

# The New Age of Machine Translation, or Mission Possible

Machine Translation, or MT, of human languages has been around for more than half a century. Nevertheless, there are still many people who criticize the technology and persistently argue that the computer cannot compete with human translators. It is interesting to note that there were moments in MT history when even the experts themselves said that the technology had no future. Yet, the energy and talent of the developers have finally conquered the skepticism, and today MT technology is rightfully acknowledged to be efficient by computer experts and millions of MT software users around the world.



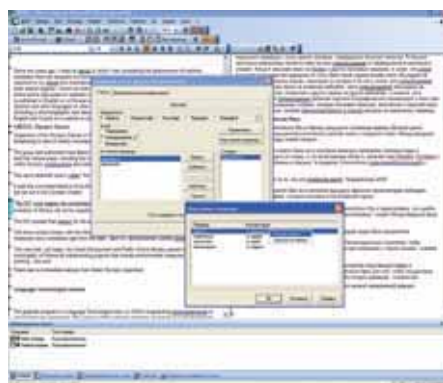
At present, there are dozens of MT developers in the world. In Russia, machine translation software is produced by the PROMT company. This company is well known to users because of its translation system PROMT, formerly named Stylus. This translation system has been on the market for 14 years, and the translation quality and functional characteristics have notably improved. At the end of last year, the company launched a new line of translation systems PROMT 7.0. The launch of the new line was accompanied by confident statements of the developer that new translation technology was created and new heights in MT were reached. Let's have a closer look at the improvements and advances in MT technology that the developers claim to have made.

### Architectural Achievements

PROMT translation systems always had a large<sup>1</sup> main dictionary (or General dictionary, as the developers call it). This can be explained by the fact that translation depends on the size and quality of the dictionary. It is important to be able to customize the dictionary, so that the system translates correctly regardless of the topic of the given text.

PROMT never allowed users to edit the General dictionary. But a user could create his/her own user dictionaries and add words with the required translations to them. For example, when it was necessary to specify that in the given context the word *switch* had to be translated into Spanish as *vara*, and not as *interrupcion*.

In the previous version, all translation variants were used in the translation process. This could sometimes make the choice of the correct variant difficult. As a consequence, the translation might be incorrect. In the PROMT 7.0 version, the situation has radically changed. The dictionary now contains two types of translations, active and inactive (Fig.1). Active translations are used for translation, and inactive translations are stored in the dictionary. A user can make an inactive translation active, and vice versa, with one mouse-click. For example, the most common Spanish translation of the English verb *go* is *ir*. This variant of translation is active (Fig.2). In addition, the dictionary gives several other inactive translation variants (*marcharse, salir, pasar, and others*), which are used less often. And this is not all! A system dictionary may contain an unlimited number of translation variants for any given word. Consequently, customization of a dictionary has to do with the selection of translation variants that will be used for translating a given text.



(Fig.1)

This new dictionary architecture is certainly a breakthrough, because it makes an MT dictionary similar to a traditional electronic dictionary. On the one hand, an MT dictionary may now contain an unlimited number of translations, on the other hand, the numerous translations are not 'piled up' (earlier the system showed all translation variants in brackets near the main translation). The new dictionary architecture is called "multidimensional" because there are two levels of translation variants – active and inactive.

As an added convenience, a user can add comments to every translation variant. In a comment, the user can specify the context where the translation variant is added, the date when the translation variant is added, or just some suggestions. The comments can be useful when more than one user is working with the system, so that everyone can see who entered each translation variant.

### Architectural Achievements

The long-time PROMT users probably remember that the previous versions of the program 'created' very peculiar word-forms. It was due to the absence of strict rules for word recognition, that the system could recognize the word as it 'wanted'. For example, the English word 'radio beacon' could be recognized in the following variants:

- radio beacon
- radio beacons
- radio beacons
- radios beacon
- radioes beacon
- radios beacons
- radioes beacons
- radioes beacons
- radios beacons
- radioes beacons



(Fig.2)

Only the first two variants in the list are correct. The other forms are 'invented' by the system itself.

In the new version, there are only canonical forms of every word in the dictionary, supplemented by the strict rules for recognition of all forms. For example, the English verb 'make' is represented in the dictionary only by the infinitive form, and the system can conjugate it correctly. In other words, the system's intelligence has substantially increased.

### Divide and conquer

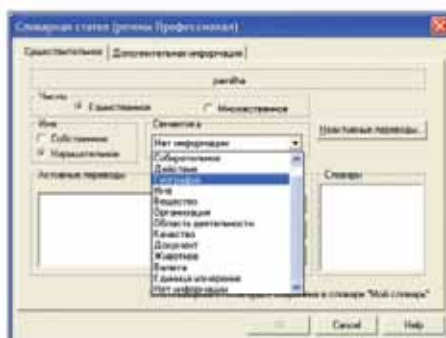
One of the main problems the MT systems encounter is polysemy of words in any natural language, as well as existence of set expressions. Sometimes it is not enough to add a set expression to the dictionary because there may be insertions between its elements. Such collocations are called phrases with insertions. For example, in the phrase *They took their clothes off* the collocation 'to take off' is split by 'their clothes'. Earlier, the translation system 'stumbled' over such expressions and gave word-for-word translations, which was incorrect (in Spanish, *Ellos tomaron su ropa lejos*). The correct Spanish translation is *Ellos quitaron su ropa*.

