Pattern REcognition-based Statistically Enhanced MT

ANNUAL PUBLIC REPORT 3

Grant Agreement number	ICT-248307
Project acronym	**PRESEMT**
Project title	**P**attern **RE**cognition-based **S**tatistically **E**nhanced **MT**
Funding Scheme	STREP – CP-FP-INFSO
Deliverable title	**Public Annual Report 3**
Dissemination level	Public
Period covered	**2012**
Responsible partner	**ILSP**

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

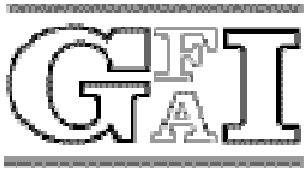
PRESEMT consortium & contact persons

Institute for Language and Speech Processing/R.C. "Athena"

Coordinator

http://www.ilsp.gr/

Contact person: **Dr. George Tambouratzis,** giorg_t@ilsp.gr



Gesellschaft zur Förderung der Angewandten Informationsforshung e.V.

http://www.iai-sb.de/iai/index.php/en/Die-GFAI.html

Contact person: **Dr. Paul Schmidt,** paul@iai.uni-sb.de

Norges Teknisk-Naturvitenskapelige Universitet

http://www.ntnu.no/

Contact person: **Prof. Björn Gambäck,** gamback@idi.ntnu.no

Institute of Communication and Computer Systems

http://www.iccs.gr/eng

Contact person: **Dr. Georgios Goumas,** goumas@cslab.ece.ntua.gr

Masaryk University

http://www.muni.cz/

Contact person: **Prof. Karel Pala,** pala@fi.muni.cz

Lexical Computing Ltd.

http://www.sketchengine.co.uk/

Contact person: **Dr. Adam Kilgarriff,** adam.kilgarriff@gmail.com

Table of Contents

1. PRESEMT overview

PRESEMT (**P**attern **RE**cognition-based **S**tatistically **E**nhanced **MT**) is an EU-funded project under the FP7 topic "ICT-2009.2.2: Language-based Interaction". It has been intended to lead to a flexible and adaptable Machine Translation (MT) system, based on a language-independent method, whose principles ensure easy portability to new language pairs. This method attempts to overcome well-known problems of other MT approaches, e.g. compilation of extensive bilingual corpora or creation of new rules per language pair. PRESEMT addresses the issue of effectively managing multilingual content, suggesting a language-independent machine-learning-based methodology.

In order for PRESEMT to be easily amenable to new language pairs, only relatively inexpensive, readily available language resources as well as bilingual lexica are used. The translation context is modelled on phrases, as they have been proven to improve the translation quality. Phrases are produced via a semi-automatic and language-independent process of morphological and syntactic analysis, removing the need for compatible, in terms of output, NLP tools per language pair.

Parallelisation of the main translation processes has been implemented in order to reach a fast, high-quality translation system.

To allow for user adaptability, all the corpora used in PRESEMT are retrieved from web-based sources via the system platform, while user feedback is integrated through appropriate interactive interfaces.

Key innovation

The PRESEMT project proposes a novel approach to the problem of Machine Translation by introducing in the MT paradigm cross-disciplinary techniques, mainly borrowed from the **machine learning** and **computational intelligence** domains.

To this end, a flexible MT system has been developed, which is enhanced with (a) **pattern recognition** techniques (such as template matching and neural networks) towards the development of a language-independent analysis and (b) **evolutionary computation** methods (such as Genetic Algorithms or Swarm Intelligence) for system optimisation.

