Translation to and from Russian: the ETAP-3 System

Igor Boguslavsky

Institute for Information Transmission Problems,

Russian Academy of Sciences

bogus@iitp.ru

1. Introduction

The ETAP-3 machine translation system has been developed by an academic institution and not by a market-oriented company. This accounts for the specific perspective in which the problem of machine translation is treated. This perspective is radically different from the one adopted by most companies or corporations strictly oriented towards developing a commercial software product

The Computational Linguistics Laboratory, which is currently a part of the Institute for Information Transmission Problems of the Russian Academy of Sciences, was founded about 25 years ago by a prominent Russian linguist Jurij Apresjan. Since the very beginning, the primary concern of the laboratory has been more theoretical than practical: in the first instance we have been interested in natural language and its computational modeling. The purpose of the model is to emulate two major linguistic abilities of humans - those of speaking and understanding. This means that it should be able to transform a meaning representation into a text bearing this meaning, and, the other way round, extract the meaning out of any correct text of the given language. This model is conceived of within the framework of the "Meaning ⇔ Text" theory proposed by Igor Mel'čuk and Jurij Apresjan. The development of this theory and its computer implementation by means of an operational model constitute the first dimension of our work.

The second one is an attempt to put such a model to practical use. The result of this activity is a multifunctional ETAP-3 system designed for machine translation and other applications.

This background explains our effort to develop a model in a way as linguistically sound as possible. We try to incorporate into the system much linguistic knowledge irrespective of whether this particular piece of information is essential for

better translation or not. In particular, we want our parser to produce what we consider a correct syntactic representation of the sentence – first of all because it is the truth about natural language. But this is not the only reason. We are convinced (and we have had many occasions to make sure that this conviction is justified) that in the final analysis the theoretical soundness and completeness of linguistic knowledge incorporated in a machine translation system will pay. After a system has achieved a certain (sufficiently high) level of translation quality, every subsequent percent of improvement requires increasingly more knowledge. One of the most important sources of quality improvement is the sophistication of linguistic knowledge.

Very often, one has to pay a very high price for this approach in terms of time and linguistic training of the developers' team. But we are happy that for a long time we could afford this investment. By now, the most difficult and time-consuming part of the dictionary encoding has been to a large extent completed. After that, the addition of new items to the dictionary mostly amounts to the introduction of terminology. This is done by means of a semi-automatic interactive procedure that does not require any special knowledge.

2. ETAP-3 Options

ETAP-3 is a multilingual and multifunctional system. It has several options of which the most important are:

- Machine translation.
- Multilingual communication via an interlingua (UNL).
- Natural language interface to databases.
- Meaning retaining paraphrasing.
- Syntactic checking and correction.

Several comments are in order.

The machine translation option deals with several language pairs: Russian – English, Russian – German, Russian – French and Russian – Korean. By far the most advanced is the first of these pairs . It is served by 50,000- strong dictionaries of both languages and comprehensive grammars. For other language pairs only small-scale prototypes are available.

Another important application on which the laboratory is working for about two years is multilingual communication and information retrieval in the Internet via an interlingua. This is a large international project carried out under the aegis of the United Nations, which aims at reducing the language barrier in the Internet. The idea is as follows. The United Nations University (Dr. Uchida) proposed an interlingua called Universal Networking Language (UNL). To give an example, the sentence

(1) *Long ago, in the city of Babylon, the people began to build a huge tower which seemed to reach the heavens soon*

