D6.1: POST-PROCESSING MODULE

Grant Agreement number	ICT-248307
Project acronym	**PRESEMT**
Project title	**P**attern **RE**cognition-based **S**tatistically **E**nhanced **MT**
Funding Scheme	Small or medium-scale focused research project – STREP – CP-FP-INFSO
Deliverable title	**D6.1: Post-processing module**
Version	**V3**
Responsible partner	GFAI
Dissemination level	Restricted
Due delivery date	31.7.2011 (+60 days)
Actual delivery date	20.9.2011

Project coordinator name & title	**Dr. George Tambouratzis**
Project coordinator organisation	**Institute for Language and Speech Processing / RC 'Athena'**
Tel	+30 210 6875411
Fax	+30 210 6854270
E-mail	giorg_t@ilsp.gr
Project website address	www.presemt.eu

Contents

Figures

Tables

1. Executive summary

The present deliverable, falling within Task *T6.1: Design and development of the post-processing module* (**WP6: Post-Processing & User Adaptation**), describes the design and implementation of the Post-processing module (**PPM**).

The particular module gives the end user the opportunity to perform modifications to the system translation output according to their preferences and feed them back to the system, which in turn can later exploit them towards automatic adaptation.

The current implementation involves two types of post-processing functionalities, namely (a) selection of lexical alternatives, where the user can substitute words (or phrases) of the system translation output with alternatives drawn from the system lexicon, and (b) free-post-editing, which entails the unrestricted alteration of the system output, ranging from reordering to modification and insertion and deletion.

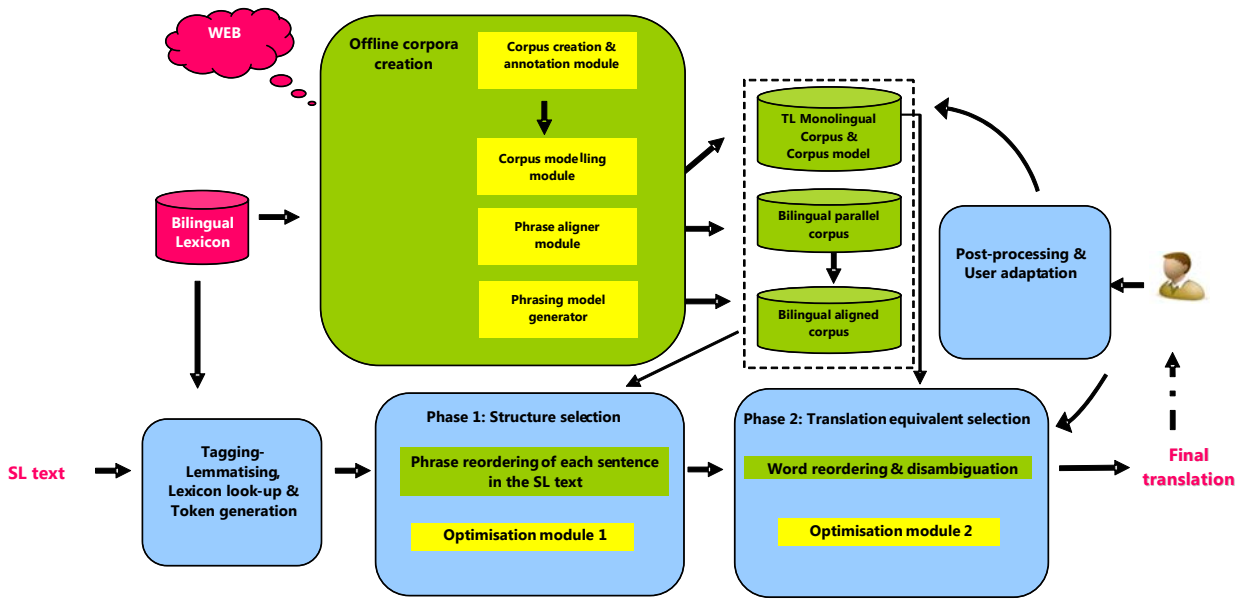
The module design also foresees the development of another functionality (*not yet implemented*), which gives the user the opportunity to save their changes for future use or revisions.

The deliverable has the following structure: Section 2 provides a brief description of the post-processing stage in PRESEMT as a whole and of the Post-processing module in particular. Section 3 contains a concise survey of post-editing interfaces provided by other MT systems. Section 4 presents the implemented functionalities of the Post-processing module, while Section 5 describes the module's architecture. References are listed in Section 6.

