# Designing a client-server architecture for the MT system POLENG

## Konstruowanie architektury klient-serwer dla systemu tłumaczenia automatycznego POLENG

Krzysztof Jassem, Mikołaj Wypych

Adam Mickiewicz University
student at the Faculty of Mathematics and Computer Science
Matejki 48/49, Poznań, Poland
wypych@amu.edu.pl
jassem@amu.edu.pl

ABSTRACT

This paper describes the structure and the modules of a new version of the POLENG Machine Translation system. The structure of this version has been designed to support many new services (including translation via the Internet). POLENG will be developed in two forms: distributed and local. The distributed form will make it possible to translate texts via e-mail and by means of a www viewer. The local version will support translation in a dedicated editor as well as in a popular word editor (e.g. Word for Windows). Among the most noticeable new features are: procedures for the treatment of unknown words and translation of HTML/XML documents.

The POLENG project has its own homepage accessible at www.poleng.pl.

STRESZCZENIE

W artykule zaprezentowano strukturę oraz moduły nowej wersji systemu tłumaczenia automatycznego POLENG. „Nowy” POLENG posiada strukturę zaprojektowaną z myślą o wielu rodzajach usług (w tym tłumaczenia poprzez Internet) System rozwijany będzie w dwóch wersjach: rozproszonej i lokalnej. Wersja rozproszona umożliwi tłumaczenie poprzez e-mail oraz przy pomocy przeglądarki WWW. W wersji lokalnej możliwe będzie tłumaczenie w edytorze dedykowanym oraz popularnym edytorze tekstów (np. Word for Windows). Nowymi cechami systemu będą również obsługa wyrazów spoza słownika oraz tłumaczenie dokumentów sformatowanych przy pomocy języka HTML/XML.

Projekt POLENG posiada własny serwis WWW pod adresem www.poleng.pl.

## 1. Preface

In comparison with the state of the art reported in [1], progress has been achieved in various aspects of the POLENG MT system. There are linguistic improvements in transfer rules (cf. [2]) and lexical resources (cf. [3]). New linguistic modules have been developed, (e.g. the module for treatment of unknown words, cf. [4], [5], procedures for character codepage autodetection), or are currently developed (e.g. the module for sentence delimitation and procedures for handling markup languages: HTML, XLS). Besides, the architecture of the system has been completely redesigned. For this reason, the name of POLENG2 has been given to the new version described here.

At the present stage of development of the system, it has become clear that there is need to make it possible for potential users to test the system. It has been assumed that the system should offer the possibility of using it via the Internet as well as a stand-alone application that might be distributed as a CD-ROM package. The client-server architecture of POLENG2 should satisfy both requirements.

## 2. General premises

The necessity of developing two versions of the POLENG2 system (a remote access version and a local version) is a result of existence of two different groups of MT users. One group consists of users that do a lot of translation and possibly want to keep their documents private. Such users would need an autonomic system that can be fully installed on a local workstation. The other group – users who need to translate documents occasionally – would probably prefer to access the translation services via the Internet. The latter group could benefit from the constantly updated linguistic resources of the system. Moreover, the Internet contact between the users and the system is of great importance to the developers of the system because the real users' texts intended for translation form a corpus of texts that is useful for further research.

The Internet version of POLENG2 will have the client/server architecture (probably powered by CORBA[1] or DCOM[2] architecture). The server application is designed to work effectively on multiprocessor computers so as to take advantage of parallel processing. The system platforms are Windows NT/2000 for the server applications and Windows95-compliant system for client applications. System clients that work in other environments will be developed in future. For the time being users of systems different from Windows are limited to the possibility of translation by e-mail or a WWW form.

[1] Common Object Request Broker Architecture

[2] Distributed Component Object Model

## 3. The structure of POLENG2

Figure 1. shows the structure of POLENG2. Arrows represent directions of data flow in the system. The diagram presents the Internet (distributed) version of the system. The structure of the local version (which in fact is a special case of the distributed version) is not discussed here. Next paragraphs describe each module in their functional and conceptual aspects.

