Users' Experiences with a Set of Domain-Specific Dictionaries for the Stylus Machine Translation System

Svetlana Sokolova

Abstract

STYLUS is currently the most widely used machine translation software for the Russian language. There are more than 15,000 registered users of the STYLUS system to date. One of the most important features of STYLUS is that it allows the user to customise the system through the management and editing of the dictionaries. The linguistic database of the STYLUS system contains three different dictionary categories: general-purpose dictionaries, domain-specific dictionaries, and user-defined dictionaries. In the current version of STYLUS, the English-Russian and Russian-English systems contain a set of domain-specific dictionaries that are arranged into groups. This grouping of several different dictionaries gives a better overview of all dictionaries. Each of the domain-specific dictionaries can also be used individually.

The Technical group provides a good illustration. This group is made up of eleven domain-specific dictionaries entitled 'Telecommunications', 'Software', 'Automotive', 'Mining', 'Building and Construction', 'Military', 'Oil & Gas', 'Electrotechnical', 'Home Appliances' and 'Navy'. This article discusses the data we obtained from a questionnaire sent to registered users.

Dr Svetlana Sokolova, Ph.D.

Graduated from Leningrad State University in mathematics. Obtained a Ph.D. in computer science. Has been involved in several machine translation projects for state organisations as head of a software design group. In 1991, established the company "PROject MT" Ltd. of which she is currently the president.

AO PROject MT Ltd.

The PROject MT team consists of mathematicians, programmers and linguists. The company staff now includes 52 employees. The leading specialists of the company have extensive experience in designing machine translation systems. Several versions of STYLUS have been launched on the market since the company was set up five years ago. The first version, called PROMT- PROgrammers Machine Translation, could translate software documentation from English into Russian. The current version of STYLUS can translate from French into Russian, and from Russian into English, German and French. STYLUS runs on Macintoshes and on PCs under DOS, Windows and Windows 95.

Dr Svetlana Sokolova
AO PROject MT Ltd
PO Box 632
St Petersburg, 199053 Russia

Tel: +7 812 275 78 87, Fax:+7 812 2757893
E-mail: svetlana@promt.spb.su

What is STYLUS?

By way of introduction I should like to explain what STYLUS is and how it works. STYLUS is a machine translation system that can translate from Russian, and into Russian, for a number of European languages (English, German and French). It is a commercial product designed for PCs and runs under DOS, Windows, and Windows 95. There is also a Macintosh version. STYLUS is currently the most widely used software in machine translation for the Russian language. We have more than 15,000 registered users of the program. They are individuals, small and big companies, universities, and state and government organisations. Our users include the Administration of the President of Russia, the Russian Space Agency, NASA, the Central Intelligence Agency, AT&T, the BBC World Service, Inmarsat, Siemens, Lockheed Corp., Chevron, NPO ”ENERGIA”, the US Air Force, the FBI, Volvo, and Ernst & Young.

The popularity of STYLUS can be ascribed to its user-friendly linguistic interface designed for end users, the high speed of the translation process and, indubitably, the intelligibility of the output. One of the most important features of STYLUS is the fact that the system can be customised to users’ own needs by managing and updating the dictionaries.

Dictionary structure

The linguistic database of STYLUS contains three different types of dictionary: general-purpose dictionaries, domain-specific dictionaries and dictionaries created by users themselves.

As a rule, a general-purpose dictionary in one direction (English-Russian, for example) contains about 60,000 entries and is consulted as the system’s basic dictionary during the translation process. It contains entries for the most frequent words and phrases of the source language. These entries have a functionally sophisticated collection of semantic and syntactic tags that are used for the translation algorithms.

The domain-specific dictionary contains not only terms specific to the corresponding domain but also general-purpose words, if they have some specific function in that domain.

For the correct translation of software documentation, for example, words such as “program” or “application” need specific information. The volume of a domain-specific dictionary varies from 10,000 to 40,000 stems.

A user dictionary is created by the user himself. There is no limitation in the system on the number of user dictionaries called up during the translation process. Normally, if a user translates texts in different domains, he organises his own terminology into different user dictionaries, and then selects the appropriate user dictionaries along with the general-purpose dictionary, and, where necessary, domain-specific dictionaries.