Features

The core features of PRESEMT are listed below:

1. Development of a novel method based on **generalised clustering techniques**, for creating a **language-independent phrase aligner** adaptable to phrasing principles defined by the end users
2. Use of **pattern recognition** approaches for defining **syntactic structure**
3. Employment of techniques inspired by **functional biological systems** for **disambiguating** between candidate translations
4. Study of **automated optimisation techniques** to define a mature system for methodically **optimising** system parameters
5. Application of **machine learning** methods for allowing system **adaptation**
6. Use of **parallel computing** architectures as well as mainstream multi-core architectures for PCs to achieve substantial improvements in **translation speed**

2. PRESEMT system description

The PRESEMT system, the architecture of which is depicted in Figure 1, roughly comprises 3 components, each of them having a modular structure (cf. Table 1):

1. **Pre-processing stage:** It involves the compilation of resources needed for the MT system to operate, i.e. the collection and appropriate annotation of corpora, the elicitation of phrasing information as well as the extraction of semantic and statistical data.
2. **Main translation engine:** This component, being the core part of the system, translates a source language (SL) text to a target language (TL) one, drawing, in stepwise mode, on the information obtained in the Pre-processing stage.
3. **Post-processing stage:** This stage offers the user the opportunity to modify the system translation output according to their preferences. These modifications can then be endorsed by the system so as to adapt itself to the given input.

