*6th International Conference on Applied Informatics*
*Eger, Hungary, January 27–31, 2004.*

# Russian-Hungarian Translation System

**József Holovács**

Department of Computer Science, Eszterházy Károly College
e-mail: holovacs@ektf.hu

**Abstract**

The article contains the description of a Russian-Hungarian translation system; the structure and the language background of the system. The translation is based on the knowledge base. The knowledge base contains the rules of grammar, which can be added continuously to the system improving its translating capability. The system takes into consideration cases when morphological and syntagmatic multi-meaning occur. The algorythms are written in Delphi system. Data and rules are stored in VFP files (.DBF).

**Categories and Subject Decriptors:** I.2.7 [Natural Language Processing]: *Machine translation.*

**Key Words and Phrases:** translation system, artificial intelligence

## 1. Introduction

Machine translation is a complicated problem still waiting to be properly solved. Though researches in this field started in the 60's they still couldn't make a machine translation system (MTS) that would be able to produce high quality translations automatically (without human interaction). This is why typical systems can produce only human aided machine translations. Even these systems are of much practical importance. These programs are successful if the text is homogeneous (it is about one field, for example informatics, economics etc.). Modern MTSs using knowledge bases belong to the field of artificial intelligence [2].

**Linguistic background of MTS.**

The most regular and natural translation process consists of three stages:

**ANALYSIS → TRANSFER → SYNTHESIS**

*Aim of analysis*: to build up the linguistic structure of the sentence in the source language and to store it in a certain internal form that can be useful at the forth-

coming stages.
*Aim of transfer*: to transform the structure of the sentence in the source language into the structure of the target language.
*Aim of synthesis*: to create the sentence in the target language based on its grammar structure.

In translation systems linguistic systems are made up of *dictionaries, sets of grammatical rules and formal representations of grammatical units.* These may include *conceptual nets* as well. An important characteristic of MTS is the modular structure of the system. This means that the analysis is executed step by step (morphological analysis, syntactical analysis, semantic analysis, conceptual analysis). The steps of synthesis are the same as those of analysis the difference is that they are applied in reversed order. Modular characteristics also mean that the dictionaries, sets of grammatical rules and formal representations are separated.

*Dictionaries:* the dictionaries of analysis are monolingual, containing a set of words. There are certain information linked to the words – such as word class, inflections, etc. dictionaries of a special field can be separated from each other. Transfer uses bilingual dictionaries.

The operation of automatic translation systems may be based on other principals as well, but this paper does not say more about them. To have more information on them turn to literature specialized in this field [1, 2].

In most of these systems one or both parts of word pairs belong to a world language (English, Russian, German, French). But there are also multi-language systems. Here are some examples:

*DLT system (Holland).* Languages: English, German, French, Italian, Japanese, Chinese, Arabian, Russian.

*SZILOD (Soviet Union).* Languages: English, Spanish, French, Japanese, Chinese, Turkish, Russian.

In the past few years – as the result of the use of internet – the SYSTRAN system became quite wide-spread.

Since there is no *Russian – Hungarian Translation System* (RHTS) on the market, we decided to prepare one. Regarding the complexity of the task, human resources are highly required. For example there were about 100 experts taking part in the development of the *LINGVO* Russian-English (English-Russian) system. Our opportunities are not that wide and we can introduce on the first experimental version of the RHTS now.

## 2. Short description of RHTS

The RHTS does *rule-based transfer translation.*
Rule-based systems contain analytical and translation rules. The software first analyses the sentence of the source language on the basis of grammatical rules given in advance, then executes the transfer tasks and finally, according to the adequate rules, it does the synthesis of the sentence in the target language.

Components of the system:

- Database
- Knowledge Base
- Program Modules
- Explanation Module (will be added in the future)

The database contains the sets of words:

- Set of Russian words
- Set of Hungarian words
- Russian-Hungarian vocabulary

and the related files.
Russian words were prepared according to [3], Hungarian words were prepared according to [4].
The linguistic background is based on *Expert System* elements. This expert system contains a knowledge base. The knowledge base is made up of *rules.*
A rule is able to translate a typically structured phrase.