A user can add new words or phrases and change the meanings of existing entries in both the general-purpose and the domain-specific dictionaries. However, all these changes will be saved in the user dictionaries, because neither the general-purpose nor the domain-specific dictionaries can be accessed by users. They can contain

information which is hidden from the user, and to prevent basic linguistic algorithms from employing this hidden information, the system saves all user-customised information in the user dictionaries.

Dictionary Manager and Entry Editor

The various dictionaries are handled by a very flexible feature in the STYLUS interface - the Dictionary Manager. This makes it possible to connect or disconnect the domain-specific and user dictionaries when translating a text and to change priorities for dictionary interrogation. New user dictionaries can be created and user or domain-specific dictionaries can be opened for consultation (see Fig. 2).

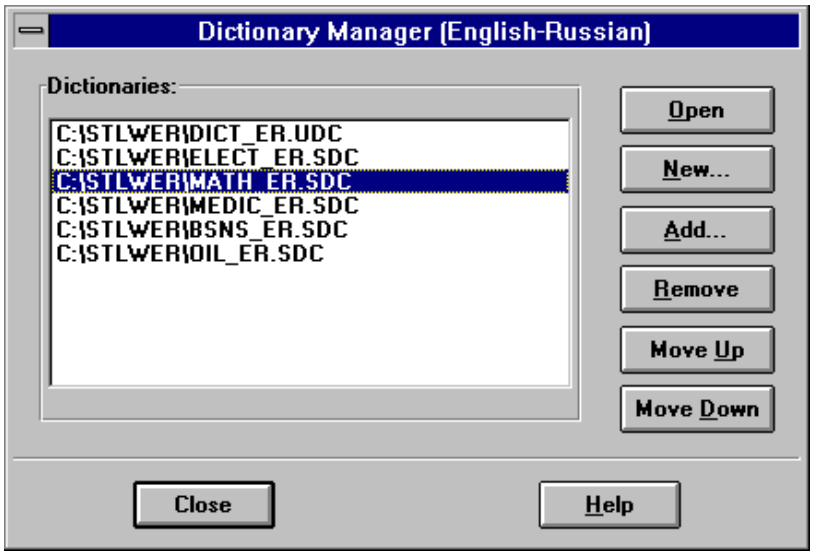

Fig.1 Dictionary Manager

When a dictionary is open, the user can copy an entry into his own dictionary by the "drag and drop" method, change the translations and grammar tags of the entry at his own discretion, and enter new words and phrases. It is possible to work simultaneously with several dictionaries.

These features offer a unique opportunity to customise the system to an individual topic and to the specific text being processed.

Another important item of the dictionary interface is the Entry Editor.

Each STYLUS dictionary is a bilingual dictionary, and all the dictionary entries have an identical structure. An entry includes the word stem, the grammatical description of this stem, and its equivalent translation in the target language. A machine translation dictionary contains highly specific information not included in a conventional dictionary. It is not easy for users to manipulate these dictionaries without a thorough knowledge of the linguistic methods employed. Hence, the STYLUS system includes the Entry Editor, which provides user-friendly access to the entry. This Editor is a kind of expert system that submits linguistic information to

a questionnaire, automatically forms a declension, assigns a set of patterns to input entries, etc. (see Fig. 2 and Fig. 3).