Figure 1: PRESEMT system architecture

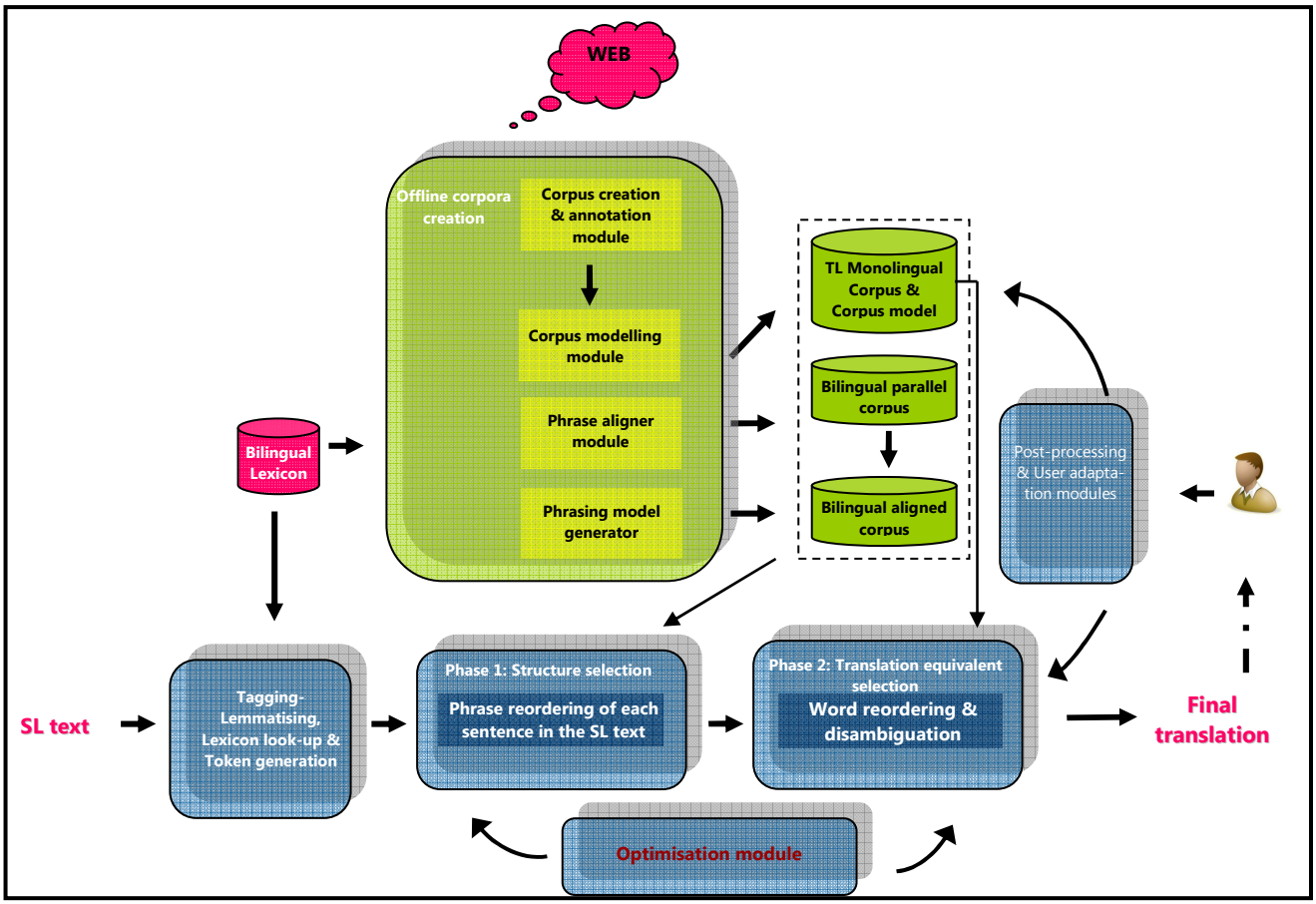


Table 1: PRESEMT basic system modules

Pre-processing stage: 4 modules	Main translation engine: 3 modules	Post-processing stage: 2 modules
Corpus creation & annotation module	Structure selection module	Post-processing module
Phrase aligner module	Translation equivalent selection module	
Phrasing model generator	Optimisation module	User adaptation module
Corpus modelling module		

Pre-processing stage

The **Corpus creation & annotation module** entails the compilation and annotation of large monolingual and small bilingual corpora to be utilised by the Main translation engine. The former are collected via web crawling, while the latter are created manually (mainly based on web resources). The collected text resources are submitted to various levels of processing (e.g. monolingual corpora: cleaning and content de-duplication; bilingual corpora: corrections / modifications) and annotation (e.g. Part-of-Speech (PoS) tagging and lemmatisation).

The **Phrase aligner module** (**PAM**), operating on bilingual corpora (cf. the aforementioned ones), performs word-and-phrase-level alignment of a bilingual corpus, one side of which is annotated only with PoS tags and lemmata, while the other one additionally bears phrasing information. In the current implementation the source language is assumed to be the non-parsed side of the language pair, while the target language is fully annotated. After determining lexical correspondences within a given language pair and on the basis of the TL parsing, the Phrase aligner proceeds to segmenting the SL corpus side into phrases. It subsequently outputs the bilingual corpus aligned at clause, phrase and word level.

The **Phrasing model generator** (**PMG**) takes as input the output of the Phrase aligner and utilises it so as to (a) generate a probabilistic phrasing model for the source language and (b) apply this model for segmenting a given SL text being input for translation. For the first task the module operates offline, whereas the second task is an online process that forms part of the actual translation procedure.

The last module of this stage, the **Corpus modelling module**, takes as input an annotated TL monolingual corpus (yielded by the Corpus creation & annotation module) and processes it so as to extract semantic-type and statistical-based information (by applying methods such as n-gram models over words and PoS tags, SOM for words and vector space models). This type of information is then utilised during the translation process for lexical disambiguation purposes.

Main translation engine

The Main translation engine is split into two phases:

The **Structure selection module** determines the optimal structure of an SL sentence, by utilising information residing in the bilingual corpus.

The **Translation equivalent selection module** disambiguates translation equivalents and microstructures, after the SL sentence structure has been established, by utilising information residing in the TL monolingual corpus.

The **Optimisation module** is responsible for enhancing the performance of the two translation phases, by optimising the values of the parameters employed.

Post-processing stage

The **Post-processing module** is a GUI via which the user can feedback their modifications to the system translation output.

The **User adaptation module** collects the user modifications and "corrects" the translation system accordingly.

Language pairs covered

The language pairs, to be examined as case studies and for evaluation purposes, have been selected on the basis of three criteria, namely (a) availability of a large TL corpus, (b) examination of different language families and (c) coverage of the consortium languages.