Table 1: List of abbreviations

Abbr	Term
GUI	Graphical User Interface
ISS	Input Source Sentence
PPM	Post-processing module
SL	Source Language
TL	Target Language
UAM	User adaptation module

Figure 1: PRESEMT system architecture



2. Post-processing stage in PRESEMT

The Post-processing stage in PRESEMT, which handles the user feedback in terms of post-editing and system modification, comprises two modules:

1. The **Post-processing module** (**PPM**), which enables the user to modify the system output for a given set of SL sentences.
2. The **User adaptation module** (**UAM**), which, by receiving the user-modified output, allows the user to customise the system behaviour.

2.1 Post-processing module

The Post-processing module receives as input the translation output of the main translation engine modules, i.e. (1) Structure selection and (2) Translation equivalent selection.

A special Graphical User Interface (GUI), has been designed, using which the end user can (a) examine the system translation output in parallel with the original text, (b) modify the output according to their preferences and (c) store their changes for future reference. The particular module features *genericness*, i.e. it is not determined by the translation domain or the type of application, and *language-independence* (cf. *D2.1: System specifications*).

Potential corrections by the user include lexical substitution, reordering, word deletion or insertion, free text edit etc.

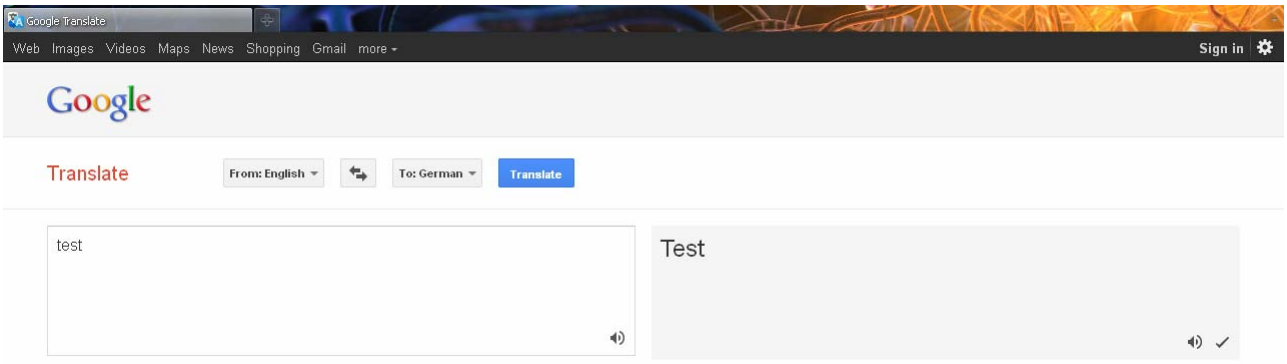
3. Literature and system survey

Before deciding on the form and functions of the Post-processing module, other post-processing interfaces of MT systems have been studied, for instance the ones provided by **SYSTRAN** (as reviewed by the PACO-MT project[1]), **GoogleTranslate[2], the GoogleTranslator Toolkit[3]**, and the **FAUST** project[4]. Moreover, we have consulted the relevant **TAUS[5]** report.

The common element of all these interfaces is that the source language text and its target language translation are represented in a conspicuous way, so that the system translation output can be easily compared to its source language input (cf. Figure 2, where the interface of GoogleTranslate is illustrated).

