COMMERCIAL MACHINE TRANSLATION SYSTEMS

by LINGVISTICA '93 and ETS PUBLISHERS LTD:

ENGLISH, GERMAN, RUSSIAN, UKRAINIAN

Dr. Michael S. Blekhman writes on the use of machine translation in Russia and the Ukraine

INTRODUCTION

Since 1991, when Ukraine declared state sovereignty, which manifested disintegration of the former Soviet Union, journalists in this part of the world, especially the politically engaged ones, have kept saying that Ukraine and Russia cannot exist on their own because economic ties were 'torn up', people 'isolated from each other', and 'scientific life stopped'.

Being a professional linguist, or, to put it more precisely, language engineer, I would like to describe the fruit of collaboration of two companies: the Ukrainian Lingvistica '93, and the Russian ETS Publishers. This paper is intended to give you facts, rather than trying to convince you of something. It is a kind of 'introduction to discussion'.

BACKGROUND

Lingvistica '93 Co. and ETS Publishers Ltd. have rather a solid position in the market of language engineering products in the territory of the former Soviet Union. These companies are led by M.S.Blekhman and I.V. Fagradiants, respectively.

As for me, I have the honour of being one of the numerous pupils of one of the most outstanding Russian linguists, Prof. Raimund Piotrowski. He has brought up dozens of specialists, among whom are the authors of the well-known Stylus translation system. During the twenty years of my professional activity, we have developed computer-based systems for information retrieval, abstracting, indexing, and, certainly, machine translation. I am also happy to list Prof. Victor Berzon and Dr. Boris Pevzner as my teachers. Prof. Berzon was one of the most authoritative specialists in discourse analysis in the former Soviet Union, while Dr. Pevzner was the first to formulate the idea of example-based machine translation in this country (in early 70s!).

My friend and partner, Igor Fagradiants, organized a unique publishing house in Moscow, ETS, which makes electronic and traditional paper dictionaries. Among others, Igor developed a series of Finnish-Russian dictionaries, whose high quality is appreciated by his Finnish colleagues.

This article gives a thorough description of some of our products. I call a spade a spade: I have no intention of trying to persuade you that these systems are perfect, or that we have solved all the principal language engineering problems. Nothing of the kind. These products are further away from perfection than from the zero point. I will try to give the readers objective information - and let them be free in their conclusions.

1. PARS-PARS/U-PARS/D-RUMP - a 4-language MT package

Ukraine is, in fact, a country where 4 languages are broadly used, though in different capacities. **Russian** is the native tongue of an overwhelming majority of people living in the towns and cities. **Ukrainian** is the official language, and it is being more and more widely used in this country, which makes its destiny to some extent similar to that of Hebrew in Israel. As to **English**, it is the language of international contacts, including Internet, CDs, and technical documentation. The last, but not the least, is **German**. Though not so broadly used as English, this language is gaining still wider application as the economic and scientific Ukrainian-German ties grow with every year that passes.

Since 1986, we have been developing the English-Russian-English PARS system (PARS is the abbreviation of the Russian name which means 'Translating English and Russian Papers'), and, since 1990, RUMP - 'Russian-Ukrainian-Russian machine translation'. PARS is presently marketed in Russia by ETS Publishers. The CD-ROMs comprise PARS and/or the Polyglossum system of dictionaries. These products have become the most popular translation systems in the huge Russian market: **more than 30,000 CDs** have been sold as of June, 1997.

In 1996, yet another system by Lingvistica '93 appeared on the market, PARS/U, for translating between English and Ukrainian. In June 1997, the alpha-version of the PARS/D German-Russian-German system was released to be followed by the beta-version (July) and the first commercial version (September).

These four systems are quite similar, so having mastered, for example, RUMP, one will easily master PARS, PARS/U, and PARS/D.

Each system runs in 2 variants: the Windows-version and the DOS-version, though only PARS/DOS and RUMP/DOS are commercially available.