In PROMT 7.0, this problem is solved once and for all. First, the system now "understands" many set expressions with insertions (e.g. *keep promise, pay bill, keep away*). Second, a user can add new phrases with insertions to a dictionary him/herself. For example, in the sentence *We will pay them back for the trick they played on us* the phrase *pay back* is a phrase with insertion. In order to enter this expression into the dictionary, highlight *pay back*, right-click and select the command *Entry* from the context menu.

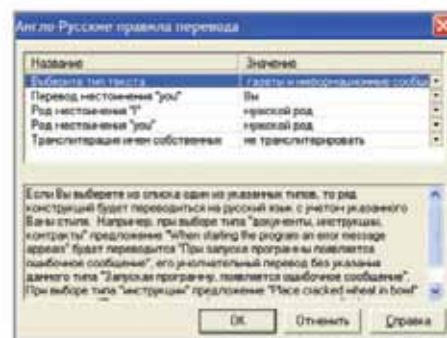
1. For example, according to the developers, the PROMT 7.0 English-Russian and Russian-English General dictionary has 680,000 words and collocations.



(Fig.3)



(Fig.4)



(Fig.5)

Next, select the user's skill as **Professional** and specify the part of speech ("verb"). Select the **Phrase with Insertion** check box (Fig. 3) and click the Choose phrase type button. In the Phrase with Insertion dialog box, select the type of phrase with insertion. After that, enter the translation in the dictionary.

### Be class-conscious

One more new feature of PROMT 7.0 dictionaries is the possibility to specify the semantic class of a word (Fig.4). Let us consider the phrase *I bought some wine*. If the program does not 'know' that wine is a substance, the translation of the word some will be the most general one, *algún* (*Compré algún vino*). If the semantic class 'substance' is specified, the translation will be correct, *Compré un poco de vino*.

In the same way, you can specify the semantic classes for geographic objects, names, animals, currencies and other objects. To do it, go to the dictionary entry of the selected word and select the semantic class from the **Semantic class** list.

Of course, even if you have not specified the semantic class or specified it incorrectly, the translation will be understandable. But this option gains in importance if you have to translate large texts, and you want the translation to be maximally correct. The less *algún vino's* generated by the system, the less time and effort spent on editing.

### Do It Yourself

As you may have noticed, enhancement of the system's controllability is the mainstream of the PROMT 7.0 new system. The developers have paid special attention to different settings which enable users to customize the system according to their needs. In addition to the selection of translation variants and specification of semantic classes, users can choose the translation rules for a specific text.

It is evident that the writing style of a cooking recipe is totally different from that of a business contract. A user can select the translation rules not only for the entire document, but also for every paragraph in particular.

To choose the translation rules for the document, select **Topic -> Document Translation Rules**. To specify the rules for the paragraph, right-click and select the **Paragraph Translation Rules** command on the context menu.

The user can choose between the following settings (Fig.5): text type (documents, newspapers, private letters), translation for the pronoun you (*usted, tú, ustedes, vosotros*), gender for the pronouns I and you, and transliteration of proper names.

In connection with translation rules, we should also mention the possibility of selecting the British or American variant of the English language (this option appeared in the previous version of PROMT).

### Translation of graphic files: a dream that has come true!

Initially, PROMT processed only text documents. Later, Microsoft Office 2000-2003 formats and HTML-pages were added. But translation of graphic files has been an impossible dream. The demand for translation of graphic files emerged after the PDF-format became increasingly popular in technical and business documents. How can a graphic document be translated? One should print, scan, recognize, convert the text into electronic form and finally translate it. The procedure is too complicated, especially if the document contains tens or hundreds of pages.

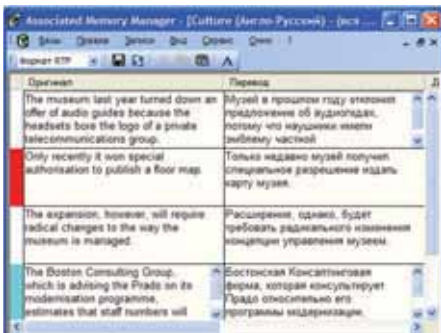
In PROMT 7.0, the problem is solved! You won't believe it, but PROMT 7.0 can translate graphic files! Thanks to the built-in OCR-system from the I.R.I.S. company (Readiris), PROMT now 'understands' (Fig.6) documents in PDF, TIFF, BMP, JPG and other formats. A new line – Image file – has been added to the list of supported formats. After you select this type of file, the built-in OCR system is launched, which extracts the text from the graphic file and transmits it to the translation system. The recognition does not take much time (the speed of recognition of PDF-documents depends on how much graphics is in the document). In one or two minutes you obtain the document in electronic form. Then you can translate, edit and save it as a text document.