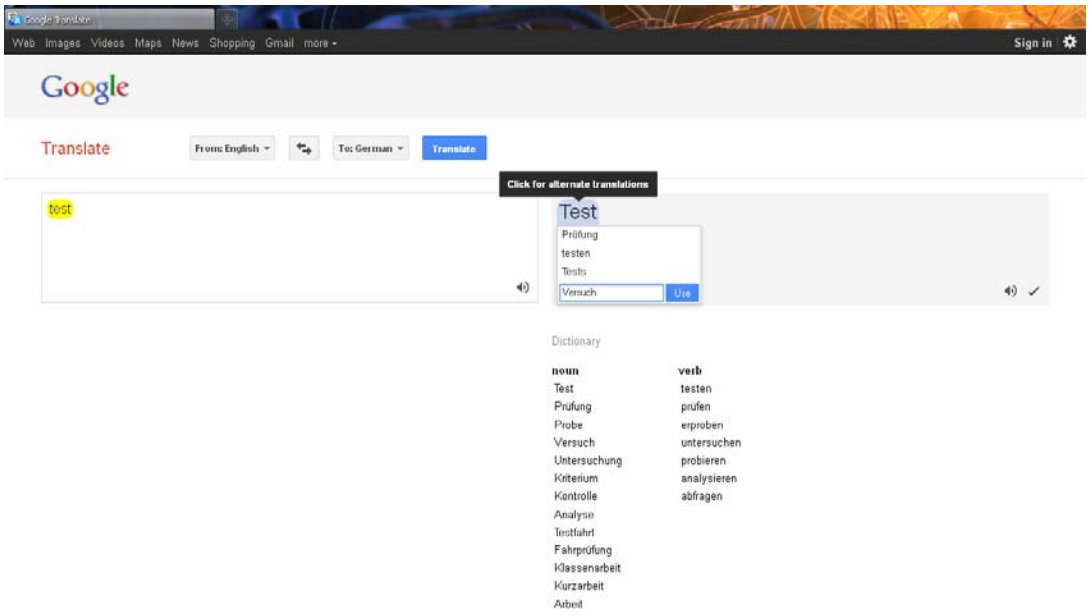
[1] http://www.ccl.kuleuven.be/Projects/PACO/paco.php
[2] http://translate.google.com/
[3] http://translate.google.com/toolkit
[4] faust-fp7.eu
[5] http://www.translationautomation.com/

Figure 2: Translation interface of GoogleTranslate



The same holds for the post-editing process, where it is important for the user to view both source and language text simultaneously, when making changes to the system translation output (cf. Figure 4, where an example of the post-editing interface in GoogleTranslate is illustrated).

Figure 3: Post-editing in GoogleTranslate



Regarding the post-processing functionalities provided to the user, these can be subsumed under two types:

1. **Lexical substitution:** the user can substitute the translation of a word (or a phrase) with a different one of their choice or with an alternative one as suggested by the MT system
2. **Reordering, deletion, insertion of elements:** the user can delete or insert words (or phrases); they can likewise change the word order

Since there is not much variability wrt the post-processing interfaces and functionalities, it has been decided that the PRESEMT post-processing module focusses on what is novel in PRESEMT, which is the User adaptation module, and that it provides a lean but yet user-friendly and efficient user interface (Figure 4).

Figure 4: Post-processing interface in PRESEMT

4. Post-processing functionalities

The PRESEMT system provides the user with two types of post-processing functionalities, (a) the **selection of lexical alternatives** and (b) **free post-editing**. For system-internal reasons the two functionalities are displayed in two discrete post-processing windows, while they are offered to the user consecutively, i.e. the post-editing process becomes available, only after the selection of lexical alternatives has been completed.

4.1 Selection of lexical alternatives

This functionality entails the use of a different translation for a word (or phrase) than the one produced by the system. By highlighting a specific word (or phrase) in the system output, a drop-down menu opens, listing all the lexical alternatives provided by the system lexicon, from which the user may select one and replace the original translation[6].

Figure 5 and Figure 6 illustrate an example of selecting lexical alternatives for the word "*test*", when translating from English to German.

Figure 5: Selection of lexical alternatives

[6] The selection-of-lexical-alternatives functionality might have to be deactivated for certain language pairs if the bilingual dictionary used has publication restrictions in its licensing.

Figure 6: Replacing the original translation with a lexical alternative

After selecting a different translation, the original translation is subsequently listed among the lexical alternatives, so that the user can revise their decision and return to the original one.

