System Demonstration

PARS AND PARS/U MACHINE TRANSLATION SYSTEMS: ENGLISH-RUSSIAN-ENGLISH AND ENGLISH-UKRAINIAN-ENGLISH

Michael S. Blekhman

Lingvistica '93 Company
94a Prospekt Gagarina, ap. 111
Kharkov 310140
Ukraine
Tel.: (0572) 27-71-35; (0572) 40-00-36
Fax:(0572)40-06-01.
E-mail: blekhman@lotus.kpi.kharkov.ua

1. General

Lingvistica '93 Co., a private firm based in Kharkov, Ukraine, a partner of Polyglossum, Inc., USA, and ETS Ltd., Moscow, Russia, presents the following commercially sold bi-directional machine translation systems for IBM PCs:

- PARS 3.1 — a Russian-English-Russian system;
- PARS/U 1.1 — a Ukrainian-English-Ukrainian system.

Both systems run under MS Windows; PARS also runs under MS DOS. Both PARS and PARS/U run in stand-alone and multi-user environments.

Up to 4 dictionaries can be used in the translation session.

The systems are compatible with all text processors supported by Windows: the user may copy the text portion to be translated to Clipboard, have it translated by PARS or PARS/U in the "Clipboard" mode, and the target text will be written to the Clipboard automatically.

The systems are also "embedded" in MS Word 6 and MS Word 7: if a text is opened in WinWord, the "Translate" option in the editor main menu lets the user have the text translated, after which the target text is written to another window, placed under the first.

The source text format is fully preserved in the target file, including fonts, styles, paragraphs, and tables.

Each polysemantic word is marked with an asterisk: if the user makes a double click on the asterisk, a panel of synonyms will be displayed, which lets the user have a word substituted in the target text.

2. Dictionary updating

PARS and PARS/U are designed in such a way that the end-user may customize the system easily to his/her subject area. Customizing is ensured in either of the two ways:

a) by updating system dictionaries, and/or

b) creating any number of new dictionaries.

PARS and PARS/U feature the dictionary extending option which is unique for operational MT systems:

- a word or phrase can be input into the dictionary immediately from the WinWord windows;
- the dictionaries are invertible, so if, for example, a Ukrainian word is entered with its English translation, the system sets the English-Ukrainian correspondence; invertibility may be disabled if the user considers it necessary;
- the Auto-encoding option assigns grammatical characteristics to the Slavonic words entered into the dictionary: part of speech, gender, tense, declination, conjugation types.

PARS and PARS/U use large grammatical dictionaries of the Russian and English languages, correspondingly, as well as the so-called grammatical indexes; both are used for automatic encoding words on analogy with previously entered ones.

Any number of translation equivalents may be assigned to a word/phrase in the dictionary; the translations may be transposed in the word-entry so that the most relevant one occupied the first position: PARS(/U) will include the first translation variant into the target text, while the others will be displayed as synonyms on the user's demand.

PARS dictionaries feature, among others, the so-called "distant" phrases such as "make...decision", the presence of which simplifies source text analysis.

Two dictionaries can be merged using the "Import" option. This is used for compiling a dictionary on several workstations.

PARS features a large English-Russian-English lexical database; the dictionaries included over 500,000 word-entries in each part as of July, 1996. The topics covered are business, computers, machine building, aerospace technology, aviation, medicine, business, mining, electronics, ecology, etc.

PARS/U had 80,000 word-entries in each part as of July, 1996.

Two ways of creating new dictionaries are used:

a) running a representative corpus of English, Ukrainian, and/or Russian texts and entering new words and phrases into the dictionaries. A lexical unit is considered "new" if it is not present in the dictionary, or if its translations don't correspond to the user's subject domain. Besides, the dictionary officer transposes translations of polysemantic words if necessary. He/she also can create any number of new dictionaries;

b) using traditional (printed) dictionaries as the basis for the PARS(/U) dictionaries.

Lingvistica '93 is engaged in large linguistic projects of creating lexicographic databases for large subject areas. One of such projects is PARS/Avia, carried out to meet MT requirements of several Aviation giants within Ukraine and abroad, such as The Antonov Aviation Design Bureau, Kiev, Ukraine, or Boeing in USA. With this end in view, three kinds of MT dictionaries are compiled:

a) "nucleus zone" dictionaries, which comprise kernel terminology: construction of aviation vehicles, specific materials, etc.;

b) "adjacent zone" terminology, which comprises terms of the subject areas such as rocket engineering or aerospace medicine;

c) "peripheral zone" dictionaries: polytechnical, radioelectronics, etc.

PARS is compatible with the Polyglossum dictionary support system developed by the ETS company, Moscow, Russia. All the PARS dictionaries are converted to Polyglossum format, so the end user may choose between fully automatic translation and machine-aided translation, which raises product flexibility. These systems are supplied on diskettes and on CD-ROM. The latter version is distributed by Polyglossum, Inc. in the USA.

Quite a number of Ukrainian-English bi-directional dictionaries have been and are being created for PARS/U. This problem is not easy because Ukrainian computer lexicography is in the first stage of development so far. Lingvistica '93 has made agreements with leading Ukrainian lexicographers to use their dictionaries in PARS/U. This collaboration let us compile MT dictionaries for the following subject areas: computers, ecology, telecommunications, measurements. Besides, a middle-size general dictionary of 35,000 word-entries in each part has been created.

3. PARS/DOS

PARS/DOS is a user-friendly tool for translating ASCII files between English and Russian. It features a built-in 2-windows text editor; the screen can be split either vertically or horizontally. The user may:

- transpose words and change case by one keystroke;
- enter new words into the dictionary directly from the screen;
- mark "standard" and vertical blocks;
- have a block or the whole text translated;

• have a polysemantic word substituted with a synonym.

PARS/DOS also features a built-in Cyrillic driver that lets the user type and display Russian texts. The driver automatically switches over between Latin and Cyrillic depending on the translation direction.

4. Translation quality

The translation quality attained by PARS and PARS/U may be characterized as "rough but not dirty". In other words, the systems come out with information quality translations, while Russian/Ukrainian to English translations have a higher quality than English to Russian/Ukrainian.

PARS is based on the FTA translation philosophy: "first translate, then analyze". This consists in grammatical analysis of the word-for-word translation text and subsequent synthesis of the target text. Semantic characteristics, such as "time", "nationality", etc., are also used. PARSes use no tree-structure representation of the source sentences for two reasons:

a) this would require a much more complicated word-entry structure as semantic characteristics are necessary to provide such a representation; in this case, dictionary extending would be a serious problem for the end user;
b) as numerous experiments have shown, traditional transfer rules act like the little girl in the well-known nursery rhyme: when they are good, they are very, very good, but when they are bad, they are horrid: transformations of the sentence structure may cause misunderstanding and non-understanding of the target sentence generated; this is especially true in real-life conditions, when unrestricted texts are processed using large, not toy-like dictionaries containing dozens of thousands of word-entries.

"Local transfer" is used instead, which consists in transposing words, inserting articles and auxiliary words, etc. This has proven to be an acceptable compromise between output quality and dictionary extending simplicity.

Translation speed is about 1-1,5 sec. per page (250 words) on a machine like Pentium/100.