The left-hand column of the following table illustrates the language pairs that have been handled for the development of the first two versions of the system prototype, whereas the right-hand column lists the language pairs that have been used for system assessment.

Table 2: Language pairs covered by PRESEMT

Language pairs (development phases 1 & 2)	**Language pairs** (development phase 3)
* Czech ⇒ English	* Czech ⇒ Italian
* German ⇒ English	* English ⇒ Italian
* Greek ⇒ English	* German ⇒ Italian
* Norwegian ⇒ English	* Greek ⇒ Italian
* Czech ⇒ German	* Norwegian ⇒ Italian
* English ⇒ German	
* Greek ⇒ German	
* Norwegian ⇒ German	

3. Activities within the 3rd year of the project

During the third year of the project the four work packages revolving around the design and implementation of PRESEMT modules, namely WP3, WP4, WP5 and WP6, have been completed; so all remaining activity in improving the technical modules has been within the work package focussing in the integration of the system modules into a single platform, i.e. WP7. The relevant activities have culminated at the end of the third project year into the release of the final versions of the different modules. In addition, the second and third PRESEMT prototypes have been released after months M24 and M36.

During the latter part of the third year, the validation and evaluation activities have also been initiated. Validation involved investigating possible performance errors of the modules developed, whereas evaluation concerned assessing, via objective and subjective (i.e. by humans) metrics, the quality of the translation produced by the PRESEMT system prototype. The human evaluation in particular has been performed by consortium-external users, who have been recruited for this purpose in all 5 countries represented in the project (Czech Republic, Germany, Greece, Norway and UK).

In the remainder of this section, the main results obtained and objectives achieved during the 3rd year are summarised per work package.

WP7: Integration

The main aim of WP7 has been to integrate the modules developed into the technical work packages (i.e. WP3, WP4, WP5 and WP6) into one working prototype and to exploit parallelisation opportunities towards the efficient application of the prototype to multi-processor/multi-core architectures.

To support this effort, the Apache Subversion revision control server established has been used in joining together all the modules developed by the partners. This has led to the release of the 2nd and 3rd PRESEMT prototypes in April 2012 and February 2013 respectively, together with the associated documentation and user manual. The new prototypes achieve improved translation accuracy in comparison to the first version in terms of objective MT quality metrics (BLEU, NIST, TER and Meteor).

Besides, the parallelisation of the translation engine has led to a reduction in execution time by a factor of 2 or more.

WP9: Validation & Evaluation

The purpose of the specific work package was twofold, namely to validate the PRESEMT system and certain individual modules in terms of performance as well as evaluate the quality of the translation output.

Validation: The validation process was carried out in two phases at each partner's site. PRESEMT team-external validators were engaged for checking the translation interface and its underlying functionalities, i.e. the translation process, the post-processing and the user adaptation. Validators were requested to document their experimentation with the system by filling in the respective validation forms, which have been compiled for this purpose.

