Pattern REcognition-based Statistically Enhanced MT

ANNUAL PUBLIC REPORT 1

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 1**
Dissemination level	Public
Period covered	**2010**
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 (Pattern REcognition-based Statistically Enhanced MT) is an EU-funded project under the FP7 topic "ICT-2009.2.2: Language-based Interaction". It is 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 will address the issue of effectively managing multilingual content and is expected to suggest a language-independent machine-learning-based methodology.

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

Parallelisation of the main translation processes will be investigated in order to reach a fast, high-quality translation system. Furthermore, the optimisation and personalisation of the system parameters via automated processes (such as GAs or swarm intelligence) will be studied.

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

Key innovation

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

To this end, a flexible MT system will be developed, which will be enhanced with (a) **pattern recognition** techniques (such as extended clustering or 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** also 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 the **functional biological systems** for **disambiguating** translations
4. Extensive use 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 for substantial advances in **translation speed**

2. PRESEMT system description

The PRESEMT system, as envisaged and implemented up to this point, is depicted in Figure 1. It 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 perform, 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

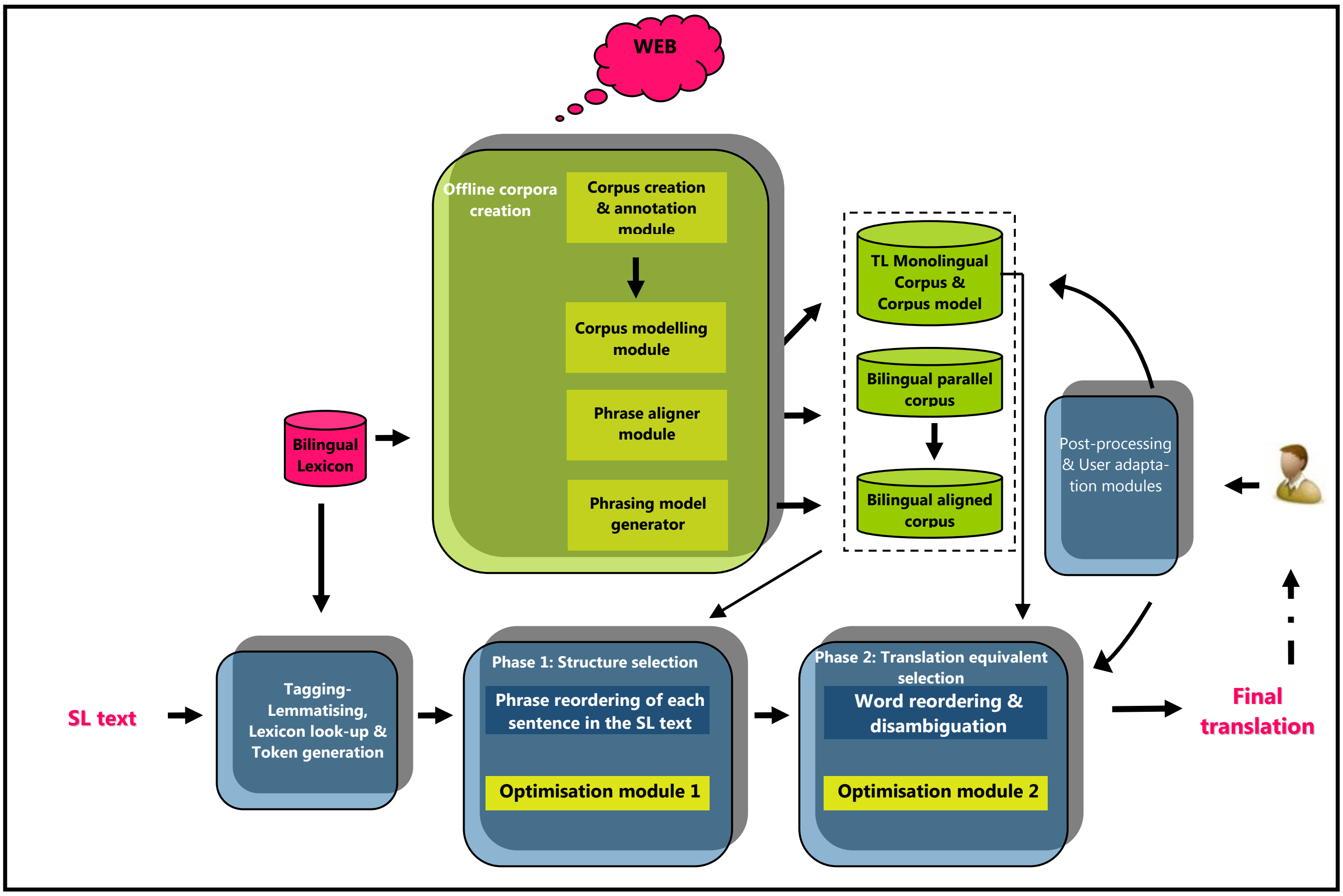

WEB
Offline corpora creation
Corpus creation & annotation module
Corpus modelling module
Phrase aligner module
Phrasing model generator
TL Monolingual Corpus & Corpus model
Bilingual parallel corpus
Bilingual aligned corpus
Bilingual Lexicon
Post-processing & User adapta-tion modules
SL text
Tagging-Lemmatising, Lexicon look-up & Token generation
Phase 1: Structure selection
Phrase reordering of each sentence in the SL text
Optimisation module 1
Phase 2: Translation equivalent selection
Word reordering & disambiguation
Optimisation module 2
Final translation

Table 1: PRESEMT basic system modules

Pre-processing stage: 4 modules	Main translation engine: 4 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 1	User adaptation module
Corpus modelling module	Optimisation module 2	

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 deduplication; bilingual corpora: corrections / modifications) and annotation (e.g. Part-of-Speech (PoS) tagging and lemmatisation).