General form of Rules:

$$\mathrm{r}_1\ \mathrm{r}_2\ \ldots \mathrm{r}_k \Rightarrow \mathrm{h}_1\ \mathrm{h}_2\ \ldots \mathrm{h}_m$$

$\mathrm{r}_i$ = either a constant (concrete Russian word) or a variable (word class with the used form).
$\mathrm{h}_j$ = either a constant (concrete Hungarian word) or a variable (word class with the used form).

The Rules contain also the word order of corresponding words.

• Example:
НА окне(noun, 6. form) ⇒ A(z) ablakon (noun, 13. form).

| Russian Side: | Hungarian Side: |
|---|---|
| HA − constant | A(z) − constant |
| окне − noun variable | ablakon − noun variable |

The rules are typed in by the *expert*. The expert does not input the grammar rule itself but examples of it and the system makes up the rules. It is a good way of improving a system without linguists; those can do it as well who know both languages but are not linguists. Until now no linguists have been involved in the project, so I had to input the basic elements of syntax into the knowledge base but in the future for the further improvement of the system experts will have to take part in it as well.

*Tasks of Program modules*

- **Lexical Analysis** (at the Russian side)
  - turns the russian sentence to words.

- **Morphological Analysis** (at the Russian side)
  - determines the word class, the base form and the used form in the input sentence.

- **Preliminary Direct Translation** of individual words
  Looks up the corresponding Hungarian word.

- **Syntactical Analysis**
  Based on Rules.
  The module selects the minimal-number sequence of rules, which covers the sentence. The result of the analysis is the syntactical structure of the Hungarian sentence based on the rules of the Hungarian side and containing the constants and variables.

- **Generation of a Hungarian Sentence**
  Based on Syntactical Analysis and Preliminary Direct Translation.
  The module substitutes the variables with concrete words on the Hungarian side with the
  help of the chosen rules.

The biggest problem occurring when using the RHTS is the presence of words with more than one meaning.
We have to take into consideration the *morphologically, syntactically and semantically homographic* words.

**Morphologically homographic words**:

| Singular | Plural |
|---|---|
| место | места |
| места | мест |
| месту | местам |
| место | места |
| местом | местами |
| месте | местах |

Three forms of the word are identical. In this case the system cannot choose the correct one from the knowledge base for sure. So we have to take into consideration the preposition of the noun. If it does not help, then we use the Knowledge Base. The System chooses the word which may belong to the longest rule.

**Syntactically homographic words:**

If a word belongs to more than one word class.

- белые (adjective) облака – fehér fellegek
- начинают белые (noun) – a fehérek kezdenek

The System uses the knowledge base again. The System chooses the word which may belong to the longest rule.

**Semantically homographic words:**

- коса kasza
- коса hajfonat, copf
- коса földnyelv

This is the most complicated case.
We do not deal with this problem.
The user may choose from a pop-up list of possible meanings.

The Program Modules are developed in Delphi.
The Database and the Knowledge Base are stored in *.DBF files.

# 3. Appendix

A few examples on the operation of the system.