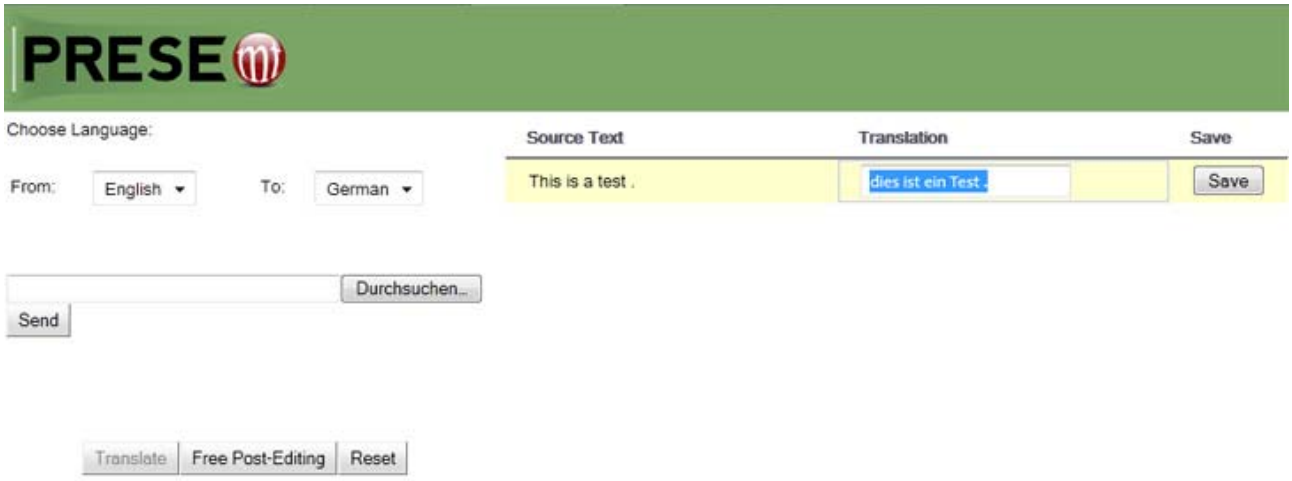
4.2 Free post-editing

The specific functionality receives as input the output of the selection-of-lexical-alternatives functionality. Once the free post-editing has been initiated, the user cannot return to the first functionality, because any modifications made during the free post-editing destroy the links between the original translation and the lexical alternatives.

Source and target language texts are displayed in two vertical columns and are aligned sentence-wise (Figure 7), resulting in the creation of bilingual sentence pairs that can later be fed into the User adaptation module as an additional bilingual corpus. Free post-editing is done in a free-input text box (Figure 8).

Figure 7: Free post-editing page

Figure 8: Free post-editing text field



4.3 Memory functionality

The **'Save'** button allows the user to save their modifications (*this functionality is not yet implemented*). More specifically, the system stores triples[7] consisting of

* the Input Source Sentence (ISS)
* the original system translation &
* the output of the post-processing

The pair consisting of the ISS and the output of the post-processing can be used by the User adaptation module for system enhancement. The original system translation is stored for evaluation purposes. It allows monitoring and evaluating changes in system behaviour after user adaptation has taken place. Besides, it allows the user to review previous translations and their modifications (*this functionality is not yet implemented*).

Furthermore, the user may modify future translations of the PRESEMT system by feeding back translation modifications into the system. These modifications can be used to establish new translation patterns that might improve later translations (*this functionality is not yet implemented*).

4.4 Multi-user functionality

A multi-user functionality is also a requirement that stems from the user and is of particular importance if web applications are considered. But even for an MT system as a stand-alone system, multi-user applications might be relevant if corporate applications are considered. However, since it is not the main task of the PRESEMT research project to re-programme multi-user environments that have been implemented in other systems, it is planned to restrict the multi-user functionality to a fixed number of users in order to provide the proof of concept that the PRESEMT architecture can handle multiple users.

[7] It is not decided yet what the contents of the third slot will be if there is no user modification; it can either be left empty or be filled with the system translation.

5. System architecture and implementation of the interface

The post-processing interface is implemented as a web interface with a client-server architecture. The server provides the output of the PRESEMT translation phases in such a form that it also includes lexical alternatives of word translations.

The system can be run as a web application with access to a server installed at one of the project partners' machines – most likely ICCS – or it can be run locally on a desktop computer or workstation after downloading the complete system. Since it is standard nowadays that computers have a web browser installed, the web interface can also be used if the system is run locally.

For the implementation of the user interface the Google Web Toolkit (GWT)[8] has been used.

6. References

Knight, K. and Chander, I. (1994): Automated postediting of documents. Proceedings of the 12th National Conference on Artificial Intelligence, AAAI Press, Menlo Park, CA.

Michel Simard, Cyril Goutte & Pierre Isabelle (2007): Statistical Phrase-based Post-editing, in: Proceedings of NAACL HLT 2007, pages 508–515, Rochester, NY, April 2007. Association for Computational Linguistics.

Vandeghinste, V., and Martens, S. (2009): Top-down Transfer in Example-based MT. Proceedings of the 3rd International Workshop on Example-based Machine Translation, Dublin City University, Dublin, Ireland, pp. 69-76.

Vandeghinste, V., and Martens, S. (2009): Paco-MT Project, D5.1 Report on State of the art in Post-Editing Interfaces.

Postediting in Practice (2010): A TAUS Report, TAUS BV, De Rijp, The Netherlands.

[8] code.google.com/webtoolkit