DOS VERSIONS

One of the outstanding qualities of these systems is the user (translator)-oriented **built-in two-window editor**. It features specific functions that correspond to the most frequent text-editing operations made by professional translators:

a key-stroke transposition of neighbouring words;

a key-stroke change of register (substitution of capital letters with small ones, and vice versa);

marking polysemantic words and phrases in the target text with asterisks; the user may easily substitute a translation variant;

search for the next 'new' word, i.e. a word not found in the dictionary;

the possibility of entering 'new' words into the dictionary directly from the text editor; according to the principle, 'dictionary first', the user opens the dictionary and enters the next 'new' word into it, while the word entered is highlighted in the text so that the user can see the context and give the right translation(s).

The screen may be split either horizontally or vertically, and the user may scroll either both windows synchronously, or the active one only. The target text may also be exported to another text editor supporting ASCII files, such as PenEdit, a pen editor developed by the Kiev-based team led by Dr. Alexander I. Kazakov.

TWO WINDOWS VERSIONS

These systems work under Windows 3.1 and Windows 95, and they translate files in such formats as WinWord, HTML, as well as Windows Help-files.

Each system may be activated directly from MS Word 6.0 and MS Word 7.0 (versions for MS Office '97 are under way) Once the MT systems have been installed, the main menu of MS Word will have the item 'Translate', with the option for running the corresponding system: PARS, PARS/U, PARS/D, RUMP. The user opens the source text in the editor and starts one of the systems, after which the machine translation appears in the bottom window created by MS Word. Formatting of the source text, such as fonts, styles, and tables is preserved in the target text. Polysemantic words and phrases are marked with 'asterisks', just as in the DOS versions.

'New' words and phrases may be entered into the dictionary directly from the screen. This differs from the DOS-version inasmuch as the user marks the word/phrase to be entered, clicks the 'New word' button, and the word/phrase is written to the dictionary. Unlike the DOS-versions, the principle is 'text first', rather than 'dictionary first'. Another difference is that not only separate words, but also **phrases** may be entered into the dictionary directly from the text.

Users may also translate on-screen Help and texts of Internet WWW-pages in the HTML format. This is done via Clipboard: the text portion to be translated is copied to the Clipboard, the MT system is started, and the target text appears in a separate window under the source text.

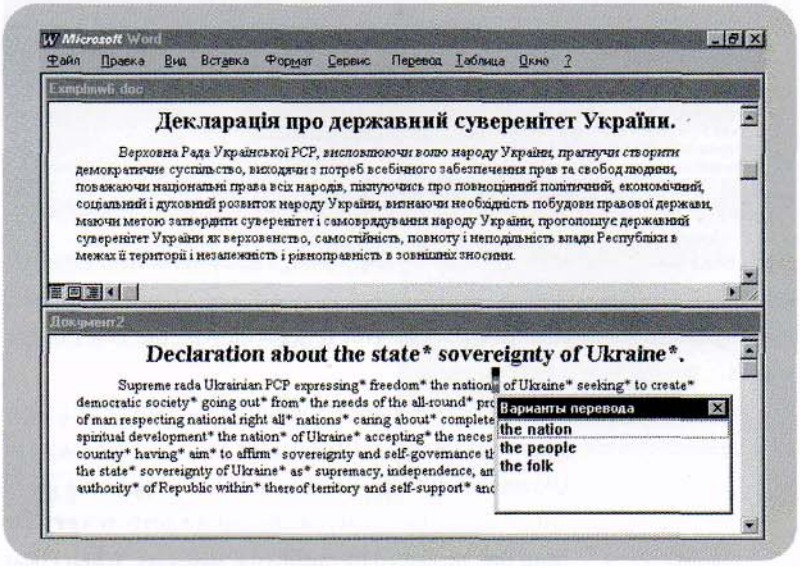

Fig.1. PARS/U has translated 'Declaration of State Sovereignty of Ukraine' from Ukrainian into English

The machine translation may be saved as a separate file.

It is important to add that Lingvistica '93 and ETS have developed a beta-version of the new PARS project: **PARS has been 'imbedded' directly into Netscape Navigator**.

To finish the functional description, I would like to add that all the above systems, both in DOS and Windows versions, run in stand-alone and **network** modes.