The **Phrase aligner module**, 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 non-parsed member of the language pair is the source language, 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** 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 mode 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 (such as n-gram models over words and PoS tags, SOM for words, word space models, frequency-of-occurrence of specific lexical orders etc.). This type of information will then be utilised during the translation process. For developing the specific module a series of methodologies are currently being explored.

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 a 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 a monolingual corpus.

The two **Optimisation modules** are 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" itself 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 will be handled for the development of the first two versions of the system prototype, whereas the right-hand column lists the language pairs to be 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 1st year of the project

During the first year of the project two main phases were evident. The first one, corresponding to the first four months of the project, coincided with work package WP2, which involved two main tasks, (a) the elicitation of the system specifications and (b) the delineation of the future system validation and evaluation activities. The second phase started in month M5 and signalled the initiation of work packages related to (a) the design and implementation of PRESEMT modules, namely WP3, WP4, WP5 and WP6, and (b) the integration of the system modules into a single platform, i.e. WP7.

In the remainder of this section, the main results obtained and objectives achieved during the 1st year are summarised per work package.

WP2: System specifications

The WP2 objectives were to (a) define the system specifications and architecture and (b) set-up the system validation and evaluation processes.

To this end the system architecture has been established and the constituent modules identified (cf. Section 2 of the present document). It is noteworthy that the architecture is modular, so as to support the system flexibility. Additionally, a UML scheme has been created, which is intended to be used for the development of the PRESEMT prototype software.

The foundations for performing validation and evaluation of the system prototype have been set, by specifying the procedure for validating the system prototype as well as various scenarios of testing per system functionality. In particular, eight functionalities have been identified and the profile of the corresponding validators sketched.

Furthermore, the evaluation (both automatic and human) methodology to be followed with regard to the translation output has been outlined. In addition, the human evaluation parameters and the automatic evaluation metrics to be used, as well as methods for meta-evaluation have been identified.

WP3: Corpus extraction and processing algorithms

WP3 relates to the design and development of the Pre-processing modules.

