero insertion force connector

"soedinitel' s maloj siloj vstavki."

For such strings of substantives ETAP-2 always generates structures of the type



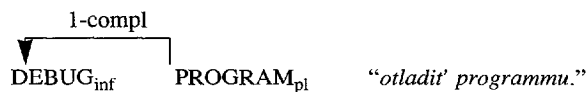
Such a structure is correct for the first two collocations but is wrong for the third. Translating a wrong structure is out of the question, because the result will be senseless. It is precisely in such cases that reinterpretation rules are put into action. They remodel incorrect structures into correct ones on the basis of lexical considerations. In our case the lexeme ZERO “*nul*” by the very nature of things cannot be semantically related to the lexeme INSERTION “*vstavlenie*”: the latter denotes an action, and actions cannot be measured. The most probable candidate into semantic heads for the noun ZERO is the noun FORCE “*sila*.” Force can be measured and it may have a zero value. Consequently, to get the correct structure for the third collocation it is sufficient to subordinate the noun ZERO to the noun FORCE in the context of the noun INSERTION. Using this lexical information, the rule of reinterpretation effects the following restructuring:



which provides for the right translation at later stages.

(4) The stage of transforming the normalized English structure into the normalized Russian structure, or the stage of translation proper, tackles just one task — that of generating the initial structure of the future Russian sentence. To effect such a transition it is necessary a) to replace the morphological characteristics of the English words by the corresponding Russian ones; b) to remodel English syntactic constructions into the corresponding Russian constructions; c) to translate all the lexemes.

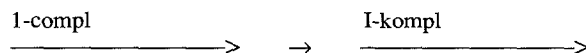
As is obvious, in some cases these three operations are quite trivial and may be carried out independently. Consider the following fragment of a normalized structure:



To generate the corresponding Russian fragment it is sufficient to carry out separately the morphological transformations

$\text{inf} \rightarrow \text{inf}(\text{initiv}), \text{pl} \rightarrow \text{mn} (\text{ožestvennoe čislo}),$

the syntactic transformation



and lexical substitutions

$\text{DEBUG} \rightarrow \text{OTLADIT'}, \text{PROGRAM} \rightarrow \text{PROGRAMMA}.$

It is no less obvious that in other cases, which are apparently quite numerous, the transformations to be carried out may be rather complicated and intrinsically interwoven. We shall adduce several examples of such non-trivial and sometimes non-autonomous transformations that do away with the still remaining morphological, syntactic, and lexical peculiarities of the English language.

a) Non-trivial replacements of morphological characteristics:

PRESENT, PERFECT, PROGRESSIVE	} →	PROŠ (EDŠEE VREMJA),
PAST, PERFECT, PROGRESSIVE		NESOV (ERŠENNYJ VID)

As a result such forms as **has been working** and **had been working** are transformed into forms like *rabotal*.

b) Non-trivial transformations of syntactic constructions, to be illustrated informally:

John being late we decided to start working without him → *Poskol' ku Džon opozdal, my rešili načat' rabotu bez nego;*

They talked me into accepting their proposal → *Oni ugovorili menja prinjat' ix predloženie;*

as many as three years → *celyx tri goda;*

two inches long → *dlinoj v dva djujma;*

five years older → *na pjat' let starše.*

As can be easily observed, non-trivial transformations of syntactic constructions are accompanied, as a rule, by morphological and lexical transformations.

c) Non-trivial lexical transformations, especially in set expressions based on lexical functions in the sense of I.A. Mel'čuk, and in terminological idiomatic collocations, e.g.

on Monday → *v ponedel'nik,*

pay attention → *obratit' vnimanie,*

space resonance → *raspredelennyj rezonans,*

controlling field → *napravljajuščee pole,*

interturn capacitance → *emkost' катуški induktivnosti,*

iron chute → *želob dla zalivki čuguna.*

In this case lexical transformations also go hand in hand with morphological and syntactic transformations.

(5) The next stage, that of expanding a full-fledged Russian structure, tackles the task of generating all the lexico-syntactic peculiarities of the Russian language. It is broken into a number of smaller tasks of which we shall mention the following four:

a) All the necessary lexemes are restored, among them the strongly governed "empty" prepositions and auxiliary elements in analytical constructions; cf.

zaviset' 1-kompl > temperatura →

zaviset' 1-kompl > ot predl > temperatura,

ČITAT', nesov, bud → BYT' analit > ČITAT'.

b) Syntactically non-conditioned grammatical characteristics are generated; for example, in the course of transformation that has just been considered the generated auxiliary verb *BYT'* is supplied with the tense characteristic *bud*, to provide for the generation of such forms as *budet* (*bud*) *čitat'*, *budem* (*bud*) *čitat'*, etc.