2. Translation: general principles and problems

As experience shows, it is rather hard to draw a demarcation line between the 'classical' translation strategies: direct and transfer-based. PARS, PARS/U, PARS/D, and RUMP have dozens of transfer rules, though it's hard to call them purely transfer-based systems.

I prefer a different kind of terminology, distinguishing between the following two translation principles:

FAT - 'First Analyse - Then Translate', and

FTA - 'First Translate - Then Analyse'.

Our products are FTA-type systems, just like quite a number of very well known PC-MT systems,

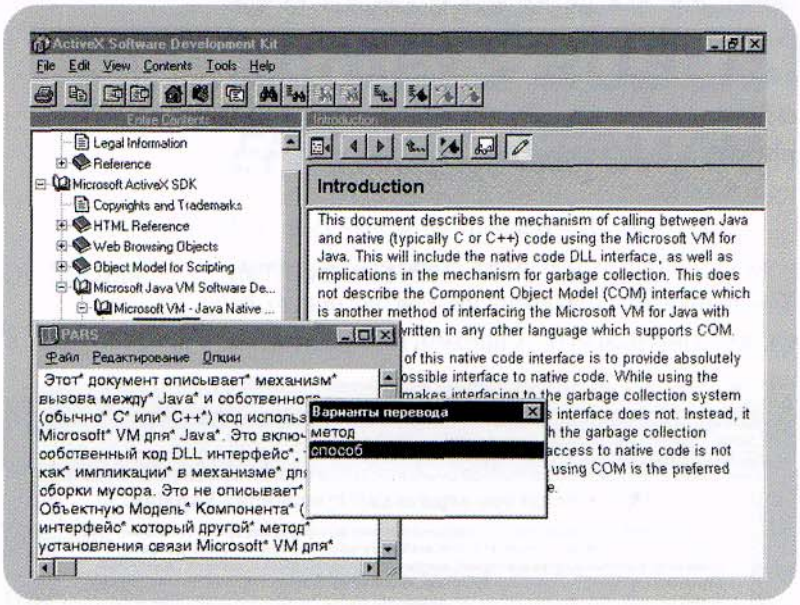

Fig. 2. PARS has translated a Help file from English into Russian

such as those by Globalink. The system first translates the source text 'word by word' and 'phrase by phrase', and then tries to edit it according to the rules of the target language.

Let's be honest: if the source and target languages are not as closely related as, for example, Russian and Ukrainian (although Russian and Ukrainian have a lot of differences), the quality of the output texts is **very** far from that produced by qualified translators. When I hear or read that an MT system ensures '80-90-percent accuracy', I am inclined to consider such a statement a mere advertising trick. Yes, machine grammars are being constantly improved, but, being a professional language engineer, I can hardly imagine that computer programs will ever be able to compete with qualified humans.

SO, WHAT CAN THE SYSTEMS BY LINGVISTICA '93 DO?

RUMP does translate texts in such a way that they are 70-80%, sometimes even 90% ready for publication, the quality of Russian-Ukrainian translation being somewhat higher than that of Ukrainian-Russian. As to PARS and PARS/U, they are, and PARS/D will be, used to let the user have a general idea of the document, for example, when browsing large databases, i.e. 'scan' the text; create a draft for subsequent polishing, i.e. turning the draft into **translation**.

The option of selecting translation variants essentially simplifies editing of the machine translation. This option also provides **transliteration of proper names**: for example, the Russian name 'Ivanov' is not translated into English by PARS, but its transliterated form is suggested as a translation variant.

At the same time, numerous users, among which there are professional translators, say that **it is very hard to edit machine translations in MS Word**. I will explain this statement.

The main disadvantage of the FTA-type programs translating between languages, one of which belongs to the German group and the other to the Slavonic one, is that, more often than not, the word order is not observed. The translator has to change it according to the rules of the target language. The reason is that observing word order requires very serious transfer rules, based not only on grammatical, but also on **semantic** characteristics of the words. Using semantics in machine translation is a task for a new generation of commercial MT systems. Speaking of editing machine translations in MS Word, the only option that can be used for transposing words is tiresome work with text blocks. That is why a Windows-version of PenEdit is being developed. It allows the user to transpose words very easily, using an electronic pen or a digitizer.

