The PARS MT Family: Practical Usage

Michael S. Blekhman (Paper not presented)

PARS is a family of machine translation systems developed by Lingvistica ’93 Co. for the following language pairs:

- Russian to and from English, German, and Ukrainian;
- Ukrainian to and from English and German.

The systems run on IBM PCs and have the following characteristics:

- they run under Windows 3.1 and above, Windows 95, Windows NT, in stand-alone and network modes; DOS versions are available for PARS/ER, PARS/U, and PARS/RU;
- PARSes are designed to make draft translations of scientific, technical, business, and socio-political texts in the subject areas covered by their dictionaries; in particular, PARS/ER bidirectional dictionaries relate to such subject areas as business, mathematics, physics, chemistry, aviation, engineering, automobile building, oil/gas, etc., altogether over 900,000 in each part, English-Russian and Russian-English; **a very large general dictionary of about 300,000 words and idioms** will be made available to the customers later this year;
- the translation modes are «from file to file», «from Clipboard to Clipboard», and «from WinWord to WinWord»; in the latter mode, PARS is started directly from WinWord 6,0, WinWord 7.0, or WinWord 97, and the target file is placed in a separate Word window under the source one, preserving the source text formatting; in the «Clipboard to Clipboard» mode, PARS can translate HTML files and Windows screen Helps, as well as files generated by all Windows-based text processors such as Word Perfect, Write, etc;
- polysemantic words and phrases are marked with asterisks in the target file so that the user could select a more appropriate translation from the panel of optional translations;
- each system has a flexible dictionary editing program;
- a special dictionary compilation technology is used to develop new dictionaries; one of the major sources of terminology is a set of Polyglossum dictionaries supplied by our partners, ETS Publishers, Moscow, Russia.

Practical usage

I made a kind of classification to outline the circle of PARS users. Unfortunately, any kind of serious statistics is impossible due to awful computer piracy in ex-Union. The only thing I can be «proud of» is that PARS as well as Polyglossum are very popular with the pirates: Igor Fagradiants, director of ETS Publishers, claims that about 300,000

piratic compact disks with our systems have been made in Russia since 1996 up to the present time.

Individual users

A very numerous subgroup is made up by *students who need their diplomas and other kinds of papers to be translated from Russian into Ukrainian*. We hope to meet their requirements with the COPERNIC CD-ROM and convince at least some of them to abstain from using piratic disks. COPERNIC is a project launched last year jointly by the Ukrainian Ministry of Education and Lingvistica '93 Co. The disk comprises the basic versions of each of the 5 PARSes (without specialist dictionaries), it is supplied with a user's guide, and costs $12 for school, college, and university students, and $28 for the rest of customers.

Some people want to communicate with people living abroad. PARS/U, is bought, in particular, by Americans and Canadians wishing to communicate with their friends and relatives residing in Ukraine. One of them told me: «They speak Ukrainian, while I speak English. The only way to communicate is to use a computer program». I wonder if one of the international pen pal organizations might be interested in using PARSes for communication purposes. It would certainly require serious modifications to the systems in order to take into account peculiarities of this style, but the idea itself seems rather promising to me.

Professional free-lance translators make up another subgroup, though less numerous. Their language pairs are mainly English, German, French, Italian to and from Russian. Some of them like MT systems, some prefer MAT software (electronic dictionaries such as Polyglossum), while others buy both. My opinion is, however, that the majority of this group are still our **potential** clients. The fact is that the foreign languages departments of Ukrainian universities train people who are good at languages but have no idea of the computer as translator's everyday tool. Introducing elements of language engineering at such departments would contribute a lot to expanding the circle of our conscientious clients!

There is a group of *individual users who require Russian to English translation of scientific texts.* Here is an example. A scientist asked me to translate his medical paper for submittance to a serious British journal. When I looked through the text, there was only one thing which I understood - I could not do without PARS because the paper was abundant in «awful» medical terms. I faced a dilemma: either to translate the text manually looking every second or third word up in the Polyglossum Russian-English medical dictionary, or to let PARS make a draft translation and post-edit it. I chose the latter variant, and the paper was accepted.

Corporate users

MT and MAT systems seem to be very popular with corporate users. Generally speaking, all kinds of *organizations, both state-owned and private, use PARS/RU for translating official documentation, including that of financial, scientific, and technical nature, between Russian and Ukrainian.*

Many *Ukrainian banks use PARS/RU for translating financial documentation, such as official instructions, between Russian and Ukrainian*. Here is another example. In autumn, I installed PARS/RU in one of the banks in the town of Saki, the Crimea. They

use it to translate megabytes of instructions they receive electronically from the Ukrainian National Bank. Those texts are written in Ukrainian, the country's state language, and the problem is that many people in the Southern and Eastern parts of Ukraine doesn't even **understand** Ukrainian, to say nothing of speaking it.