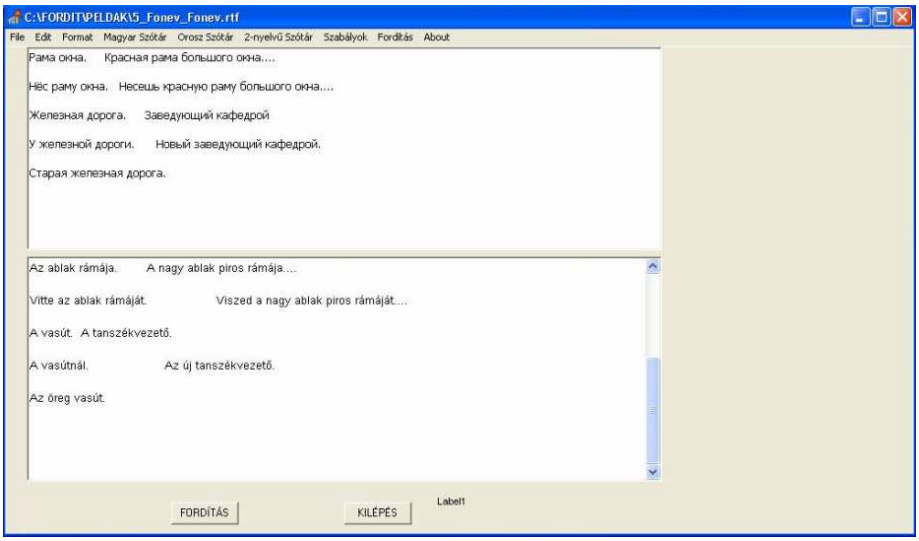


Figure 1: Translation

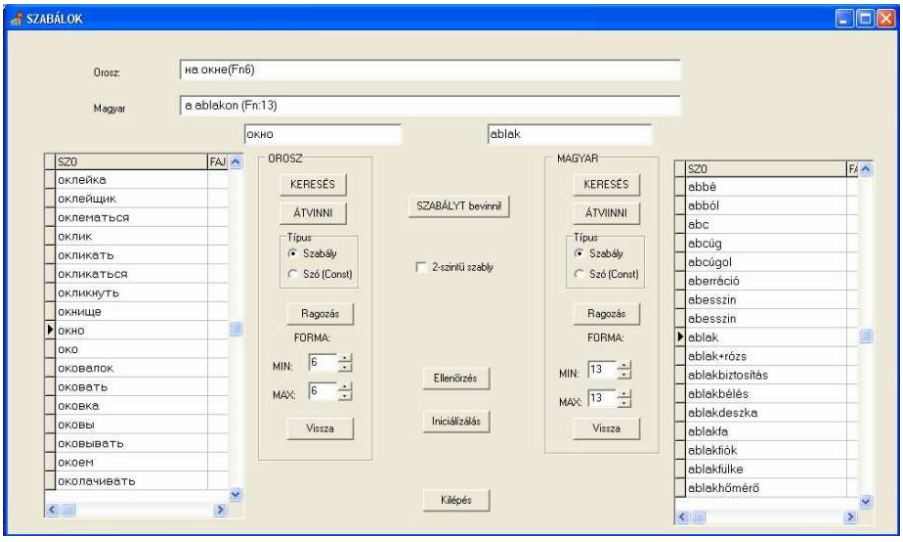


Figure 2: Rule input

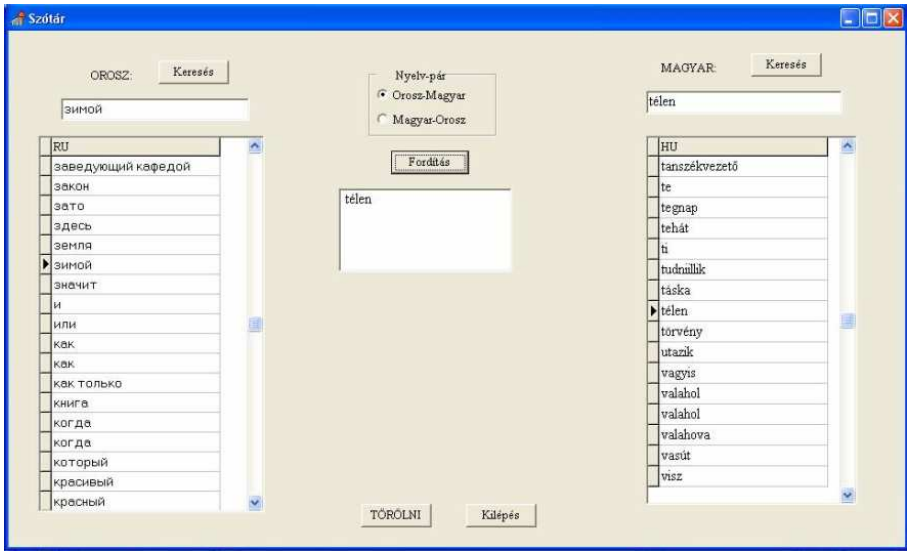


Figure 3: Bilingual dictionary

# References

[1] Prószéky Gábor- Kis Balázs: *Számítógéppel – emberi nyelven. Intelligens szövegkezelés számítógéppel*, SZAK Kiadó, 1999.

[2] Искусственный интеллект: ред. Попова Э.В., Москва, Радио и связь, 1990.

[3] Зализняк А. А.: Грамматический словарь русского языка, Москва, Русский, язык, 1977.

[4] Elekfi László: *Magyar ragozási szótár*, MTA Nyelvtudományi Intézete, Budapest, 1994.

## Postal address

**József Holovács**
*Department of Computer Science*
*Eszterházy Károly College*
*Eszterházy tér 1., 3300 Eger*
*Hungary*