MT systems by Lingvistica '93 may use up to 4 dictionaries in the translation session, and the user may set their priorities. When translating, the system looks the word (phrase) up in the dictionary that has the highest priority, then, if it was not found there, in the following one, etc. As it turns out, this approach has not only advantages, but also drawbacks. Let's discuss the latter.

a) To begin with, PARS comprises quite a number of dictionaries, which requires linking up more than 4 dictionaries in some translation sessions. For example, the following dictionaries may be used for translating aviation texts:

general;
aviation;
aerospace;
mathematical (mathematical modeling in aircraft building);
computer;
aviation medicine;
radio electronics;
ground and space communications;
polytechnical.

b) Having found a word in one of the dictionaries, the system stops looking it up in the rest, which may cause incorrect translation simply because one and the same word may be present in different dictionaries and thus have different meanings.

www.logos.it/language_today

c) Another large problem consists in the difficulty of correctly assigning priorities to the dictionaries. For example, PARS translated an English medical text into Russian using the medical and general dictionaries in the indicated order of priorities. The word 'flow' was translated as 'menstruatsiya' (menstruation) instead of 'potok' (flow), the latter being suggested as a translation variant; but if the general dictionary had a higher priority, the translation would be correct, and the wrong variant would be given as a variant.

3. Dictionaries

It seems to me that one of the most important criteria of evaluating a commercial MT system is its dictionary support subsystem: the easier it is to extend dictionaries supplied with the system as well as create user's dictionaries, the better the system is in general.

3.1. User options

1) Dictionaries in Lingvistica '93 systems are fully bi-directional. For example, if the user enters an English word with its Ukrainian translation into a PARS/U dictionary, the system automatically sets the opposite correspondence, Ukrainian-English.

2) Any dictionary may be browsed and edited in either language direction. For example, English-Russian or Russian-English.

3) It is very important that a word/phrase can have a practically limitless number of translations. This permits a wider range of translation variants, from which to choose in the target text.

4) Dictionary entries in the systems by Lingvistica '93 are similar to those in traditional dictionaries. The main difference is that in 'paper' dictionaries it is the head word which is replaced with a tilde in a phrase, this word bearing the main sense of the word string, while in PARS, PARS/U, PARS/D, and RUMP dictionaries the first word is considered the headword.

NB: Since Russian and Ukrainian are inflective languages, word endings are separated from the stems with vertical lines.

The user may use the one keystroke transposition option in the dictionary entry to assign a higher priority to the translation that is considered the most likely one for the subject area. For example, in the PARS general dictionary, the Russian word 'obshchestvo' has two English translations - 'society' and 'company'. For translating sociopolitical texts, it is advisable to put the translation 'society' in the first place in the dictionary entry, and then the word 'company' will be placed under an asterisk as a translation variant. For translating financial-legal texts, the word order is the reverse.

5) These systems feature **fully-automated indexing** (tagging) of Slavonic words being entered into the dictionary. The system automatically assigns grammatical characteristics, such as part of speech, declension, conjugation, and subclass characteristics, such as gender. If the program is unsure of how to index a word, the user can choose from several options. For example, the system is not sure about the declension of the Russian word 'benzozapravshchik' - as 'avtomat' (i.e. subject) or as 'inzhener' (i.e. human).

3.2. Lingvistica '93 dictionaries

Most of the dictionaries used have been compiled by professional lexicographers. The lists of dictionaries supplied with the systems include not only the quantitative characteristics, but also **the names of authors**.