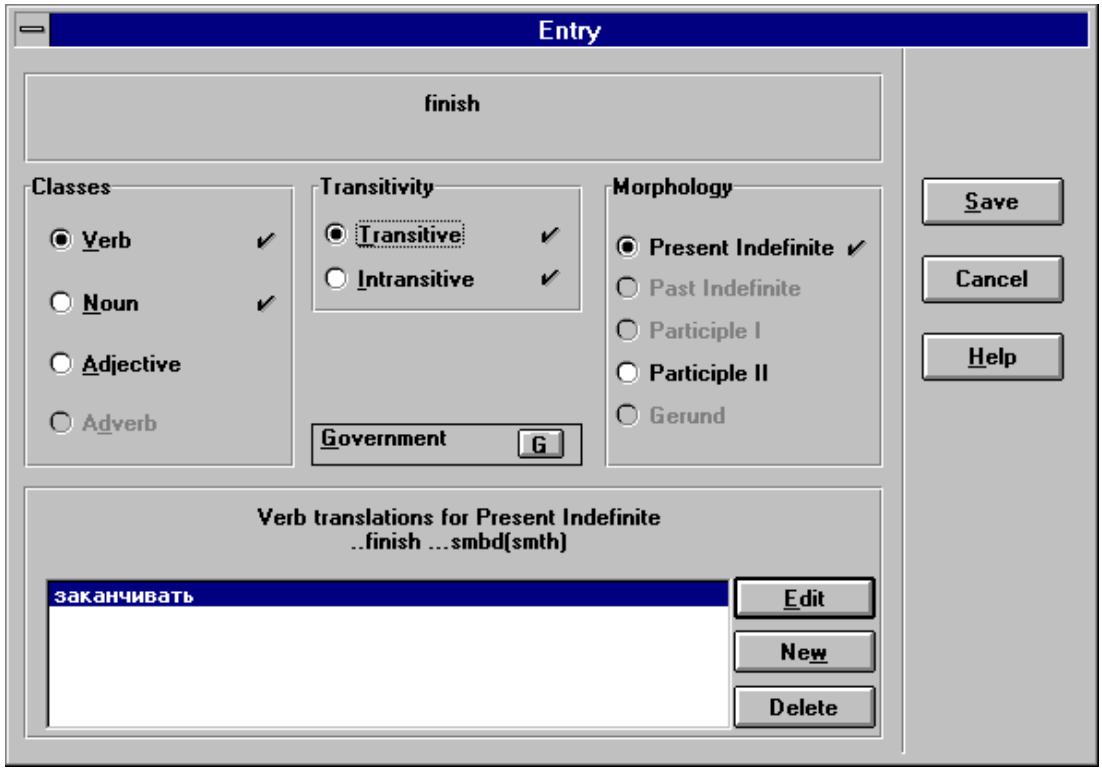

Fig.2 Dictionary Entry Questionnaire

The Entry Editor has two operating modes: Beginner and Expert. If the user chooses Beginner mode, his interaction with the system will be minimal. In Expert mode he can actively intervene in the updating process. For example, he can introduce “Government” by himself, change semantic information or correct automatically produced word forms. In this case the user should be familiar with both the source and the target language grammar. Automatic declension is a very important feature of the system, because STYLUS employs full morphology description for all the languages processed: 800 morphological types for Russian, 230 morphological types for English, more than 300 morphological types for German and French. Thanks to this automatic declension, the stems to be entered can be defined very quickly, cutting down the among of routine work involved in dictionary adaptation.

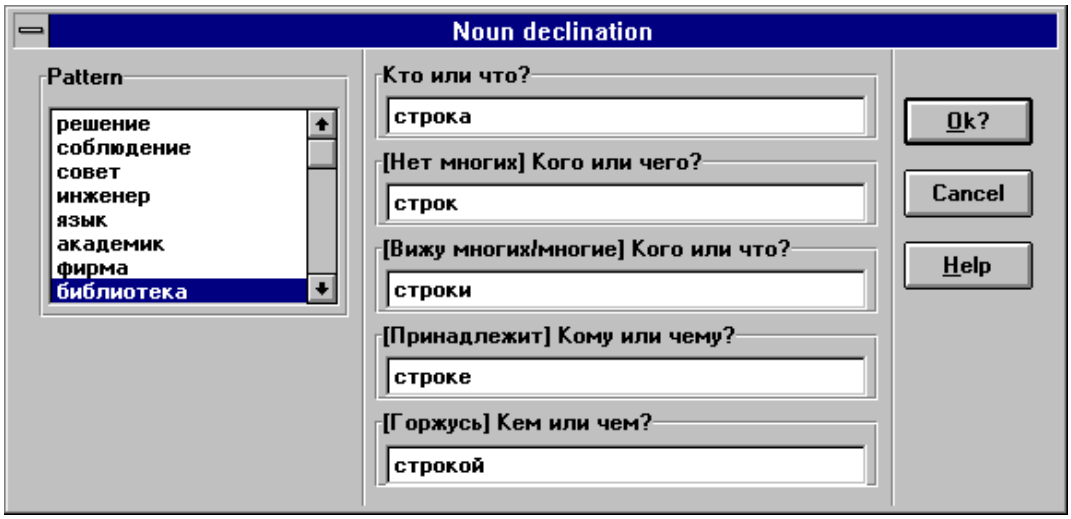

Fig.3 Automatically generated declension

Sets of Domain-specific Dictionaries

The current version of the STYLUS English-Russian and Russian-English systems has a large number of domain-specific dictionaries grouped together in sets. This grouping of several dictionaries in a set facilitates dictionary management. Obviously, each of the domain-specific dictionaries can also be used separately.