Up to this point several text resources, monolingual and bilingual ones, have been collected over the web and annotated with PoS and lemma information. The monolingual corpora have been collected by crawling the web with the Heritrix crawler. The web corpora have then been processed with the ***jusText*** algorithm developed at MU (see Pomikálek (forthcoming[1])) for removing duplicate content and boilerplate, i.e. non-informative parts outside of the main content of a web page, typically machine-generated and repeated across the web pages of the same website. The bilingual corpora are substantially smaller in size (~200 sentence pairs each) and have been manually created by eliciting multilingual content from the web.

The following tables provide an overview of the text resources collected so far.

[1] Pomikálek J. (forthcoming). Removing Boilerplate and Duplicate Content from Web Corpora (PhD thesis). Masaryk University, Brno, Czech Republic.

Table 3: Statistical data on PRESEMT monolingual corpora (1st year of the project)

Language	English	Italian	German	Greek
Corpus name	enTenTen / BiWeC	itTenTen	deTenTen / BigDeWaC	
final size (*in tokens*)	3,658,726,327	3,076,812,674	2,874,779,294	
downloaded volume (*gzipped*)	129 GB	272 GB	291 GB	157 GB
downloaded volume uncompressed (*estimate*)	645 GB	1360 GB	1455 GB	785 GB
downloaded (*unique*) URLs	13,638,928	33,459,999	43,160,992	
original main content (*words*)	6,805,296,135	7,740,199,568		
original boilerplate content (*words*)	7,843,420,305	14,814,859,910		
docs (*after removing exact duplicates*)	3,357,252	5,335,839	8,237,310	
tokens (*after removing exact duplicates*)	4,765,119,530	5,017,409,779	4,880,335,291	
duplicate 10-grams (*after removing exact duplicates*)	212,864,389	295,369,351	331,258,395	
docs (*after removing duplicate text blocks*)	2,838,738	4,020,968	5,752,857	
tokens (*after removing duplicate text blocks*)	3,658,726,327	3,076,812,674	2,874,779,294	
duplicate 10-grams (*after removing duplicate text blocks*)	6,757,185	12,230,555	7,196,387	
	words = whitespace separated character strings (wc-w)			
	tokens = unitok.py (universal tokeniser) output			

Table 4: Statistical data on PRESEMT bilingual corpora (1st year of the project)

		Target Language		
		English	German	Italian
		corpora size (*in tokens*)		
Source Language	Czech	3,496	3,871	
	English			
	German	6,767		
	Greek	6,659	8,705	
	Norwegian	---[2]	---	

[2] Currently being developed

The first versions of the Phrase aligner module and the Phrasing model generator have been released.

The **Phrase aligner module** operates in a two-step mode, (a) performing alignment of SL words to TL ones, based on correspondences provided by a bilingual lexicon, and (b) grouping the unaligned SL words into phrases by taking into consideration grammatical features like number, case etc. The module has been tested so far on two bilingual corpora, Greek → English and German → English, with accuracy rates of 89.99% and 83.33% respectively.

For the **Phrasing model generator,** experiments involving subsets of the Greek → English parallel corpus were performed, using two different packages, namely *CRF*++ and *MALLET*.

Given that the **Corpus modelling module** will provide input to the 2nd phase of the translation process, development & test data sets and scoring tools of SEMEVAL 2010 have been examined for word translation disambiguation (WTD) purposes. Moreover, n-gram models have been created, which will be the baseline of the novel language modelling methods to be investigated in the project.

WP4: Structure selection & WP5: Translation equivalent selection

These two work packages involve the design and development of the two translation phases. The consortium is currently investigating two different algorithmic implementations in order to identify their comparative strengths and weaknesses.

With respect to the optimisation, experiments have been conducted regarding the optimisation of real-valued parameters used in the structure comparison phase of MT systems similar to PRESEMT, and specifically the parameters used in the METIS II MT system. For the optimisation of the parameters the SPEA2[3] multiobjective evolutionary algorithm is used.

WP6: Post-processing & User adaptation

Within WP6 an extensive study of the literature and of concepts for post editing has been initiated to make concrete the idea of having the user improve the system by post editing. Furthermore, three modes of adaptation have been identified, these being (a) adaptation through optimisation, (b) core system adaptation and (c) caching adaptation.