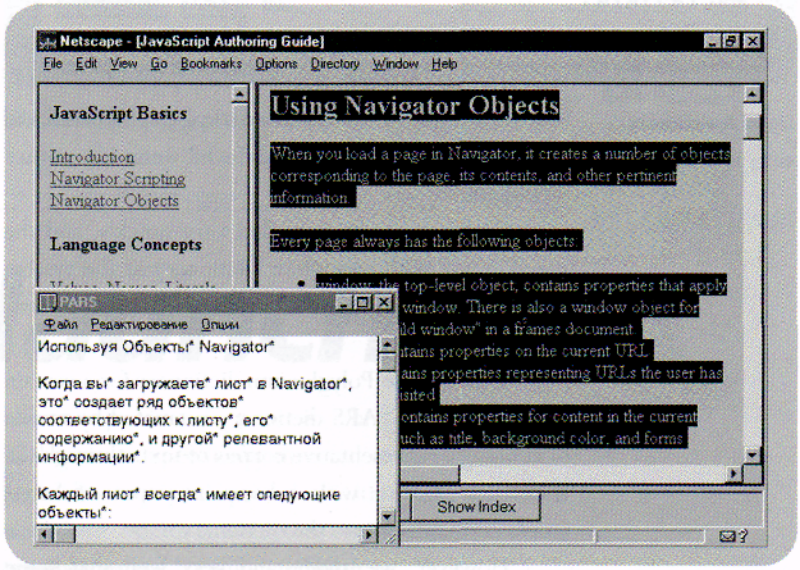

Fig. 3. PARS has translated an HTML file from English into Russian

PARS features a large spectrum of English-Russian-English specialist dictionaries, the subject areas being technology, business, medicine, space engineering, electronics, mathematics, chemistry, automobile building, etc. The total number of terms as of June 1997 is above 800,000 words and phrases in each language direction - English-Russian and Russian-English.

Such great volumes could never be compiled without the collaboration of Lingvistica '93 and ETS. Under the joint PARS+Polyglossum project, the dictionaries of the world's largest English-Russian dictionary basis, Polyglossum, are semi-automatically converted into the PARS format. The procedure of semi-automatic processing consists of three stages.

First, the Polyglossum dictionary is imported into PARS.

Then, the Russian words of the new dictionary are encoded in a batch mode according to the **coincidence principle**: the word acquires the same grammatical characteristics as in the PARS dictionary that was set as the prototype.

At the last stage, the dictionary officer looks through the dictionary entries and encodes the words that were not encoded by the batch mode program. In this case, the program uses the **analogy principle**:

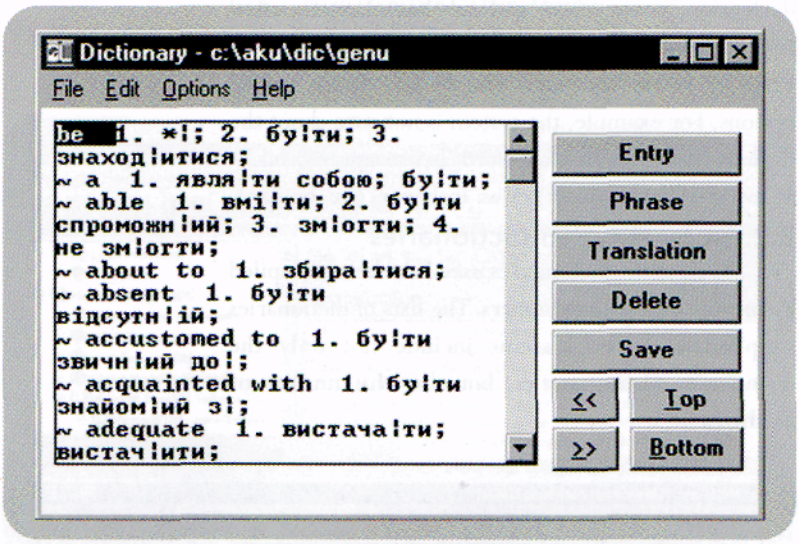

Fig. 4. A dictionary entry in PARS/U

the word acquires those grammatical characteristics as similar words that were entered into the other dictionaries.

Dictionaries are compiled very quickly, and the speed grows with every new dictionary as the system has more and more encoded words to compare the new ones with.

If there is no Polyglossum dictionary for a certain subject area, the PARS dictionary is created by means of running a representative corpus of texts through the translation system with subsequent input of 'new' words and phrases into the dictionary.

