PRESEMT project: Publishable Summary

Project objectives

The objective of the PRESEMT project is to develop a flexible and adaptable MT system, based on a language-independent method, which is easily portable to new language pairs. This method attempts to overcome well-known problems of other MT approaches, e.g. bilingual corpora compilation 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.

The key aspects of PRESEMT involve syntactic phrase-based modelling, pattern recognition techniques towards the development of a language-independent analysis and evolutionary algorithms for system optimisation. PRESEMT is intended to be of a hybrid nature, combining linguistic processing with the positive aspects of corpus-based approaches, such as SMT and EBMT. 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 an automatic and language-independent process of morphological and syntactic analysis, removing the need for 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 genetic algorithms or swarm intelligence) will be studied. To allow for user adaptability, all the corpora used in PRESEMT will be retrieved from web-based sources. User feedback will be integrated through the use of appropriate interactive interfaces. PRESEMT is expected to be easily customisable to both new language pairs and specific sublanguages.

Work performed since the project start

The work performed within the 1st year of the project relates primarily to the PRESEMT system design and implementation. In particular, the PRESEMT **system specifications** have been finalised. This involved defining the constituent modules, their mode(s) of operation, their input/output and their interconnection. This work falls within WP2.

Figure 1 illustrates the architecture of the PRESEMT system, where three processing stages have been identified, (a) the **Pre-processing stage**, where the resources needed for the translation process are created, (b) the **Main translation engine**, involving the translation process per se and its optimisation and (c) the **Post-processing stage**, which allows users to modify the system output.

Besides, a first version of the majority of the **system modules** (WP3 – WP6) has been designed and implemented. The remaining modules are under development.

The consortium has also collected NLP **tools** as well as various **text resources** (monolingual and bilingual corpora) needed for the project **(**WP3).

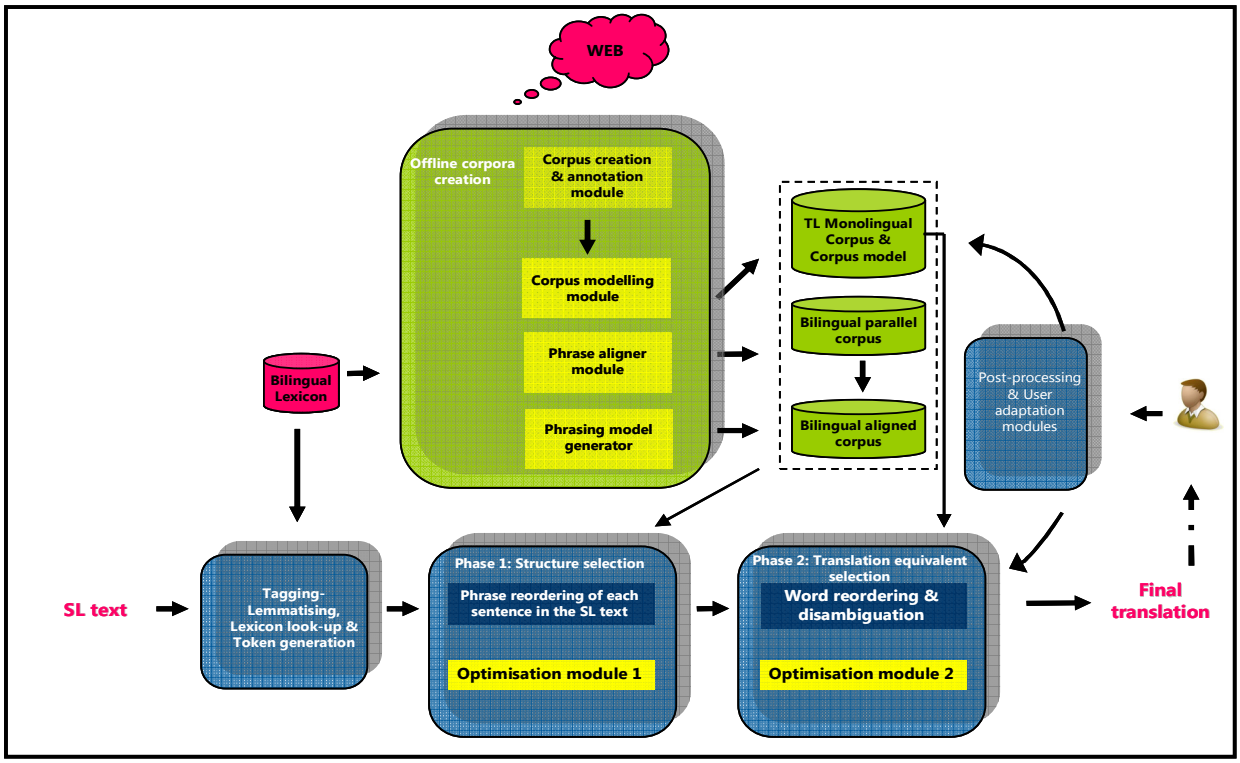
For **integration** purposes a platform has been set up, while some first paper experiments have been carried out in order to investigate the issue of **parallelisation** (WP7).

Furthermore, the consortium has delineated the processes and the actors for **validating** the system's performance as a whole and per module. A framework for **evaluating** the quality of the translation output has also been established. Both validation and evaluation activities will be carried out within WP9, which is yet to begin. The aforementioned work falls within WP2.