c) Words that do not fit the syntactic construction in which they occur owing to the lack of the necessary syntactic function are replaced by the words that do have it; for example, the word *DVA* in the construction *dva sutki*, derived from the English original **two days and nights**, is replaced by the word *DVOE*, since only collective numerals like *DVOE*, *TROE*, *ČETVERO*, etc. can be combined in Russian with such *PLURALIA TANTUM* nouns as *SUTKI* "day and night," *NOČ NICY* "scissors," or *PLOSKOGUBCY* "flat-nose pliers."

d) Constructions which for some reason or other are not admissible in Russian are substituted for by the equivalent admissible constructions. For example, the construction



derived from the absolutely well-formed English construction **He was followed**, cannot be directly translated into Russian and should be replaced by a semantically related although syntactically slightly different construction *Za nim sledovali*.

Let it be emphasized that stages (3)-(5) solve a very important task of syntactic synthesis as well, that of linearizing the processed syntactic structure. As has already been pointed out, a syntactic structure is a linearly ordered dependency tree. The linear order of nodes is given by their ordinal numbers; all the nodes are numbered from left to right. The rules of stages (3)-(5) may bring about some changes in word order which may be elementary or non-elementary. Elementary changes affect separate words. This is the case of certain replacements (**problems discussed** → *obsuždavšiesja problemy*), deletions of articles and “empty” prepositions, addition of new nodes (for instance, the node *BYT* in the analytical future or the node *BY* in the subjunctive mood). Non-elementary changes involve the movement of whole syntactic groups. For instance, to translate the collocation **ferrite-cored anchor** “*jakor’ s ferritovym serdečnikom*,” the attributive group ‘*s ferritovym serdečnikom*’ replacing the modifier group **ferrite-cored** should be transferred from preposition into postposition. Such transformations are accompanied by a renumbering of all the words in the sentence. In this way the linear order of the processed dependency tree is controlled at every stage of its gradual translation from English into Russian. All the changes in word order are attained by local rules. ETAP-2 has no global word-order rules.

(6) Considering the work on linearization that is completed at the preceding stage of translation, there are only two tasks left for the stage of syntactic synthesis: the morphologization of the syntactic structure and the placement of punctuation marks.

As is known, all the morphological characteristics of wordforms are divided into semantically loaded (number in substantives, tense, aspect and mood in verbs, degrees of comparison in adjectives and adverbs) and syntactically controlled (case of substantives, gender, number, and person of verbs, gender, number, and case of adjectives, etc.). Most of the semantically loaded characteristics are supplied for the nodes at the preceding stage of transformation of the syntactic structure. The main task of syntactic synthesis is the generation of the syntactically conditioned (or controlled) morphological characteristics of nodes.

The rules of Russian syntactic synthesis capitalize on the idea that syntactically controlled characteristics of wordforms fall into two main types (Sannikov 1980): they are either canonical, that is, typical for the given syntactic function (cf. the accusative case of the direct object), or else they are the product of regular syntactic alternations of grammatical characteristics in definite contextual conditions (cf. the replacement of the direct object accusative case by the genitive case in the context of negation: *čitat’ knigi*, but *ne čitat’ knig* “not to read books”). It has been observed that in Russian syntax one and the same alternation takes place in different contextual conditions. For instance, the accusative case of a verbal complement or an adverbial modifier is replaced by the genitive case under negation, within certain quantitative groups, and in the partitive context. That is why it appeared reasonable to try and organize syntactic synthesis as a two-stage procedure. At the first stage the system should generate only the canonical form *X* which, at the second stage, should be transformed into *X’* by a single rule taking

into account all the different contexts with the uniform effect upon X. That is precisely the way the rules of morphologization were written in ETAP-2.

Let us turn to punctuation rules. Generally speaking, the principal solution for this problem would be elaboration of the global punctuation rules for Russian. However, for a number of theoretical and practical reasons we decided to do with local rules of punctuation. As has been pointed out above, a similar decision, on similar grounds, has been adopted with regard to word order rules.

In the process of transition from English to Russian ETAP-2 preserves all the semantic punctuation marks: full stops, exclamation and question marks, colons, semicolons, inverted commas, brackets and dashes. The only type of punctuation marks that are deleted in the input English text and generated in the output Russian text are commas. They are generated in most trivial instances: in compound sentences before the subordinate conjunction or the syntactic group of the conjunctive word, in sentences with parenthetical, participial and gerundial constructions, etc. Apart from this there is a small block of punctuation rules closing the rules of syntactic synthesis that takes care of dashes and colons in cases when those two punctuation marks could not have been inherited from English, namely in constructions with the zero copula (*Razrabotka sistemy — bol'soe dostizenie* "The elaboration of the system is a great achievement") and in the so-called clarifying constructions.

