



PANACEA (Platform for Automatic, Normalised Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies)

Funding agency: European Commission

Funding call identification: FP7-ICT

Type of project: STREP

Project ID number: 248064

<http://www.panacea-lr.eu>

List of partners
Universitat Pompeu Fabra, Spain (coordinator)
CNR – ILC, Italy
ILSP – R.C. “Athena”, Greece
University of Cambridge, United Kingdom
Linguattec, Germany
Dublin City University, Ireland
ELDA, France

Project duration: January 2010 — December 2012

Summary

A strategic challenge for Europe in today's globalised economy is to overcome language barriers through technological means. In particular, machine translation (MT) systems are expected to have a significant impact on the management of multilingualism in Europe.

PANACEA addresses the most critical aspect for MT: the so-called language resource (LR) bottleneck. Although MT technologies may consist of language-independent engines, they depend on the availability of language-dependent knowledge for their real-life implementation, i.e., they require LRs.

The objective of PANACEA is to build a factory of LRs that automates the stages involved in the acquisition, production, updating and maintenance of LRs required by MT systems and by other applications based on language technologies, and simplifies eventual issues regarding intellectual property rights. This automation will cut down the cost, time and human effort significantly. These reductions of costs and time are the only way to guarantee the continuous supply of LRs that MT and other language technologies will be demanding in the multilingual Europe.

In its first year, PANACEA has developed the first version of the LR factory, which allows the final user to acquire LRs by designing workflows (acquisition pipelines) that connect web-services provided by the different partners. For example, to create a parallel corpus for a given domain, the user can call the multilingual crawler (provided by ILSP) with a set of keywords of the domain, and connect its output to any of the sentential aligners (provided by DCU).