WP7: Integration

The main aim of WP7 is to integrate the modules developed into the technical work packages (i.e. WP3, WP4, WP5 and WP6) into one working prototype. Besides, fine-tuning of the proposed system will also take place, to make use of parallelisation opportunities towards the application of multi-processor/multi-core architectures.

To this end, an Apache Subversion revision control server has been set up, in order to keep a record of all versions of the implemented source code and documentation, enabling at the same time the sharing of the source code among partners. The Subversion server allows partners collaborating on a specific module to be able to instantly view the work of the others or contribute their own code, without resorting to emailing or uploading their source code on a site.

As regards parallelisation, interactions between algorithms, data structures and execution platforms have been and continue to be analysed in order to shed light to the best possible parallelisation strategy. Alternative implementation tools and programming languages have also been investigated.

[3] E. Zitzler, M. Laumanns, L. Thiele, "SPEA2: improving the strength Pareto evolutionary algorithm," Swiss Federal Institute of Technology (ETH). Zurich, Switzerland. Technical report TIK-Report 103, 2001

4. Dissemination activities

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

General dissemination activities

Activity name	Place	Date	Details / Comments
PRESEMT website	N/A	January 2010	Date of launch (the website is constantly updated)
Project logo	N/A	January 2010	It is available for download from the project website. It is also used in the cover page of all the project reports and deliverables, while it is planned to accompany any info material that will produced during the project lifecycle.
PRESEMT Facebook group	N/A	January 2010	
PRESEMT Fact sheet	N/A	February 2010	A first version has been compiled. It is planned to be regularly updated throughout the project lifecycle.
PRESEMT Project Presentation	N/A	March 2010	A first version has been compiled. It is planned to be regularly updated throughout the project lifecycle.
Concise presentation of PRESEMT	N/A	February 2010	This text has been posted under the webpage of EC's Unit E1 "Machine Translation & Language Technologies", following a respective request from the Unit
Large corpora gathered within PRESEMT, and related web services	N/A	November 2010	LCL and Masaryk University have been setting up the large corpora gathered within PRESEMT, and related web services, so that they may be used for showcasing PRESEMT and within an open access approach.

Publications

Publication title	Authors	Type	Publication details	Date	Status
Comparable Corpora Within and Across Languages, Word Frequency Lists and the KELLY Project	Adam Kilgarriff	Conference paper	Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010 (Valletta, Malta, May 22, 2010), pp. 1-5	2010	Published
Fast syntactic searching in very large corpora for many languages	Miloš Jakubíček, Adam Kilgarriff, Diana McCarthy, Pavel Rychlý	Conference paper	Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, PACLIC 24 (Tokyo, November 4, 2010), pp. 741-747	2010	Published
Evolutionary Algorithms in Natural Language Processing	Lars Bungum & Björn Gambäck	Conference paper	Proceedings of the 2nd Norwegian Artificial Intelligence Symposium (Gjøvik, Norway, November 22, 2010)	2010	In print
Studying the SPEA2 Algorithm for Optimising a Pattern-Recognition Based Machine Translation System	Sokratis Sofianopoulos & George Tambouratzis	Conference paper	IEEE Symposium on Computational Intelligence in Multicriteria Decision-Making (MCDM 2011), Paris, France, April 11 -15, 2011	N/A	Submitted *(Review results: January 2011)*