is represented in UNL as follows:

```
tim(begin(icl>event).@entry.@pred.@past, long ago)
nam(city(icl>place).@def, Babylon(icl>city))
ppl(begin(icl>event).@entry.@pred.@past, city(icl>place).@def)
agt(begin(icl>event).@entry.@pred.@past, people(icl>human).@def)
obj(begin(icl>event).@entry.@pred.@past,
build(equ>construct,agt>human,obj>structure).@pred)
agt(build(equ>construct,agt>human,obj>structure).@pred,
   people(icl>human).@def)
obj(build(equ>construct,agt>human,obj>structure).@pred,
   tower(icl>building).@indef)
aoj(huge(aoj>entity),tower(icl>building).@indef)
aoj(seem(icl>event).@pred.@past, tower(icl>building).@indef)
obj(seem(icl>event).@pred.@past,
   reach(gol>entity,obj>entity).@pred.@begin-soon)
obj(reach(gol>entity,obj>entity).@pred.@begin-soon,
   tower(icl>building).@indef)
gol(reach(gol>entity,obj>entity).@pred.@begin-soon,
   heaven(ant>hell).@def.@pl)
```

Each language should be supplied by two systems: a so-called deconverter which automatically generates texts from UNL and an enconverter which – in an interactive mode – produces UNL expressions from the texts in a given language. Once a text is "translated" into UNL, it will be available for web browsers and generation in any working language. Deconverters for a representative group of languages (Arabic,

Chinese, English, French, German, Hindi, Indonesian, Italian, Japanese, Latvian, Portuguese, Russian, Spanish, Thai) have been already developed and enconverters are under development. The number of the languages covered will grow and it is hoped that ultimately all languages spoken in the United Nations member states will be involved. Our laboratory is working on the Russian component of the UNL system which is conceived of as a module of the ETAP-3 system.

The NL interface to databases translates database queries from Russian (or English) into SQL. It can also produce reverse generation of the NL query from an SQL expression.

The paraphrasing option produces for each sentence a series of variants of conveying the same meaning. For example: *The director ordered him to write a report – The director gave him an order to write a report – He was ordered by the director to write a report – He received an order form the director to write a report.* The paraphrasing is carried out in terms of the so called lexical functions – one of the important innovations of the "Meaning ⇔ Text" theory (see also below).

The syntax checking and correction option is designed to find and eliminate a wide range of errors in syntactic agreement and government.

3. ETAP-3 Salient Features

- Rule-based Approach
- Stratificational Approach
- Transfer Approach
- Syntactic Dependencies
- Lexicalistic Approach
- Multiple Translation

All of these features will be commented upon below.

The ETAP-3 system in its present state is purely rule-based. During the processing each sentence passes through several stages and is consecutively represented by several structures: morphological structure, syntactic structure and normalized (deep-syntactic) structure. The transfer from the source language to the target language is done at the level of the normalized structure. The architecture of the ETAP-3 machine translation module is given in Fig. 1.