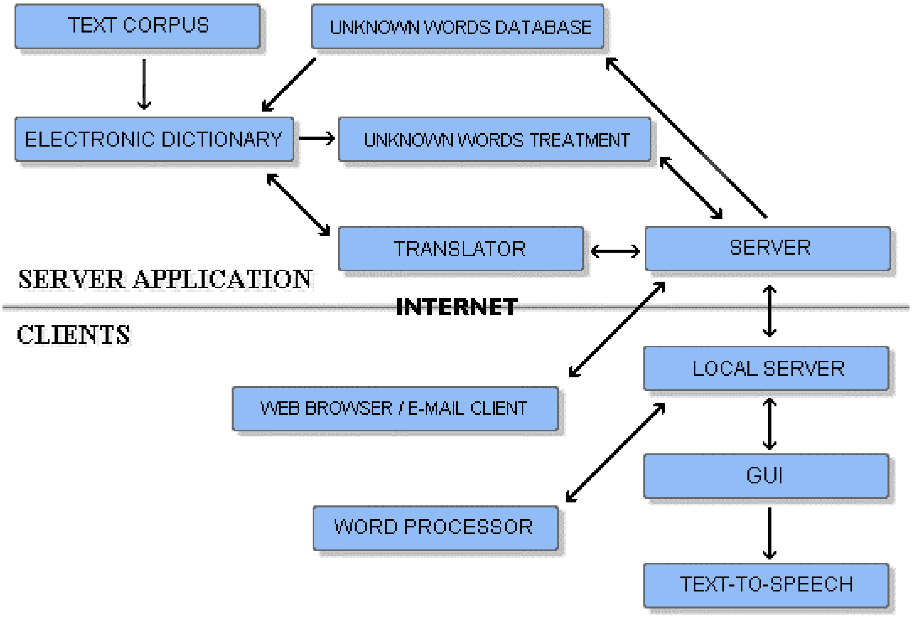

**Fig. 1. Structure of the POLENG 2 system with data flow**

### 3.1. Text Corpus

Within the confines of the project a corpus of computer texts has been collected. Source files of the corpus were collected from computer magazines, CD-ROMs and Web pages. The procedure of text selection consisted of three stages:

(1) Automatic search for files containing Polish texts (this consisted in verifying that a file was of a textual type and that frequencies of its letter clusters were characteristic of Polish). About one hundred CD-ROMs have thus been searched. This stage was completed by converting all the files to the common code standard.

(2) Automatic search for computer texts in the files selected at the previous stage. One of the most effective methods used for this purpose consisted in checking the occurrence of words

characteristic of computer science (the list of such words had been obtained on the basis of a smaller corpus collected for the needs of the previous version of the system).

(3) Final manual selection of texts.

As a result of the procedure presented above a fairly large corpus (over 1 100 000 words) was collected. The next step was to perform morphological analysis of words contained in the corpus in order to create a list of Polish lexemes that was to provide a basis for the ELECTRONIC POLISH-ENGLISH DICTIONARY. More information about the corpus can be found in [6].

### 3.2. Electronic Dictionary

The function of the ELECTRONIC DICTIONARY is to recognise morphological, syntactic and semantic features of the input words and phrases in Polish and to return adequate information for the English translations of inputs.

Originally, the dictionary has the textual SGML[3] format but it is converted to a binary form afterwards. The binary form is much more effective during the search phase. The procedures for consulting ELECTRONIC DICTIONARY are linked to a DLL[4] which makes them applicable to TRANSLATOR. More information about the dictionary can be found in [7] and [8].

### 3.3. Translator

The TRANSLATOR module is an implementation of the translating algorithm for standard Polish constructions (the grammar is based mainly on the descriptions included in [9] and [10]). The module applies information contained in ELECTRONIC DICTIONARY as well as that returned by the MODULE FOR THE TREATMENT OF UNKNOWN WORDS in case of occurrence of words from beyond the dictionary.