It should also be noted that a number of **dissemination** activities (see WP8) have been carried out, these including the launch of the project website, the preparation of dedicated factsheets and presentations and the participation in relevant conferences to inform the community of the PRESEMT project.

Regarding the **administrative** aspect of the project, the collaboration within the consortium has been unproblematic. A series of bilateral and project-level meetings facilitating the project progress have taken place, while all due deliverables and reports have been submitted on time to the European Commission.

Figure 1: Architecture of the PRESEMT system



Main results achieved so far

The main results of the project after the completion of the 1st year can be summarised as follows:

* The PRESEMT system **architecture** and constituent **modules** have been defined (see Figure 1).
* A **UML scheme** has been created to support the development of the PRESEMT prototype.
* The **validation** and **evaluation** activities to be carried out after the first half of the project have been specified.
* A first version of various system modules has been released, namely the **Corpus creation & annotation module**, the **Phrase aligner module**, the **Phrasing model generator**, the **Corpus modelling module** and the **Structure selection module**.
* The main system platform has been created.
* The project website has been launched.
* The second version of the **dissemination & exploitation plan** has been produced.

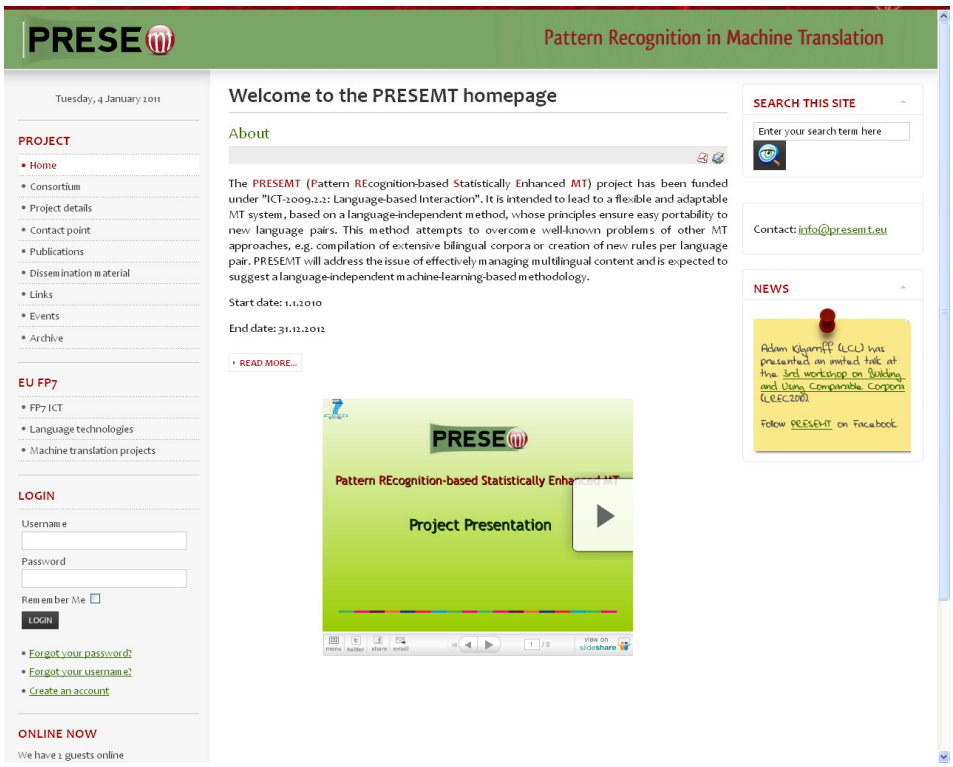
Expected results and potential impact

The project impact and results are expected to cover the research/scientific community as well as the general public. The MT approach being developed within the project has a number of innovative aspects, ranging from the method for implementing the division of sentences into phrases to the way of optimally utilising the linguistic information in the monolingual and bilingual resources. These aspects are project-specific, and thus can be expected to contribute as a whole to the state-of-the-art of the MT area. In this respect, the design of the prototype so that it allows the rapid development of new translation systems for different language pairs is of prime importance. The project results are envisaged to follow two main directions, (a) the dissemination of the MT system design as a whole and (b) the release of a prototype via the web for use by the academic community and the general public. The latter involves the release of specific modules individually, via dissemination activities, most likely following strategies defined in the exploitation plan. Hence, it is expected that a number of advances may be achieved, relating to the specific field of machine translation (entire system) as well as to other scientific areas, where mainly isolated PRESEMT modules are studied, the respective areas including general computational linguistics, pattern recognition and parallel processing, to name but a few.

Regarding the MT field, the impact is expected to be substantial. As noted by one of the PRESEMT peer reviewers, "the project is ambitious and its impact on future MT, both research and practice, could well be highly significant". Also, as noted by another peer reviewer, the aspect of meta-evaluation of the system is important as "very useful information could come out". The prime impact is of course expected to be to the machine translation community, which could filter through to the public both in terms of the final PRESEMT prototype as well as future MT systems, which may well be influenced by PRESEMT. It is hoped that advances in both the translation quality, the ease of development of systems for new language pairs and the speed of MT processing will all benefit from the project. Similar areas of impact are expected in the other research and application areas mentioned earlier, though these impacts are less evident at this point and will become more concrete as the project progresses further.

PRESEMT website

For further information and for keeping up-to-date regarding the PRESEMT project, please visit our website at **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 Informationsforschung 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