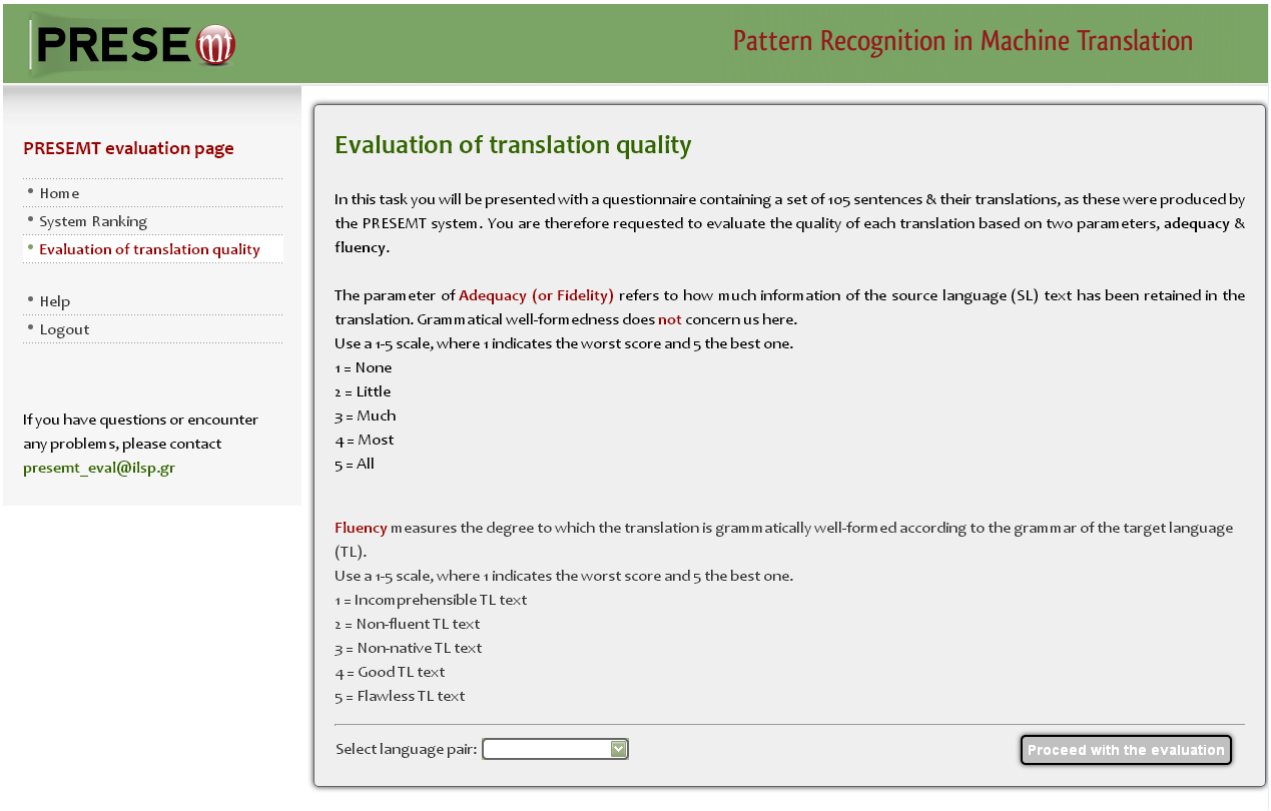
The comments of the validators mainly related to performance errors and their suggestions focussing on the user-friendliness of the interface were taken into account to a great extent for improving the corresponding system modules.

Evaluation: The human evaluation activities have been carried out for the eight main language pairs, namely from Czech, English, German, Greek and Norwegian to English and German, and spread over all 5 countries represented in the project. The evaluators have been recruited by the different partners (each partner was responsible for those pairs where their native language was the source language), and have

been called to perform two tasks: (a) rank various MT systems on the basis of their translation output on the same set of data and (b) to assess the quality of translations in terms of adequacy and fluency.

The evaluation GUI (cf. Figure 2), via which the end users accessed the data, has been finalised during May 2012. The evaluation itself has been performed on the latest version of the PRESEMT system which was valid in November 2012. The reason for that was to evaluate the advantages as well as the limitations of this MT methodology using an advanced prototype. The results obtained are currently being processed (late January 2013) and will be released within a dedicated project deliverable as well as in presentations in scientific fora.

Figure 2: Screenshot of the PRESEMT evaluation platform



4. Dissemination activities

The following tables summarise the main dissemination activities undertaken by the PRESEMT consortium members during the third year of the project, as well as any activities planned for the immediate future.