ETAP-3 MACHINE TRANSLATION MODULE

GENERAL ARCHITECTURE

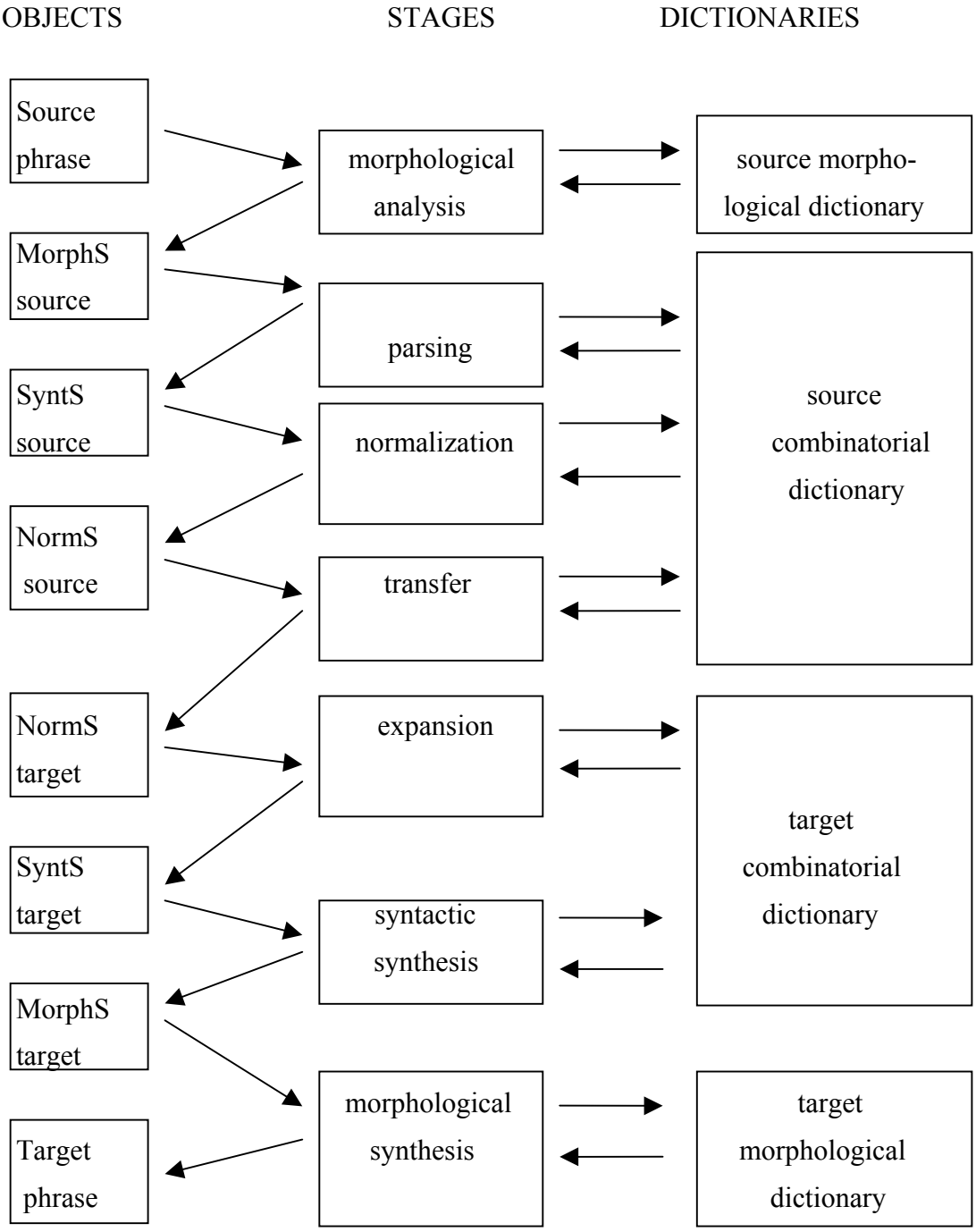

Fig. 1

Syntactic composition of sentences is described within the dependency formalism. An example of a dependency structure for sentence (1) is given in Fig. 2.