A tendency that gains popularity is making *MT systems part of integrated products*, such as PRAVO, a system very well-known in Ukraine. It is supplied on CD-ROM and comprises the full set of Ukrainian laws and decrees, with a retrieval system and our Ukrainian-Russian translation module. Later this year, Ukrainian to English and German modules will be added in May.

I am especially proud that PARS/ER is used for translating Russian medical abstracts into English for the *Medical Practice* journal published in Kharkov. I do it myself, first running the texts through PARS and then post-editing the raw translations. *Using MT systems for translating abstracts in scientific journals* may become a tendency.

Large plants and design bureaus that export their products are among the users of the PARS/ER system. The Antonov Aviation Design Bureau in Kiev as well as the Yangel Spacecraft Bureau in Dnepropetrovsk are among them. We supplied PARS/Avia to them, which includes the core Russian-English-Russian system and a number of terminological dictionaries on aviation, space, communications, etc. Their reaction is very important for me: they say that PARS is better for translating technical documentation, while Stylus by ProMT is preferable for business correspondence. Well, we'll try to be up to the mark in all the aspects!

A new tendency is using PARS *to translate Russian textbooks and lectures into English for foreign students coming to sudy at our universities* (see below).

MT can and should also be used for purely academic purposes. An example is using PARSes at Kharkov State Polytechnical University in the course of machine translation at the Department of Intelligent Information Systems. Presently, we are going to set up a department of language engineering at Kharkov Slavonic University. I plan to implement all our systems there.

Access to Internet and E-mail will contribute to a higher role of MT. However, this will require not only technical (which is comparatively simple) but also linguistic solutions because colloqial texts, which are very often to be found on web sites, to say nothing of E-mail messages, are very hard to translate automatically. I am sure that Internet and MT will stimulate each other greatly. And this application is very promising. The fact is that Internet resources are in fact unaccessible to Russian and Ukrainian speaking scientists because of the language barrier, and so are the Russian and Ukrainian publications for the English-speaking community. You should take into account that the state system of scientific information, which was the pride of the former Soviet Union, does not exist in Ukraine for a number of reasons, so Internet will be a very good, though not the only source of information if the decision will be taken to build up such a system in this country.

In 1996-1997 Olga Bezhanova, my elder daughter, used PARS to produce draft translations of large scientific and technical text corpora. She described her experience elsewhere (MT News International, Proceedings of the MT Summit VI). In this paper, I would like to summarize the results she obtained.

Post-editing PARS-made scientific translations

This work was ordered by The Russian Foundation for Fundamental Research (RFFR). Olga was supposed to translate the «RFFR Annual Bulletin» that comprised about 400 pages presented in the WinWord 6.0 format (about 1.2 MB), namely the titles and bibliographic data for several thousand research projects in the following areas:

- mathematics and information science;
- physics and astronomy;
- chemistry;
- biology and medicine;
- geosciences;
- liberal arts;
- databases and books issued in Russia.

Each document in the Directory comprised approximately 5,500 pieces of information, each consisting of the author's surname, project title, identification number, name of the institution (University, research institute, etc.), and the city/area of residence.

Thus the task generally consisted in translating not complete texts but the titles of research projects, **each title having two to fourty words**.

The customers required a similar English text presrving the source text styles and formatting. The customers also stipulated that **the surnames were to be transliterated according to the rules of the English language**, while the titles of institutions were to be translated.

The work was supposed to be done within approximately a month. Taking into account the days-off, the translation was made during 34 days, 5-7 hours a day, consisting in **post-editing the texts translated by PARS**. Numerous misprints in the sorce text (the better half of which composed Latin letters instead of Cyrillic ones in Russian words) slowed down the whole process.

A great number of scientific terms in the source text relating to numerous subject areas required using quite a number of various dictionaries. Olga says that **translating texts of such volume by one person for such a short period of time without using machine translation software would be impossible**.

Before the translation session, the Polyglossum system of dictionaries was activated on CD-ROM, which made it possible to access any of the dictionaries without exiting from WinWord.

Words not found by PARS were looked up in Polyglossum dictionaries.

When Olga was making the translation, PARS lacked several terminological dictionaries that were entered into the system later. They would have increased MT quality greatly. The names of these dictionaries are given for the corresponding subject areas.

Here are the results for each subject area.

Mathematics and information science

This chapter of the Directory comprised 800 titles of research projects in the above subject area.