PRESEMT dissemination activities							
Activity name	**Type**	**Event**	**Place**	**Date**	**Authors**	**Site**	**Status**
Machine Translation, Natural Language Interfaces	Lecture	---	Trondheim, Norway	January – May 2012	Björn Gambäck, Lars Bungum, Erwin Marsi	NTNU	---
Presentation / Publication details & Comments							
7th Web as corpus Workshop (WAC-7)	Workshop	In conjunction with World Wide Web 2012 (WWW2012)	Lyon, France	April 17, 2012	---	---	Held
Presentation / Publication details & Comments	The workshop was jointly sponsored by PRESEMT and ACL SIGWAC (Special Interest Group on Web As Corpus). Adam Kilgarriff (LCL) was member of the organising committee.						
The PRESEMT project for the Machine Translation Task	Workshop presentation	Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (**ESIRMT**) and Hybrid Approaches to Machine Translation (**HyTra**) *[In conjunction with the 13th Conference of the European Chapter of the Association for Computational Linguistics (**EACL2012**)]*	Avignon, France	April 23, 2012	George Tambouratzis, Marina Vassiliou & Sokratis Sofianopoulos	ILSP	Published
Presentation / Publication details & Comments	Oral presentation of the PRESEMT project by George Tambouratzis						

PRESEMT dissemination activities							
Activity name	**Type**	**Event**	**Place**	**Date**	**Authors**	**Site**	**Status**
Efficient N-gram Language Modeling for Billion Word Web-Corpora	Workshop paper	Challenges in the Management of Large Corpora (**CMLC**) [In conjunction with the 8th International Conference on Language Resources and Evaluation (**LREC2012**)]	Constantinople, Turkey	May 22, 2012	Lars Bungum & Björn Gambäck	NTNU	Published
Presentation / Publication details & Comments							
The PRESEMT project	Workshop presentation	5th Workshop on Building and Using Comparable Corpora (**BUCC2012**) *[In conjunction with the 8th International Conference on Language Resources and Evaluation (**LREC2012**)]*	Constantinople, Turkey	May 26, 2012	Adam Kilgarriff & George Tambouratzis	ILSP & LCL	Published
Presentation / Publication details & Comments	Invited talk given jointly by George Tambouratzis (ILSP) and Adam Kilgarriff (LCL)						
Accurate phrase alignment in a bilingual corpus for EBMT systems	Poster presentation	5th Workshop on Building and Using Comparable Corpora (**BUCC2012**) *[In conjunction with the 8th International Conference on Language Resources and Evaluation (**LREC2012**)]*	Constantinople, Turkey	May 26, 2012	George Tambouratzis, Michalis Troullinos, Sokratis Sofianopoulos, Marina Vassiliou	ILSP	Published
Presentation / Publication details & Comments							
Word Sketches for Turkish	Conference paper	8th International Conference on Language Resources and Evaluation (**LREC2012**)	Constantinople, Turkey	May 21-27, 2012	Bharat Ram Ambati, Siva Reddy, Adam Kilgarriff	LCL	Published
Building a 70 billion word corpus of English from ClueWeb	Conference paper	8th International Conference on Language Resources and Evaluation (**LREC2012**)	Constantinople, Turkey	May 21-27, 2012	Jan Pomikálek, Miloš Jakubíček, Pavel Rychlý & Adam Kilgarriff	MU	Published
Presentation / Publication details & Comments							

