



PRESEMT: Pattern REcognition-based Statistically Enhanced MT

EC Seventh Framework Programme (FP7/2007-2013)
ICT-2009.2.2: Language-based Interaction
Small or medium-scale focused research project (STREP)
ICT-248307
<http://www.presemt.eu>

List of partners	
	Institute for Language and Speech Processing/R.C. "Athena", Greece (coordinator)
	Gesellschaft zur Förderung der Angewandten Informationsforschung e.V., Germany
	Norges Teknisk-Naturvitenskapelige Universitet, Norway
	Institute of Communication and Computer Systems, Greece
	Masaryk University, Czech Republic
	Lexical Computing Ltd, United Kingdom

Project duration: January 2010 — December 2012

Summary

PRESEMT proposes a novel approach to MT by introducing cross-disciplinary techniques, mainly from the **machine learning** and **computational intelligence** domains. It aims to create a flexible, adaptable MT system, based on language-independent principles to ensure easy portability to new language pairs. This method attempts to overcome well-known problems of MT, such as compilation of extensive bilingual corpora or creation of specific rules per language pair. The translation context is modelled on phrases, produced via a semi-automatic, language-independent process of morphological and syntactic analysis, alleviating the need of compatible NLP tools per language pair, and providing an alternative to n-gram models. Parallelisation of the translation process is aimed to lead to a fast, high-quality translation system.

The MT system to be developed will be enhanced with (a) **pattern recognition** techniques towards a language-independent analysis and (b) **evolutionary computation** methods for system optimisation. To allow for user adaptability, all corpora will be retrieved from web-based sources, while user feedback will be integrated through appropriate interactive interfaces.

During the first year, PRESEMT focussed on the development of the modules required for the MT system. Currently, these modules are being integrated to create the first prototype which is expected to be completed by September 2011 and thereafter form the basis for evaluation activities over multiple language pairs.