The dictionary sets cover Business, Polytechnic and Science. The Business dictionary set includes three domain-specific dictionaries: Business, Banking and Finance; Law; and Software.

The Science dictionary set includes five domain-specific dictionaries: Mathematics; Physics; Electrotechnics; Chemistry; and Computer Science.

The Polytechnic set is the most representative, comprising eight domain-specific dictionaries: Telecommunications; Software; Automotive; Mining; Building and Construction; Military; Oil & Gas Industry; and Electrotechnics.

There are also a number of individual dictionaries which are not included in any collection because of their very specific application: Military; Navy; Space Industry; Home appliances; and Medicine.

For the German-Russian and Russian-German systems there are only two domain-specific dictionaries: Business documentation and Software documentation. Dictionaries for Law and Medicine are under development.

For the French-Russian and Russian-French systems domain-specific dictionaries are under development.

How to use the set

A novice user will often start off by attaching as many domain-specific dictionaries as he can connect simultaneously. However, meaning is often lost by doing this, because the system interrogates the dictionaries according to the priorities that are fixed in the dictionary manager.

Take, for example, a text concerning a business agreement involving a software product for the management of a chemical process. Let us suppose that three dictionaries are connected during the translation session: the business dictionary (priority 1), the software dictionary (priority 2), and the chemical dictionary (priority 3). The general-purpose dictionary is employed in the background and has the lowest priority.

The program will always translate the word “file” as a business term, the word “exception” as a software term and the word “facility” as a chemical term. Hence, it is important to select priorities very carefully when connecting dictionaries to the translation process.

Sometimes, the only way to solve a dictionary conflict is to create a user dictionary and save the appropriate meanings for the conflicting words in that user dictionary, which is then given the highest priority.

User questionnaire

We have received very interesting feedback from our registered users on their experiences with the dictionary set in their translation work. This information is extracted from the questionnaire we sent to users. About 800 questionnaires were returned, which we consider to be quite a good result. The percentage figures below are in relation to the number of responses received.

According to our users’ experiences, the best way to obtain a good translation is to use one of the domain-specific dictionaries combined with several user dictionaries.

Most users customise the system by creating their own user dictionaries:

- 29% of users create less than 5 entries a month
- 16% of users create 5-10 entries a month
- 30% of users create 10-20 entries a month
- 15% of users create 20-50 entries a month
- 9% of users create 50-100 entries a month
- 6% of users create more than 100 entries a month.

More than 52% of users consider the dictionary editing procedure to be simple enough and easy to use. About 17% of users find this work difficult, and the remainder did not answer this question.

More than 86% use the machine output as the basis for a final translation, 46% as an aid to understanding and classifying texts, and 6% use the system for other purposes, such as checking their texts in a foreign language by a process of reverse translation.

In response to the question “What type of text do you usually translate?”, the users answered as follows:

- about 70% translate technical documentation
- about 42% translate contracts and agreements
- about 65% translate business letters
- about 32% translate help information for software products

- about 44% translate scientific articles
- about 20% translate other types of documents.

The following question produced some interesting data:

In what way do you input your texts into the computer?

E-mail	16.5% of the users
Optical Character Recognition software	75% of the users
Typing	50% of the users
Files from diskettes	65% of the users.

The other questions were related to marketing issues, which are not considered in this paper.

Summary

The STYLUS translation system first appeared on the Russian market four years ago. There are now many users of STYLUS in a number of different countries. The experiences of users show that STYLUS suits both professional translators and those with a less specialized linguistic background. When separate domain-specific dictionaries are used in combination with user dictionaries to obtain a draft translation the results are very positive, making translation work more creative and productive.