One of the important capabilities of the TRANSLATOR is the interpretation of the subset of ITF markups. ITF (Internal Translation Format) is a simple markup language, designed especially for the needs of POLENG, which aims at marking specific substrings of input texts. ITF is used for pointing out beginnings of sentences, nonlexical text fragments, HTML/XML markups and some punctuation marks. The TRANSLATOR module uses ITF also for returning an option lists of some words (or phrases) in the translated texts.

[3] Standard Generalised Markup Language

[4] Dynamic Linked Library

Transfer rules are written in Prolog. The Prolog engine is linked to the SERVER module, which considerably increases the speed of the internal data transfer.

For more information see [2] and [7].

### 3.4. Unknown Words Treatment

The module for UNKNOWN WORDS TREATMENT serves as a spell-checker of the input texts and makes hypotheses about the morpho-syntactic features of words from beyond the dictionary. This module extends the flexibility of the system in situations such as typos or proper names in source texts.

More about the module can be found in [4].

### 3.5. Server

The main tasks of the SERVER module are text preprocessing and organising work for the other components of the system and services.

Text preprocessing tasks carried out by SERVER are: sentence delimitation, auto-detection of character codepage and processing HTML/XML formatted texts. Auto-detection of the codepage of a text is required because of the variety of standards for coding Polish diacritic characters (like ‘ą’, ‘ę’ and ‘ł’) using 8-bit codings. SERVER recognises the standard of a text and converts the text into the unified character set. Another task executed in the preprocessing stage is translating HTML/XML markups into ITF (interpretable by TRANSLATOR). This enables the translation of WWW pages.

The coordination of work for the other components consists in the support of the Internet access to the system, controlling jobs, multiprocessing support and monitoring the resources. SERVER is a place where any MT transaction in POLENG2 has to be registered. SERVER has also special conservation tasks that will be useful in the process of upgrading linguistic resources.

One of the main requirements for SERVER is its efficiency in parallel processing – this module should be able to take advantage of multiprocessor server computers.

### 3.6. Unknown Words Database

The database contains words missing from the dictionary that occurred in input texts. The main function of the database is to facilitate the updates of the dictionary. UNKNOWN WORDS DATABASE makes possible the creation of the frequency lists for words from beyond the dictionary and found in source texts. It seems to be the best method for determining a list of lexemes for dictionary update.

UNKNOWN WORDS DATABASE is fully controlled by SERVER.

### 3.7. Local Server

The LOCAL SERVER application is a kind of a protocol translator. LOCAL SERVER is a remote client for the POLENG2 server application (through Internet). On the other hand, LOCAL SERVER is a server application for local clients (applications running on clients' computers).

LOCAL SERVER is a central part of the client's software for the POLENG2 system. All client applications are connected to LOCAL SERVER. Integrating LOCAL SERVER with some of the SERVER services will result in a local, stand-alone version usable on a single workstation. There emerges an issue of computational balance between SERVER and LOCAL SERVER in the distributed version. In order to distribute computations it seems desirable to shift the maximum load from SERVER to LOCAL SERVER. This, however, has two drawbacks: the time of downloading the client's software is lengthened and the upgrade of the tools and data in LOCAL SERVER is no longer possible.

### 3.8. GUI

GRAPHICAL USER INTERFACE consists of two synchronised text edit windows - one for a source text and one for the translated text. Both windows support standard editing techniques as well as some features characteristic of the application.

When the translation command is called, a text from the source edit window is sent to LOCAL SERVER and further to SERVER where the translation process continues. Then, SERVER sends the source text equipped with prompts received from the UNKNOWN WORDS TREATMENT module and the translated text in English back to the client's edit windows. The graphical interface displays the prompts in the source edit window and variant English equivalents in the target edit window.