Conferences & workshops				
Event title	**Place**	**Date**	**Presentation title**	**Details / Comments**
7th Language Resources and Evaluation Conference (LREC 2010)	Valletta, Malta	May 19-21, 2010	N/A	Distribution of PRESEMT leaflets on site
3rd Workshop on Building and Using Comparable Corpora (LREC 2010)	Valletta, Malta	May 22, 2010	Comparable Corpora Within and Across Languages, Word Frequency Lists and the KELLY Project	Invited talk: Adam Kilgarriff (LCL) has presented work on comparable corpora, word lists and domains, closely tied to the ideas on domains to be further explored in PRESEMT.
ACL 2010	Uppsala, Sweden	July 15, 2010	N/A	Attended by NTNU members
ACL workshop on Companionable Dialogue Systems	Uppsala, Sweden	July 11-16 2010	N/A	Co-organised by NTNU
COLING 2010	Beijing, China	August 23-27, 2010	N/A	Attended by NTNU members
European Summer School in Logic, Language and Information (ESSLLI 2010)	Copenhagen, Denmark	August 9-20, 2010	N/A	Attended by NTNU members
2nd Norwegian Artificial Intelligence Symposium	Gjøvik, Norway	November 22, 2010	Evolutionary Algorithms in Natural Language Processing	Presentation given by Lars Bungum & Björn Gambäck
25th European Conference on Object-Oriented Programming (ECOOP 2011)	Lancaster, UK	July 25-29, 2011	The PRESEMT project for the Machine Translation Task	Invitation received from the ECOOP conference for a presentation of PRESEMT at the "**Research Project Symposium**" (http://ecoop11.comp.lancs.ac.uk/?q=calls/symposium) *(Submitted presentation proposal by ILSP; acceptance notification: March 1, 2011)*
15th Annual Conference of the European Association for Machine Translation (EAMT-2011)	Leuven, Belgium	May 30-31, 2011	N/A	This publication is expected to focus on work carried out on the Phrase aligner module. *(Planned submission by ILSP; submission deadline: February 18, 2011)*
ICWSM (Web and Social Media) and NAACL-HLT conferences	Washington DC & Los Angeles, USA	May & June 2010	N/A	While primarily for information-gathering purposes, opportunities were taken where they arose to describe and discuss the work being undertaken within PRESEMT. Adam Kilgarriff chaired WAC-6 (6th Web-as-Corpus workshop), a highly salient event, and gave a brief outline of PRESEMT there.
The Second Sketch Engine Workshop (SKEW-2)	Brighton, UK	March 16-17, 2011	N/A	Adam Kilgarriff is organising, or involved in the organisation of event. Event includes PRESEMT presentations.
Research Models in Translation Studies II	Manchester, UK	April 29-May 2, 2011	Rethinking Corpus-based Translation Studies in the Web Era (*title of a panel discussion*)	Adam Kilgarriff & Silvia Bernardini placed a proposal for a panel discussion at the conference, which has now been accepted. Kilgarriff will contribute with a paper describing PRESEMT work.

Conferences & workshops

Event title	Place	Date	Presentation title	Details / Comments
Annual Lexicon Summer School in Lexicography and Lexical Computing	St Petersburg State University, Russia	June 14-19, 2011	N/A	Adam Kilgarriff is organising, or involved in the organisation of event. Event includes PRESEMT presentations.
Linguistics Institute 2001	University of Colorado, USA	July 7-August 2, 2011		Lexical Computing Ltd. will sponsor the Linguistics Institute, a prestigious annual four-week summer school organised by the Linguistics Society of America. It is typically attended by around 600 academics and graduate students This year, "*the Institute focus will be on interdisciplinary, empirically based approaches to language*", which fits well with PRESEMT methods as well as LCL expertise. Adam Kilgarriff will be giving several lectures, to include and showcase aspects of PRESEMT work.
The Second e-lexicography conference	Bled, Slovenia	November 10-12, 2011	N/A	Adam Kilgarriff is organising, or involved in the organisation of event. Event includes PRESEMT presentations.