In order to translate this chapter, the following dictionaries were set up in PARS (in a descending order of priorities): computer dictionary (25,000 terms in each part, Russian-English and English-Russian), technical (76,000 terms), general (35,000 words and phrases).

The mathematics dictionary (85,000 terms) was made later.

It is to be noted that the dictionaries used did not cover the subject area completely, so Olga had to refer to the Polyglossum dictionaries when post-editing the documents on mathematics. As to those on information science, they were translated by PARS fairly well.

The main difficulty when working with this chapter consisted in translating phrases comprising surnames of «foreign» mathematicians, as, for example, *Langevin equation*. Because many similar phrases were absent both in PARS and in Polyglossum, Olga had to look them up in The Great Soviet Encyclopaedia, which presents names of well-known scientists in their native languages.

Physics, astronomy

The following dictionaries were set up in PARS for translating this chapter (consisting of 1290 titles): technical, radioelectronics (50,000 terms), microelectronics (20,000), general.

Due to the absence of a special dictionary on physics and astronomy in PARS (the dictionary on physics, 80,000 terms, was developed later), post-editing this chapter was more difficult. The Polyglossum dictionaries, comprising about 1,500,000 terms of various subject areas, were of great help.

The main problems arose in rendering the names of the planets and their satellites, which Olga managed to find in the English-Russian astronomy paper dictionary.

Chemistry

For translating this chapter (659 titles), the technical and general dictionaries were set up in PARS. The chemical dictionary (50,000 terms) was not yet present in the system.

This chapter was the most difficult to translate since it comprised quite a lot of specific chemical terms, such as *фталоцианин (phthalocyanine), редокс (oxidation-reduction), рацемат (racemoid), аценафтен (acenaphthene), гваяцил (guaiacyl)*, etc.

Olga had to look up the words not found either by PARS or by Polyglossum in the Russian-English Dictionary of Chemical Reactions and in the English-Russian Dictionary of Petroleum Chemistry and Processing because, generally, the difficulty consisted in the spelling of the unknown chemical terms.

For example, it was clear that the English translation of the term «стирил» could not differ seriously from the Russian variant, but the translator was not sure whether it was «styril» or «stiril». She found the word «styryl» in one of the paper dictionaries, which put an end to the troubles.

It was very difficult to translate complex terms consisting of several components, for example, «винилхалькогенополигалогенбензол». PARS failed to translate such words, that is why it took Olga 6 days to post-edit this comparatively short chapter.

Coming across a word consisting of several components, the translator usually broke it in sense-bearing parts and translated them in turns. Thus, the term

винилхалькогенополигалогенбензол

was broken into *винил, халькоген, полигалоген,* and *бензол*. The resulting «simple» words were translated and united in one. It's only natural that such work was very labor-intensive and occupied much time.

Biology, medicine

This chapter (908 titles) was the second most difficult to post-edit. The following PARS dictionaries were used for translation: medicine (20,000 terms), aviation medicine (24,000 terms), technical, general.

The chemical dictionary as well as that on biotechnology (10,000 terms) was entered into PARS later.

The main difficulty consisted in translating the names and genders of animals and insects. Again, Olga could not do without the Russian-English paper dictionary by A.I. Smirnitski, in which she found such terms as «иглокожие», «ракообразные», - «echinodermata», «crustacea», etc. The dictionary comprises quite a number of biological terms.

Geosciences

This chapter comprised 752 projects in such subject areas as geology, paleontology, archaeology, ecology, etc. PARS translated this chapter very well, owing to the presence of geological and ecological dictionaries; this raised the translation quality several times as compared with translating such chapters as «Chemistry» and «Biology, Medicine».

The chapter was translated in two stages. It was split into two portions of nearly equal sizes that were translated by PARS using the following dictionaries:

The first portion: geological (11,000 terms), technical, general.

Embarking on the translation of this chapter, Olga didn't yet know that it comprised many documents on ecology, that is why the ecological dictionary was not chosen for translating the first portion. When post-editing it, she saw that the better half of the words not found in the system dictionaries related to ecology, so she also set up the PARS ecological dictionary (18,000 terms) for translating the second portion.

The second portion: technical, geological, ecological, general.

By the way the dictionary on oil and gas (70,000 terms) was developed later. More than that, the geological dictionary was extended to comprise 27,000 terms.

When post-editing this chapter, Olga actively used The Random House Unabridged Dictionary to clear up the spelling of such words as *Темис (Tethys), пегматит (pegmatite)*, and many others. In particular, she made use of the table of geological periods given in this dictionary. It is to be noted that this was the only source where she managed to find translations of a large number of geological terms. It only took her 4 days to translate this chapter, which would have been impossible without using the Random House dictionary.