However, we understand very well that some dictionaries are to be updated much quicker in order to include the most up-to-date terminology. This relates, for example, to such a 'terminologically flexible' sphere as telecommunications. The only way to do this is to cooperate with the companies that generate new terminology, and **I would like to suggest such collaboration to all those interested in translating technical documentation between Russian, Ukrainian, English, and German.**

4. Translation technology: the PARS+Polyglossum tandem

Experience shows that the most efficient way of translating from Russian into English and from English into Russian in our case is by using PARS and Polyglossum to complement each other. The fact is that the Polyglossum system has a flexible program for dictionary look-up, and the word entries in its dictionaries contain numerous explanations and commentaries. That is why Polyglossum is not only a source of new PARS dictionaries, but it also serves for translating technical terms which PARS fails t translate, or for choosing a more appropriat translation variant if the human translator post-editir the raw machine translation needs an explanation of term.

The 'freshest' example of the realisation of th technology (February - March, 1997) was th translation from Russian into English of a collection lectures on various branches of the aviation industr The work was done by Lingvistica '93 for Kharkc State Aviation University and the Kharkov Aviatic Plant. The total volume of texts to be translated w several hundred pages. Translation was carried ou using both PARS and Polyglossum. The dictionari used and the number of terms in each are indicate below:

a) PARS:
general (40,000);
polytechnical (76,000);
concise aviation dictionary (7,000);
aerospace (60,000);
mathematical (80,000);
computers (20,000);

b) Polyglossum: polytechnical dictionary - 300,000.

The source texts were received as DOS an WinWord files and both the DOS and Windov versions of PARS were used for translation. The sourc texts were first translated by PARS. After the machir translation of each document, 'new' terms (practical all of them were then found in the Polyglossu polytechnical dictionary), were entered into th corresponding PARS dictionaries, which substantial improved the quality of the translation of subsequer documents.

Machine translation underwent human post-editing the purpose of which was to create **informativ though maybe stylistically imperfect,** Englis text. Editing was made in two stages:

primary editing made by **two translators** who knov English grammar rather well but do not specialize in translating texts on aviation;

final editing - verification of terminology by an experienced translator of aviation texts.

In the context of this work, we tried to determir the efficiency of using the two systems, PARS an Polyglossum, at the stage of primary editing. The questio was: 'Is it easier, and, if so, to what extent is it easier t edit the machine translation output than to translate th text manually?' The translators gave the following answe **it is 3-4 times easier**. Using PARS and Polyglossur each translator prepared 20-30 pages a day.

The reader may refer to the Summer issue of M News International: the paper by Olga Bezhanova give

www.logos.it/language_today

a detailed account of the translation process.

Another goal of this work was to improve the translation algorithm and determine the most frequent operations made by the translator when editing machine translations from Russian into English. This will permit A.Kazakov's group to fine-tune PenEdit.

5. How are the systems supplied?

Lingvistica '93 supplies the MT systems both on the 'buy-and-go' and 'registered user' principles. In the latter case, the customer pays 350 Ukrainian grivnas (about 180USD) on average for a single licence, or 500 grivnas (260USD) for a networked variant, after which, as a registered user, he/she gets an upgrade free of charge every 4-6 months within 2 years.

Igor Fagradiants suggested and implemented the 'buy-and-go' idea. ETS sells PARS and Polyglossum in Russia, and RUMP in Ukraine, on CD-ROMs, at very low prices that correspond to the low wage levels in the former Soviet Union. The average prices are 15 USD (!) for a disk.

We are providing more information on our efforts to those who attend the MT Summit in San Diego, California from October 29 to November 1 this year. We hope that we will be able to answer all the questions raised there! [*Editor: there will be a report on the MT Summit in the next issue of Language Today*] ■

Dr. Michael S.Blekhman
Director, Lingvistica '93 Company,
Kharkov, Ukraine;
Head of The Laboratory for Machine Translation,
Kharkov State Polytechnical University;
Vice-President, POLYGLOSSUM, Inc.,
Washington, DC, USA.

94a Prospekt Gagarina, apt.111, Kharkov 310140, Ukraine.
Tel.: (0572) 27-71-35. E-mail:
blekhman@lotus.kpi.kharkov.ua

blekhman
@lotus.kpi.kharkov.ua