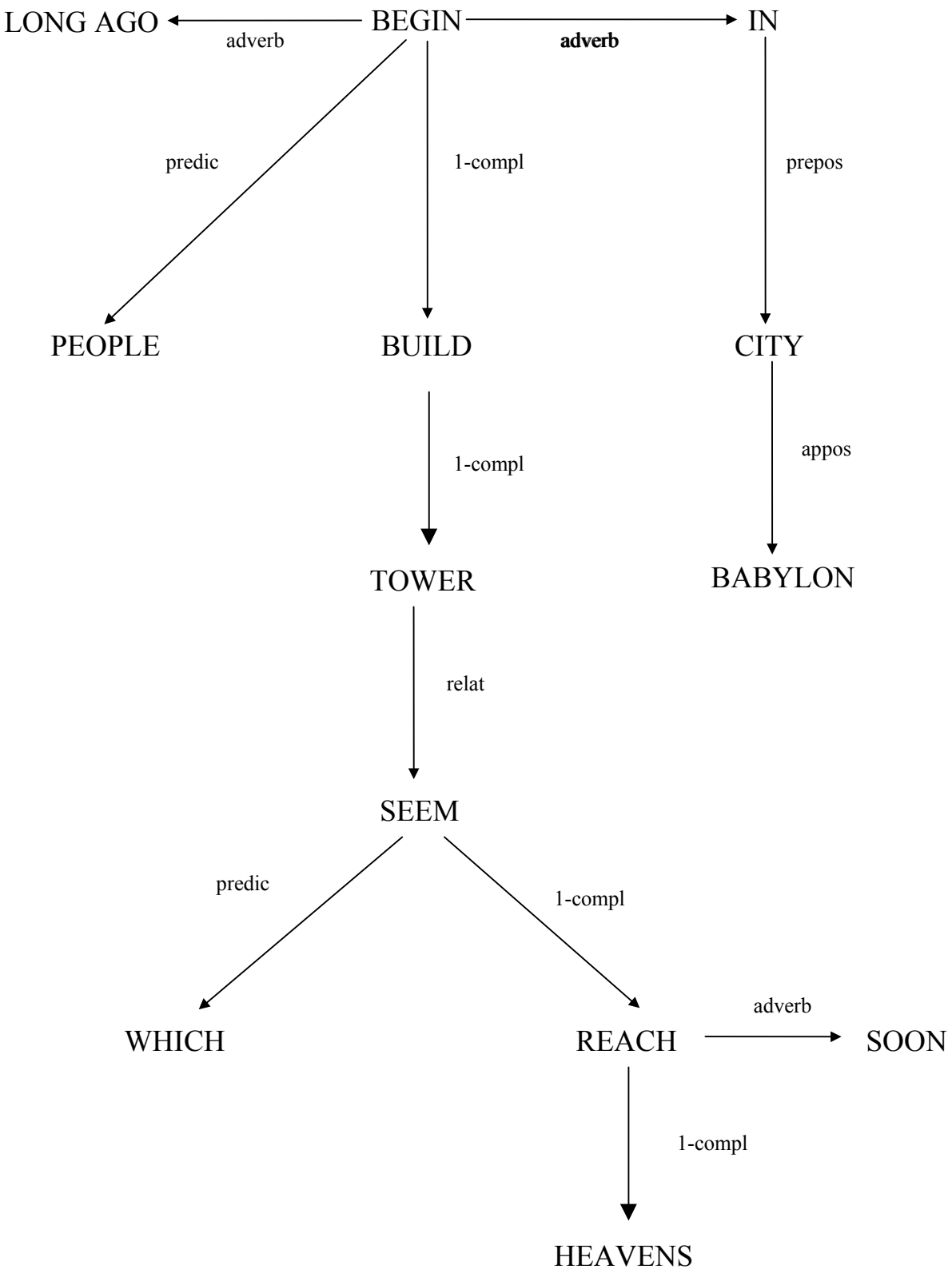

Fig.2

The ETAP-3 system takes a lexicalistic stand in the sense that the lexical information is considered as important as the information included in the grammar. In order to be able to fulfil this role, the dictionary should include much more information than is usual for the NLP systems. The information contained in a dictionary entry falls into the following categories.

- Lemma Name
- Part of Speech
- Syntactic Features
- Semantic Features
- Subcategorization Frame
- Default Translation
- Rules of Different Types
- Lexical Functions

Some of these information types are obvious, others require comments.

Syntactic features characterize the ability/non-ability of a word to make part of certain syntactic constructions. A word can have several syntactic features selected from a total of more than 200 items.

Semantic features are needed to check semantic agreement between the words. The number of semantic features used is about 50.

Subcategorization frame shows the surface marking the arguments of the predicates can have (in terms of case, prepositions, conjunctions, etc.)

Rules of different types are an essential part of the dictionary entry. All the rules operating in the system are distributed between the grammar and the dictionary. Grammar rules are very general and apply to large classes of words. The rules listed or simply referred to in the dictionary are much more restricted in their scope and are applicable to small classes of words or even individual words. This organization of the rules ensures the self-tuning of the system to processing every sentence. In processing a sentence, only those dictionary rules are activated that are explicitly referred to in the dictionary entries of the words making up the sentence.