Info days / Exhibitions / Other events

Event title	Place	Date	Presentation title	Details / Comments
Mid Sweden University seminar between Östersund (Sweden) and Trondheim(Norway)	Östersund, Sweden	January 28, 2010	Presentation of the PRESEMT project	Presentation in the session "Research status and plans overview" of the meeting held on January 27-29, 2010
Speech and Language Technology at NTNU Day	Trondheim, Norway	September 17, 2010	Domain adaptation in Machine Translation	Presentation of the work of NTNU in PRESEMT – Presented by Lars Bungum
Speech and Language Technology at NTNU Day	Trondheim, Norway	September 17, 2010	Word translation disambiguation without parallel text	Presentation of the work of NTNU in PRESEMT – Presented by Erwin Marsi
Speech and Language Technology at NTNU Day	Trondheim, Norway	September 17, 2010	Overview of Language Technology related activities at IDI	Presentation of the work of NTNU in PRESEMT – Presented by Björn Gambäck
ICT 2010: Digitally driven	Brussels, Belgium	September 27-28, 2010	PRESEMT demo	Exhibition proposal (*rejected*)

Teaching

Title	Type	Lecturer	Place	Date
Machine Translation, Natural Language Interfaces *(NTNU)*	Lecture	Björn Gambäck (NTNU)	Trondheim, Norway	March 2010 *(spring semester 2010)*
Moderne Übersetzungswerkzeuge und Fachkommunikation *(University of Saarland)*	Lecture	Paul Schmidt (GFAI)	Saarbrücken, Germany	April 2010 – July 2010 *(summer semester 2010)*

Teaching				
Title	**Type**	**Lecturer**	**Place**	**Date**
Ausgewählte Themen der maschinellen Übersetzung und der Fachkommunikation *(University of Saarland)*	Seminar	Paul Schmidt (GFAI)	Saarbrücken, Germany	April 2010 – July 2010 *(summer semester 2010)*
Moderne Übersetzungswerkzeuge und Fachkommunikation *(University of Saarland)*	Lecture	Paul Schmidt (GFAI)	Saarbrücken, Germany	October 2010 – February 2011 *(winter semester 2010)*
Ausgewählte Themen der maschinellen Übersetzung und der Fachkommunikation *(University of Saarland)*	Seminar	Paul Schmidt (GFAI)	Saarbrücken, Germany	October 2010 – February 2011 *(winter semester 2010)*
Machine Translation, Natural Language Interfaces *(NTNU)*	Lecture	Björn Gambäck (NTNU)	Trondheim, Norway	March 2011 *(spring semester 2011)*

5. Future work

Within the next reporting period, work in the PRESEMT project will focus to a large extent on progressing in terms of the research and evaluation work.

More specifically, within the second year of the project, the two translation phases (structure selection and translation equivalent selection) will be studied extensively, building on the work carried out up to now. In addition, the optimisation of parameters of these two phases will be performed, to determine their optimal values. This research is planned to take place while involving multiple language pairs, in order to reach sound conclusions regarding the effectiveness of the proposed approach.