We should note that in this version the OCR-system does not work with a scanner. It recognizes only ready graphic files.

### XML-files are no longer X-files

In addition to text documents and graphic files, PROMT 7.0 translates XML-files. At present, the XML format (Extensible Markup Language) is becoming more popular and more widely used in data presentation. An XML-file is a common text file which is created according to certain rules and which contains data and their structure about a certain object (e.g. an invoice). XML-files are most often used to process and store large volumes of documents, because the text contents and the format information are stored separately in a database.

The increasing popularity of the XML-standard resulted in the demand for translation of XML-documents. In PROMT 7.0, the developers added an application that can be used to create and edit the files with translation rules for translation of XML-documents.



(Fig.6)

Unlike files of similar formats or Word-documents, XML-files do not have a predetermined structure, that is why additional information about translation of a particular file (or a set of files with a common structure) is required for translation of XML-files. This information is contained in the file with translation rules. It describes the rules for translation of different tags and their attributes.

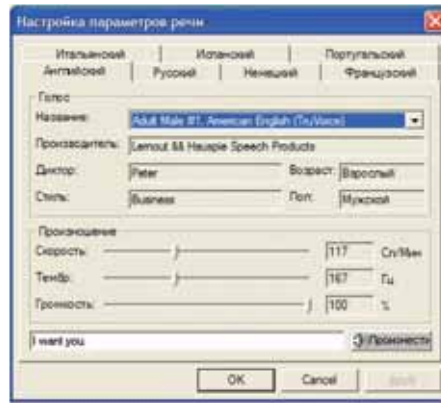
### Talking Program

In addition to recognition and translation of documents, the PROMT 7.0 program can speak the selected text using Microsoft Agent technologies and Text-To-Speech engines. The PROMT distribution set contains the speech modules for all the languages that are included in the set (for example, PROMT Giant has six speech modules for English, German, French, Spanish, Italian and Russian).

The reading program allows the user to customize voice speed, pitch and volume (Fig. 7). You can also use the mode when the text is read by a pop-up Microsoft Agent character, so that the process is more vivid. When a Microsoft Agent pop-up character is used, the spoken text is also displayed in a pop-up window. The reading option can help users save their time and not strain their eyes reading text on the screen. Of course, you will need a sound card and speakers or headphones to hear the computer talk.

### What's the moral of the story?

So, a skillful computer translation is no longer a dream. It has become reality. This review of PROMT 7.0 innovative ideas demonstrates that many of the tasks that seemed impossible some decades ago are successfully accomplished today. The system has become more convenient and user-friendly. No doubt, the developers have taken into consideration the many wishes and needs of modern users. The result is a high-quality and easy-to-use product.



(Fig.7)

### Machine Translation - Faster, Higher, Stronger!

If we look at the stages of MT technology development, we see that every new version of a translation program is a step forward in translation quality and user-friendliness.

There is one thing that remains: even the best machine translation program cannot compete with a human translator.

The MT systems have a different mission. They are designed as a reliable and convenient tool for overcoming the language barrier. In fact, a translation program solves two problems.

First, it gives a quick and rough translation, when it is necessary to understand the general sense of a text in a foreign language. Second, it serves as a powerful tool to enhance the efficiency of professionals' work (business people, engineers, professional translators).

The first task was successfully accomplished long ago.

As for the second task, the developers constantly improve the system and offer more and more settings for customizing the PROMT translator in accordance with the subject area of a text.

The more convenient it is to work with the translator, the more satisfied users there will be.



### How to customize PROMT 7.0

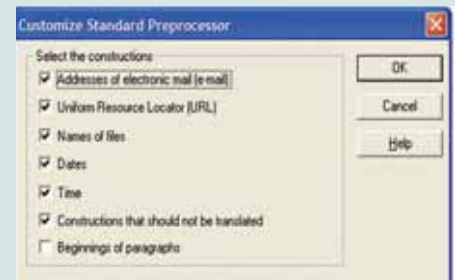
To customize the translation program for your texts and enhance translation quality, the developers suggest that you use the following options:

- 1. Attaching the specialized dictionaries** developed by the PROMT company. At the moment, there are more than a hundred dictionaries for different subject areas. The developers advise that you use the corresponding dictionary when translating a text in a specific subject area.
- 2. Creating user dictionaries.** This can be very efficient when you translate a very highly specialized text.
- 3. Preserving words that do not require translation.** A preserved word is marked in the input text as a word that does not require translation. This procedure is most often required for proper names, foreign words, etc. It is especially important to preserve proper names when they coincide with meaningful words (for example, 'Windows XP', 'Miami Beach', 'Bill Gates').

**4. Selecting the translation rules** (See "Do It Yourself").

**5. Using Associated Memory databases.** Associated Memory is used to store phrases, sentences and even paragraphs, together with their translations, and use them later during translation of texts in a given subject area.

**6. Using Preprocessors.** A preprocessor is a set of rules that are used for translation of e-mail addresses, URLs, file and folder names and other text constructions (Fig. 8).



(Fig.8)