(7) Morphological synthesis gets at the input a wholly linearized and morphologized syntactic structure of the processed sentence with all the punctuation marks inserted. Its only task is generating concrete wordforms in conventional orthography on the basis of their morphological structures, that is, the names of lexemes with exhaustive sets of morphological characteristics.

CONCLUSION

To evaluate a system of machine translation it is not enough to look at its practical performance. No less important is its bearing on theoretical linguistics. We have already noted earlier in this paper that every sufficiently sophisticated system of machine translation creates a testing range for linguistics where it can improve its theories and perfect its tools. There was a time when ideas and concepts moved from theoretical linguistics to the applied branches of science. Our time has given us a unique chance to see the beginnings of the inverse movement of ideas, concepts, principles, strategies, and concrete findings from applied linguistics to theoretical linguistics and thus to witness the rise and growth of an entirely new branch of science — experimental linguistics. It would be futile to try to enumerate everything that theoretical linguistics owes to contemporary experimental linguistics and through it to machine translation. We can only say that by a close study of the mistakes the computer makes in the process of translation a number of improvements can be introduced into linguistic theories. In this respect the role of the specifically computerish "negative" material for language study is comparable to the role of aphasia for the progress of psychology. That is why no serious theoretician, whatever his bias, can ignore nowadays the developments in the field of machine translation.

BIBLIOGRAPHY

- APRESJAN, Ju. D., I.M. BOGUSLAVSKY, L.L. IOMDIN, A.V. LAZURSKIJ, N.V. PERTSOV, V.Z. SANNIKOV, L.L. TSINMAN (1985): *Lingvističeskoe obesčenie sistemy ETAP-2*, Moskva, "Nauka," 1989, 295 s.
- APRESJAN, Ju. D., L.L. IOMDIN, N.V. PERTSOV (1978): *Ob'ekty i sredstva modeli poverxnostnogo sintaksisa russkogo jazyka*, Makedonski jazik, t. 29, s. 125-171.
- IOMDIN, L.L. (1979): *Fragment modeli russkogo poverxnostnogo sintaksisa. Opredelitel' nye konstrukcii*, Južnoslovenski filolog, t. 35, s. 19-53.

- IOMDIN, L.L., I.A. MEL'ČUK, N.V. PERTSOV (1975): *Fragment modeli ruskogo poverxnostnogo sintaksisa. 1. Predikativnye sintagmy*, NT1, ser. 2, Inform. processy i sistemy, N7, s. 30-43.
- IOMDIN, L.L., N.V. PERTSOV (1975): *Fragment modeli ruskogo poverxnostnogo sintaksisa. 2. Kompletivnye i prisyjazocnye konstrukcii*, *Ibid.*, N11, s. 22-32.
- KULAGINA, O.S. (1979): *Issledovania po masinnomu perevodu*, Moskva, Nauka.
- MEL'ČUK, I.A. (1974): *Opyt teorii lingvističeskix modelej "smysl-tekst"*, Moskva, Nauka.
- MEL'ČUK, I.A., N.V. PERTSOV (1975): *Model' anglijskogo poverxnostnogo sintaksisa*, Preprint Inst. ruskogo jazyka AN SSSR, N 64-66 Moskva.
- MEL'ČUK, I.A., N.V. PERTSOV (1987): "Surface Syntax of English. A Formal Model within the Meaning-Text Framework," *Linguistic and Literary Studies in Eastern Europe*, Amsterdam, Philadelphia.
- SANNIKOV, V.Z. (1980): *O ceredovaniax v sintaksise (k probleme sintmorfologii)*, Preprint Inst. ruskogo jazyka AN SSSR, N 137, M.
- SAVVINA, E.N. (1976): *Fragment modeli poverxnostnogo sintaksisa ruskogo jazyka. Sravnitel'nye konstrukcii (sravnitel'nye i otkluznye sintagmy)*, NT1, ser. 2, Inform. processy i sistemy, N 1, s. 38-43.
- URYSON, E.V. (1981): "Poverxnostno-sintaksiceskoe predstavlenie ruskix appozitivnyx konstrukcij," *Wiener Slawistische Almanach*, Bd. 7, Wien, s. 155-215.
- URYSON, E.V. (1982): *Napravlenie sintaksiceskoj zavisimosti v ruskix appozitivnyx konstrukcijax*, *Bull. Soc. pol. linguist.*, Fasc. 39, s. 91-107.