PRESEMT dissemination activities							
Activity name	**Type**	**Event**	**Place**	**Date**	**Authors**	**Site**	**Status**
Measuring Distance between Language Varieties	Conference paper	The Sixth Inter-Varietal Applied Corpus Studies (**IVACS**) group International Conference on Corpora across Linguistics	Leeds, UK	June 21-22, 2012	Adam Kilgarriff, Jan Pomikálek, Pavel Rychlý & Vit Suchomel	LCL	Published
Presentation / Publication details & Comments							
Implementing a language-independent MT methodology	Conference paper	1st Workshop on Multilingual Modeling (**MM-2012**) [In conjunction with the 50th Annual Meeting of the Association for Computational Linguistics (**ACL2012**)]	Jeju Island, Republic of Korea	July 13, 2012	Sokratis Sofianopoulos, Marina Vassiliou & George Tambouratzis	ILSP	Published
Presentation / Publication details & Comments							
Finding Multiwords of More Than Two Words	Conference paper	15th EURALEX International Congress	Oslo, Norway	August 7-11, 2012	Adam Kilgarriff, Pavel Rychlý, Vojtěch Kovář & Vít Baisa	LCL	Published
Presentation / Publication details & Comments							
SOM-based corpus modeling for disambiguation purposes in MT	Workshop paper	Hybrid Machine Translation Workshop [In conjunction with the 15th International Conference on Text, Speech and Dialogue (**TSD2012**)]	Brno, Czech Republic	September 3, 2012	George Tambouratzis, George Tsatsanifos, Ioannis Dologlou & Nikos Tsimboukakis	ILSP	Published
Presentation / Publication details & Comments							
Disambiguating word translations with target language models	Workshop paper	Hybrid Machine Translation Workshop [In conjunction with the 15th International Conference on Text, Speech and Dialogue (**TSD2012**)]	Brno, Czech Republic	September 3, 2012	André Lynum, Erwin Marsi, Lars Bungum & Björn Gambäck	NTNU	Published
Presentation / Publication details & Comments							

PRESEMT dissemination activities							
Activity name	**Type**	**Event**	**Place**	**Date**	**Authors**	**Site**	**Status**
User Adaptation in a Hybrid MT System: Feeding User Corrections into Synchronous Grammars and System Dictionaries	Workshop paper	Hybrid Machine Translation Workshop [In conjunction with the 15th International Conference on Text, Speech and Dialogue (**TSD2012**)]	Brno, Czech Republic	September 3, 2012	Susanne Preuß, Hajo Keffer, Paul Schmidt, Georgios Goumas, Athanasia Asiki & Ioannis Konstantinou	GFAI & ICCS	Published
Presentation / Publication details & Comments							
Hybrid Machine Translation workshop	PRESEMT Workshop	Hybrid Machine Translation Workshop [In conjunction with the 15th International Conference on Text, Speech and Dialogue (**TSD2012**)]	Brno, Czech Republic	September 3, 2012	---	---	Held
Presentation / Publication details & Comments	MU is the organiser of the workshop, with members of the PRESEMT consortium being in the program committee. Proceedings of the workshop were edited by Karel Pala, Aleš Horák, Pavel Rychlý and Petr Sojka from MU.						
Towards Retrieving and Ranking Clinical Recommendations with Cross-Lingual Random Indexing	Conference paper	CLEF 2012 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis (CLEFeHealth2012)	Rome, Italy	September 17-20, 2012	Hans Moen & Erwin Marsi	NTNU	Presented
Presentation / Publication details & Comments							
Evaluating the translation accuracy of a novel language-independent MT methodology	Conference paper	24th International Conference on Computational Linguistics [COLING 2012]	Mumbai, India	December 8-15, 2012	George Tambouratzis, Sokratis Sofianopoulos & Marina Vassiliou	ILSP	Presented
Presentation / Publication details & Comments	Presented in the main conference session by Sokratis Sofianopoulos						

5. Future work

This deliverable reports on the final period of the project. However, activity is still being carried out within the scope of PRESEMT. This involves both the analysis of the evaluation results as well as an error analysis of the MT output for the latest version of the PRESEMT prototype. These results will be used to determine if further improvements can be achieved in an economic manner. In addition, dissemination activities regarding the final results of the project are being prepared in the form of presentations in international conferences and specialist workshops as well as publications in refereed scientific journals. Besides, a further evaluation activity is planned in the near future. Finally, the project website has been improved and ported to the newer version of Joomla in order to continue serving the dissemination efforts of the project.

6. Further information

For further information and for keeping up-to-date regarding the PRESEMT system, please visit our website at www.presemt.eu.