In many aspects the GUI for POLENG2 mirrors the conceptions tested in the previous version of POLENG, as described in [11]. The biggest changes will appear in new restrictions for text viewing and editing.

### 3.9. Text-to-Speech

POLENG2 GUI is prepared to take advantage of an external English TtS system. In the previous version of the POLENG the system INFOVOX220 was used.

### 3.10. Word Processor

The WORD PROCESSOR module represents a script that makes translation available to users of a standard word processor. The execution of the script may be called by clicking a button on a toolbar. The script communicates with LOCAL SERVER in order to translate a selected text. Results can be shown behind the source text in the edit window or copied into clipboard. Such a script is planned to be written for Microsoft Word first.

### 3.11. Web Browser / E-mail Client

Apart from using dedicated software clients for the POLENG2 system, there are other ways of working with the translator. These additional possibilities are the access to the translation service by a web browser or by an e-mail client. The first feature will be realised by an HTML form on the POLENG web page (see also [12]). The e-mail service should allow for sending translation requests with a source text in an e-mail massage. The translated text is to be mailed back to the sender.

## 4. Summary

The POLENG MT system is being expanded in two directions: improving linguistic efficiency and broadening the range of MT services. POLENG2 is still being developed and it will reach its final shape by the end of 2000. It is currently possible to test the system (in its development state) by contacting POLENG2 homepage at www.poleng.pl.

REFERENCES[5]

[1] K. Jassem, *Struktura systemu tłumaczenia automatycznego z języka polskiego na język angielski, (The Structure of a Polish - English MT System),* in: Speech and Language Technology, vol.2, Poznań, 1998

[2] K. Jassem, *Dealing with Free Order and Non-Language Markers in a Top-Down-Left-First Algorithm,* in: Speech and Language Technology, vol.4, Poznań, 2000

[3] R. Korzeniewski, Speech and Language Technology, vol.4, Poznań, 2000

[4] J. Daciuk, *A module for treatment of unknown words,* in: Speech and Language Technology, vol.3, Poznań, 1999

[5] Most of the referenced papers are accesible on the POLENG project homepage at www.poleng.pl.

[5] F. Graliński, G. Krynicki, *Word-formation analysis in Polish-English Machine Translation,* in: Speech and Language Technology, vol.4, Poznań, 2000

[6] F. Graliński, *Hasłowanie korpusu polskich tekstów informatycznych (1,2 mln słów) — raport, (*Word Lemmatisation of a Corpus of Polish Computer Texts (1,2 mln words) – Report), in: Speech and Language Technology, vol.4, Poznań, 2000

[7] K. Jassem, *Examples of Applying the Bilingual Dictionary to the Translation Algorithm in System POLENG,* in: Speech and Language Technology, vol.4, Poznań, 2000

[8] M. Lison, *Model i realizacja struktury danych leksykalnych słownika komputerowego skonstruowanego dla potrzeb automatycznego tłumaczenia.* (A Model and the Implementation of a Structure of Lexical Data in an Electronic Dictionary Designed for the Purpose of MT), in: Speech and Language Technology, vol.2, Poznań, 1998

[9] S. Szpakowicz, *Formalny opis składniowy zdań polskich,* Wydawnictwa Uniwersytetu Warszawskiego, Warszawa 1986

[10] M. Świdziński, *Gramatyka formalna języka polskiego,* Wydawnictwa Uniwersytetu Warszawskiego, Warszawa 1992

[11] F. Graliński, M. Wypych, *POLENG SHELL – graficzny interfejs użytkownika dla systemu tłumaczenia maszynowego POLENG,* (POLENG SHELL – a Graphical User Interface for the Machine Translation System POLENG), in: Speech and Language Technology, vol.3, Poznań, 1999

[12] Paweł Bogusławski, *Komunikacja pomiędzy użytkownikiem i systemem tłumaczenia automatycznego poprzez Internet,* (Communication Between the User and the MT System via the Internet), in: Speech and Language Technology, vol.4, Poznań, 2000