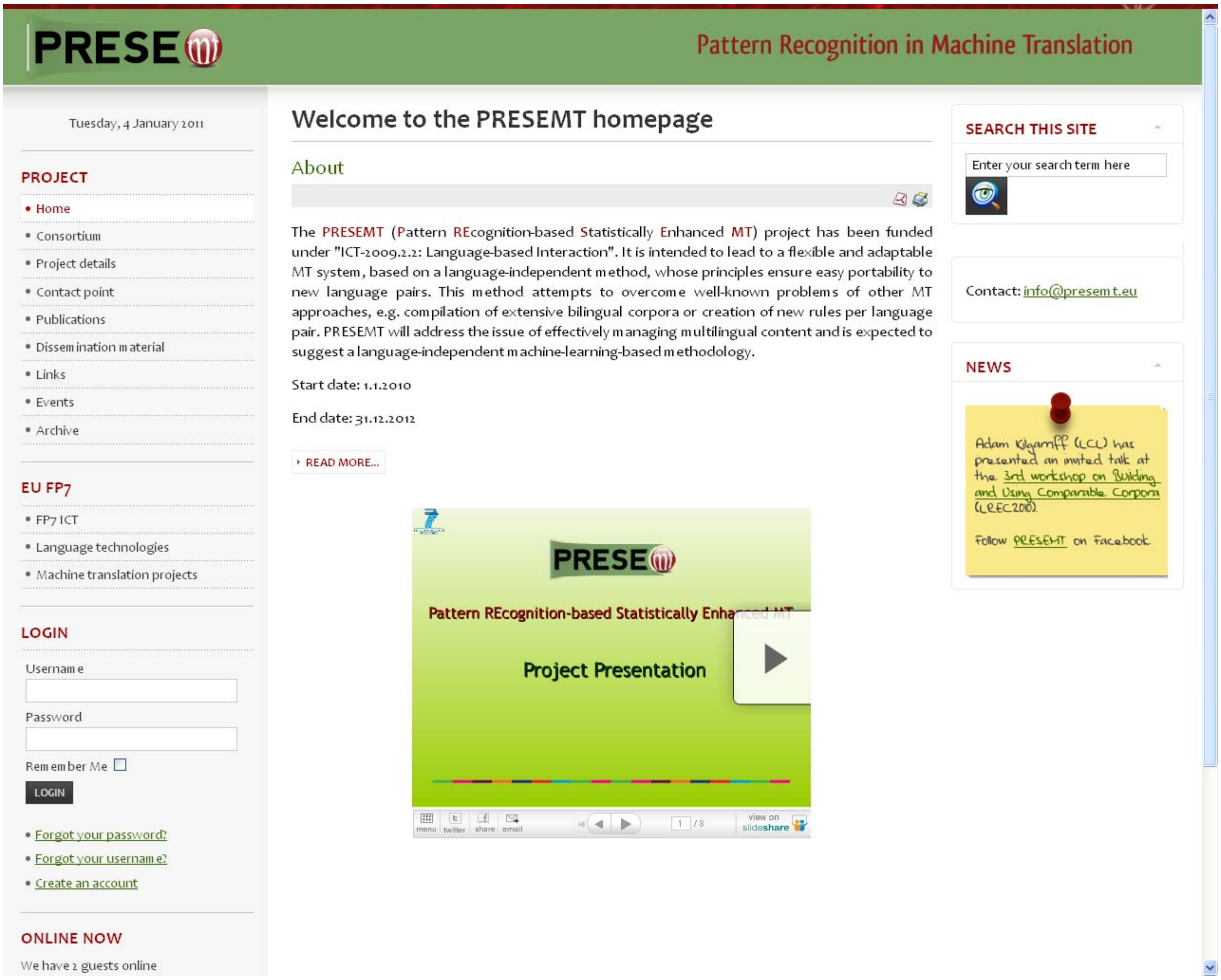
In addition, the pre-processing algorithms will be studied in detail, together with the post-processing and user adaptation algorithms.

As the modules are prepared and tested, the integration of the PRESEMT prototype will become a major activity. A first prototype, coupled with documentation and user manual, is to be released and evaluated within the second year of the project. To that end, the first user groups will be set up within the forthcoming period.

Furthermore, dissemination activities will be continued and intensified, in order to publicise the results of the project to the wider academic community as well as the prospective user groups. These activities will also be communicated via the PRESEMT website, in order to reach the relevant audience.

6. Further information

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



Tuesday, 4 January 2011

PROJECT

- Home
- Consortium
- Project details
- Contact point
- Publications
- Dissemination material
- Links
- Events
- Archive

EU FP7

- FP7 ICT
- Language technologies
- Machine translation projects

LOGIN

Username

Password

Remember Me □

LOGIN

- Forgot your password?
- Forgot your username?
- Create an account

ONLINE NOW

We have 2 guests online

Welcome to the PRESEMT homepage

About

The PRESEMT (Pattern REcognition-based Statistically Enhanced MT) project has been funded under "ICT-2009.2.2: Language-based Interaction". It is intended to lead to a flexible and adaptable 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 will address the issue of effectively managing multilingual content and is expected to suggest a language-independent machine-learning-based methodology.

Start date: 1.1.2010

End date: 31.12.2012

READ MORE...

SEARCH THIS SITE

Enter your search term here

Contact: info@presemt.eu

NEWS

Adam Kilgarriff (LCL) has presented an invited talk at the 3rd workshop on Building and Using Comparable Corpora (LREC2010).

Follow PRESEMT on Facebook.