

PANACEA

Marc Poch
Universitat Pompeu Fabra

EAMT

27th - 28th of May 2010 - Saint-Raphaël, France



**PANACEA's objective is
to join together a number
of advanced interoperable tools
to build a
factory of Language Resources**



A production line that automates the stages involved in the acquisition, production, updating and maintenance of the LR required by MT and other Language Technologies.



**Cost and time reduction by automation
is the only way to ensure
the continuous supply of LR's
that can guarantee a LT industry
covering all languages, all domains,
for current and future needs, and
in the time required by the market.**

The factory is build as a Web Service-based platform for easy integration of the latest technological components for:

- **Monolingual and Parallel Text Acquisition and Pre-processing**
- **Sentential and sub-sentential alignment**
- **Bilingual Dictionary and Transfer Grammar production**
- **Lexical Information Acquisition for rich information dictionary production.**

Project results (1/3)

1. The platform, as a virtual, distributed, production line where different interoperable components can be chained in particular workflows to produce different types of LR's, for different languages.
 - **Interoperability space**
 - A dedicated **Panacea Registry**, for the location, searching and documentation of Panacea components.
 - Dedicated **Panacea workflow editor** for defining different production chains.

Project results (2/3)

2. The automatic acquisition and production components:

- **Corpus Acquisition Component**
- **Corpus clean-up and Normalization Component**
- **Text Processing Components for sentence splitting, PoS Tagging, lemmatization, chunking and NER**
- **Sentential and subsentential aligners**
- **Bilingual dictionary extractor**
- **Transfer grammar extractor**
- **Lexical information Induction component**
- **Lexical classifiers**
- **Dictionary merger**

Project results (3/3)








3. LR's used as test and proof of the proper functioning of the factory.

- **Parallel texts**, cleaned and prepared for training-building translational models.
- **Large monolingual corpus**, PoS tagged and lemmatized for training and modelling language data,
- **Monolingual lexica** with morpho-syntactic, syntactic and lexical-class semantic information
- **Bilingual dictionary and transfer grammar**

Evaluation

**PANACEA's contribution & impact
will be demonstrated with a significant
time and cost reduction
in producing LR's.**

**A real life use case will be used to
measure the achievements**

| | | |
|---|-----------|---|
| <p>Prof. Núria Bel Universitat Pompeu Fabra - UPF</p> | <p>ES</p> |  |
| <p>Dr. Nicoletta Calzolari Consiglio Nazionale delle Ricerche - Istituto de Linguistica Computazionale – ILC</p> | <p>IT</p> |  |
| <p>Dr. Stelios Piperidis Institute for Language & Speech Processing ILSP</p> | <p>GR</p> |  |
| <p>Dr. Anna Korhonen University of Cambridge – UCAM</p> | <p>UK</p> |  |
| <p>Dr. Gregor Thurmair Linguattec -- LT</p> | <p>DE</p> |  |
| <p>Prof. Andy Way Dublin City University -- DCU</p> | <p>IR</p> |  |
| <p>Dr. Khalid Choukri Evaluations and Language Resources Distribution Agency -- ELDA</p> | <p>FR</p> |  |

Summary

**PANACEA goal is to build
a Language Resource factory
that will ensure the supply that Language
Technology industry needs to break through
problems such as Machine Translation
systems covering all languages, all domains,
for current and future needs, and in the time
required by the market.**



Keep informed at

www.panacea-lr.eu

Thanks!