Humanities and social sciences

This chapter (214 titles) was the easiest to translate. Two PARS dictionaries were used: general and economy (55,000 terms).

Despite the fact that this chapter occasionally comprised separate terms of biology, geology and ecology, the number of words not found by PARS was very small.

«Databases of the 95-96ies»

This chapter was translated by PARS very well using the general and computer dictionaries. Post-editing consisted in making minor corrections to the machine translation.

Generally speaking, the draft translations provided by PARS were of different quality, depending on the source text subject area and, accordingly, on the presence of terminological dictionaries.

A few translations required no post-editing at all or «cosmetic» post-editing. For example:

151. Офицеров В.И. Исследование структурной организации конформационных антигенных детерминант на примере белков оболочки вируса гепатита А.

Machine translation:

151. Ofitserov V.I. Research of the structural organization of conformational antigenic determination on the example of the proteins of the capsule of hepatitis virus A.

On the other hand, in some cases the translation offered by PARS had to be corrected completely to obtain the correct text. The fact is that if the system dictionary doesn't have a set expression, PARS translates it word for word, sometimes making it hard to understand. For example, the phrase «принимающий решения» was translated «receiving decisions», on analogy with «receiving letters». The phrase «образ жизни» was rendered as «image of life» instead of «way of life», etc.

One of the merits of the PARS system is the translation variants option. Here is an example.

580. Носиков В.В. Поиск и изучение антигенных детерминант, связанных с аутоиммунной деструкцией островковых бета-клеток при инсулинозависимом

сахарном диабете, с сипользованием библиотек бактериофагов, экспрессирующих широкий спектр разнообразных пептидных эпитопов.

Machine translation:

580. Nosikov V.V. Search and the studies* of antigenic determinants bound with the autoimmunity disruption of islet beta-cages* at инсулинозависимом sugar diabetes, using the libraries of bacteriophages expressing the broad* spectrum of diversified* peptide epitopes.

A double click on the asterisk will display the list of translation options for this word/phrase. Having chosen one of the variants and pressed the button, the post-editor inserts it into the text instead of the initial one.

In the above text, the following translation variants were offered: studies (research, analysis), cages (cells), broad (wide, capacious, extensive, large-scale), diversified (miscellaneous, diverse). The most important substitution is cerainly 'cells' instead of 'cages'.

Transliterations are presented as translation variants for proper names, which is very useful when post-editing machine-made translations of bibliographic documents. For example, in the above text, *Nosikov V.V.* was substituted for *Носиков В.В.*

6.1. *Post-editing PARS-made technical translations*

The order from Kharkov Aviation University consisted in translating from Russian into English a large corpus of technical texts on aircraft building for Iranian students who were coming to sudy at the University. 650 K of source Russian texts were presented as WinWord and DOS files. They were translated by PARS and then post-edited. Both PARS for Windows and PARS for DOS were used, with a number of terminological dictionaries.

The translator had very little time for doing the work. An original Russian text of 20-30 pages used to be given every day at noon, the request being to «please translate it not later than tomorrow morning!».

The situation being so tense, sacrifices as to the stylistic purity of the end-translations had to be made in order to submit them as soon as possible. The translator's was to make the texts grammatically correct and understandable, omitting a number of stylistical details, such as repetition of several «of»-clauses, misuse of articles in those cases where this did not affect understanding, etc.

This allowed Olga to come up with translations that were grammatically and lexically correct though stylistically far from ideal in a number of cases. Here are two examples of machine translations left as they were, without any post-editing:

«Requirements to the execution of the outlines of modern airplanes and assurance of inter-changeability of their aggregations».

«This advantage is especially noticeable on the larger level of loading».

Again, technical terms were the main difficulty, but the problem was solved by using Polyglossum were PARS failed to translate a term: the time needed to find in

Polyglossum a word not translated by PARS and paste it into the text is, as a mathematician would say, «negligibly little». Besides, all the «new» terms were immediately entered into the corresponding PARS dictionaries by Lingvistica ‘93 dictionary officers, which made PARS «cleverer» with each text translated.

At the same time, there were also some sentences generated by PARS that had to be changed completely, as, for example, the following one:

«After switching-on pumping station, if via 5 with pressure is not is heaved above 8 Pa actuates signaling table ...ABORT».

If all or most of the sentences had been translated so poorly, post-editing would have been much more difficult, which would have made MT quite or almost useless. However, it only took the translator about **2-2.5 hours to post-edit 30 K of texts using PARS, and, what is very important, the work itself was not so boring and tiresome as manual translation**. In other words, **editing machine-made translations was 3-4 times easier than translating the same texts manually**.

Using PARS (plus Polyglossum, if necessary), the translator can prepare 20-30 pages a day and not feel exhausted.