Lexical function is a certain meaning that is expressed in different ways depending on the word with which it co-occurs. For example, if one wants to attach the meaning 'very much' to the English nouns *rain* or *illness,* one should express it with the word *heavy* (*heavy rain*) and *grave* (*grave illness*) but not the other way round: one cannot

say **a grave rain* or **a heavy illness*. In Russian, one cannot use the word *heavy* (*tjazhelyj*) to characterize the rain. Instead, one should take the word *strong* (*sil'nyj*). This is not because *heavy* in Russian cannot have the meaning 'very much'. It can - it is just this word that is required in case of *illness* (*tjazhelaja bolezn'*). Moreover, *illness* does not co-occur with *strong,* but the corresponding verb *to fall ill* does (*sil'no zabolel*). This shows that the meaning 'very much' can be considered as a function in the mathematical sense of the term: for every word serving as an argument it returns another word which means 'very much' and co-occurs with the given word. There are several dozen lexical functions of this type widely and idiomatically represented across different languages. The co-occurrence information related to lexical functions is strictly lexically bound and should be specified in the dictionary. This information is very useful to obtain idiomatic translation. For example, if one has to translate into Russian the English sentence *He missed a fair chance* one will encounter problems because both *miss* and *fair* cannot be translated literally, being values of two different lexical functions of *chance*. To produce an idiomatic translation of the sentence, it is sufficient to identify the words *miss* and *fair* as lexical functions and find the values of these functions for the Russian translation of *chance.* This is a non-trivial operation, since the words *miss* and *fair* can take different syntactic positions with respect to *chance* (cf. *The chance that he missed seems quite fair*).

Here is a sample dictionary entry for the English noun *chance.*

01626 *CHANCE1*
POR:S
SYNT:COUNT,PREDTO,PREDTHAT
DES:'FACT','ABSTRACT'
D1.1:*OF*,'PERSON'
D2.1:*OF*,'FACT'
D2.2:*TO2*
D2.3:*THAT1*
SYN1:*OPPORTUNITY*
MAGN:*GOOD1/FAIR1/EXCELLENT*
ANTIMAGN:*SLIGHT/SLIM/POOR/LITTLE1/ SMALL*
OPER1:*HAVE/STAND1*
REAL1-M:*TAKE*
ANTIREAL1-M:*MISS1*
INCEPOPER1:*GET*
FINOPER1:*LOSE*

```
        CAUSFUNC1:GIVE<TO1>/GIVE
      **************************
ZONE:R
        TRANS:SHANS/SLUCHAJ
REG:TRADUCT2.00
TAKE:X
LOC:R
R:COMPOS/MODIF/POSSES
CHECK
1.1 DEP-LEXA(X,Z,PREPOS,BY1)
N:01
CHECK
1.1 DOM(X,*,R)
DO
1 ZAMRUZ:Z(PO1)
2 ZAMRUZ:X(SLUCHAJNOST')
N:02
CHECK
2.1 DOM(X,*,*)
DO
1 ZAMRUZ:Z(SLUCHAJNO)
2 STERUZ:X
TRAF:RA-EXPANS.16
LA:THAT1
TRAF:RA-EXPANS.22
```

The last feature of the ETAP-3 system worth mentioning is its ability to present multiple translations when it encounters an ambiguity it cannot resolve. By default, the system produces one parse and one translation that it considers the most probable. If the user selects the option of multiple translation, the system remembers the unresolved ambiguities and provides all mutually compatible parses and lexical choices. To give one example from the real ETAP-3 output: the sentence *They made a general remark that...* , when submitted to the multiple translation option, received two Russian translations that correspond to radically different syntactic structures and different lexical meanings: (a) *Oni sdelali obshchee zamechanie, chto...* (≈ They made some common remark that ...) – (b) *Oni vynudili generala otmetit', chto...* (≈ They forced some general to remark that ...).

To conclude, a few words on the implementation of the system and some practicalities.

Currently, the system runs on a MicroVax computer under the VMS operating system. Its porting under WindowsNT environment is under way. The system is

supplied by a lexicographer's toolkit that allows the user to effectively update the dictionaries of the system.