SEVENTH FRAMEWORK PROGRAMME

PANACEA

SEVENTH FRAMEWORK PROGRAMME
THEME 3
Information and communication Technologies

PANACEA Project

Grant Agreement no.: 248064

Platform for **A**utomatic, **N**ormalized **A**nnotation and
Cost-**E**ffective **A**cquisition
of Language Resources for Human Language Technologies

2011 Annual Report

Dissemination Level:	Public
Delivery Date:	November, 2011
Status – Version:	
Author(s) and Affiliation:	UPF – Núria Bel

Table of Contents

1 Project Objectives & Challenges

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, making it possible to translate the huge quantity of (written or oral) data produced, and thus, covering the needs of hundreds of millions of citizens. PANACEA is addressing the most critical aspect for MT: the, so-called, language-resource bottleneck. Although MT technologies may consist of language independent engines, they highly depend on the availability of language-dependent knowledge for their real-life implementation, i.e., they require Language Resources (LR). In order to equip MT for every pair of European languages, for every domain, and for every text genre, appropriate language resources covering all these languages, domains and genres must be found, processed and supplied to MT developers. These should be provided in the format and with the information demanded by their systems. At present, this is mostly done by hand. Moreover, a Language Resource for a given language can never be considered complete or final because of the characteristics of natural language: language changes and new knowledge domains and new language varieties emerge at rapid pace. What is needed is an automatic system for compiling, producing and validating LR; a system conceived as integrated machinery for the production of LR.

The objective of PANACEA is to build a factory of LR that automates the stages involved in the acquisition, production, updating and maintenance of LR required by MT systems, and by other applications based on Language Technologies. This automation will cut down the costs, in terms of time and human effort, significantly. These reductions of cost and development time are the only way to guarantee a continuous supply of LR that Machine Translation and other Language Technologies may demand in a multilingual Europe.

In order to address this objective, PANACEA is working in the following areas:

1) The creation of a platform, which will be designed as a dedicated workflow manager, for the composition of a number of processes for LR production based on combinations of different web services.

2) The automatic production of massive amounts of LR for MT and other Language Technologies by the use of advanced components for the acquisition and normalization of corpora, monolingual and parallel corpora, the alignment of parallel corpora; the derivation of bilingual dictionaries out of subsentential aligned corpora; and the production of monolingual rich information lexica using corpus based automatic methods.

3) The evaluation of the platform and the LR production chain within the framework of both R&D and industrial settings.

PANACEA's contribution & impact will be demonstrated with a significant reduction in time and cost when producing LR's. A real life use case will be used to measure the achievements.

PANACEA WP 8 has defined a specific use case for evaluation, which is the adaptation of a MT system to a specific / specialized domain. This is a very complex use case; however it does not cover all of the PANACEA tools nor all of the PANACEA languages. In turn, it has practical relevance, as the production of MT systems is one of the major industrial applications of Language Technologies.

2 Summary of activities

Right after the kick-off of PANACEA, a six-month period of analysis began in order to assess and decide technologies to be used in the project and, also, how they will be used. The following challenges were discussed at the first technical meeting, which was held in Athens in April 2010.

- PANACEA must be convincing about the industrial use of available acquisition technologies by introducing ready to use tools as web services, with confidence indicators and which give priority to high precision and show objective measures of current capacities and cost reduction.
- PANACEA must invest in research for improving accuracy in automatic LR acquisition and production technologies

PANACEA reached the first Delivery Milestone in June 2010 with precise programs of work and objectives for the rest of the project for facing these questions. These reports which were the first delivery of all technical work packages can be found at PANACEA web site: http://www.panacea-lr.eu//en/deliverables/list

Immediately after, the consortium started the 1st cycle development which was mainly characterized by the deployment of different web services related to monolingual and parallel text crawling, pre-processing and annotation of texts in 6 different languages (EN, ES, IT, DE, EL and FR), as well as alignment of parallel texts. The web services were deployed according to PANACEA's platform specifications, that is, according to common interfaces (CI) for the same operations. In order to allow the chaining of different web services into workflows, web services integrated have used specific converters for accepting and producing (input/output texts) the formats agreed as standards for travelling objects, i.e., resources travel from web service to web service. The results of this first cycle were delivered in February 2011 and were evaluated by two external reviewers in March 2011 who considered the progress of the project as "very good". As date of today, users can find more than 80 different services in the PANACEA registry.

The second cycle of development started in March 2011 and has ended in October 2011. This cycle was mainly characterized by the validation that the resources produced by the PANACEA services could indeed feed MT applications. Alignment workflows were implemented as to test that from crawling to alignment produced LR's that could be run in a SMT environment (see the tutorial video "How to build a workflow from scratch" at http://www.vimeo.com/24790416).

Figure 1: Shared workflow (under CC license) for bilingual sentence alignment for crawled data EN and EL in myPanaceaExperiment

3 The PANACEA platform

PANACEA's platform is an interoperability space that joins together advanced interoperable tools to build a factory of LR. These tools will be offered as web services that, with a well defined functionality, will carry out particular tasks. By the selection of the appropriate web services, the user will be able to chain different production lines that **automate** the stages involved in the acquisition, production, updating and maintenance of the LRs as required by MT and other Language Technologies.

Figure 2: myPanaceaExperiment the Social Platform to share NLP workflows

4 PANACEA use scenarios

The main users of the PANACEA platform are assumed to be linguistically trained persons who collect/need new resources to improve or extend their linguistic applications (new domains, new languages etc.). Such people need to have skills in computational linguistics and some programming experience, but they don't need to be 'hard-core' programmers.The factory offers them 'services', that is, ready to be used components which do not need to be installed locally. Each service performs a precisely-defined operation of which the user needs only to know the input and output format and restrictions. Web services are announced and described in a Registry that will allow for searching and accessing them.

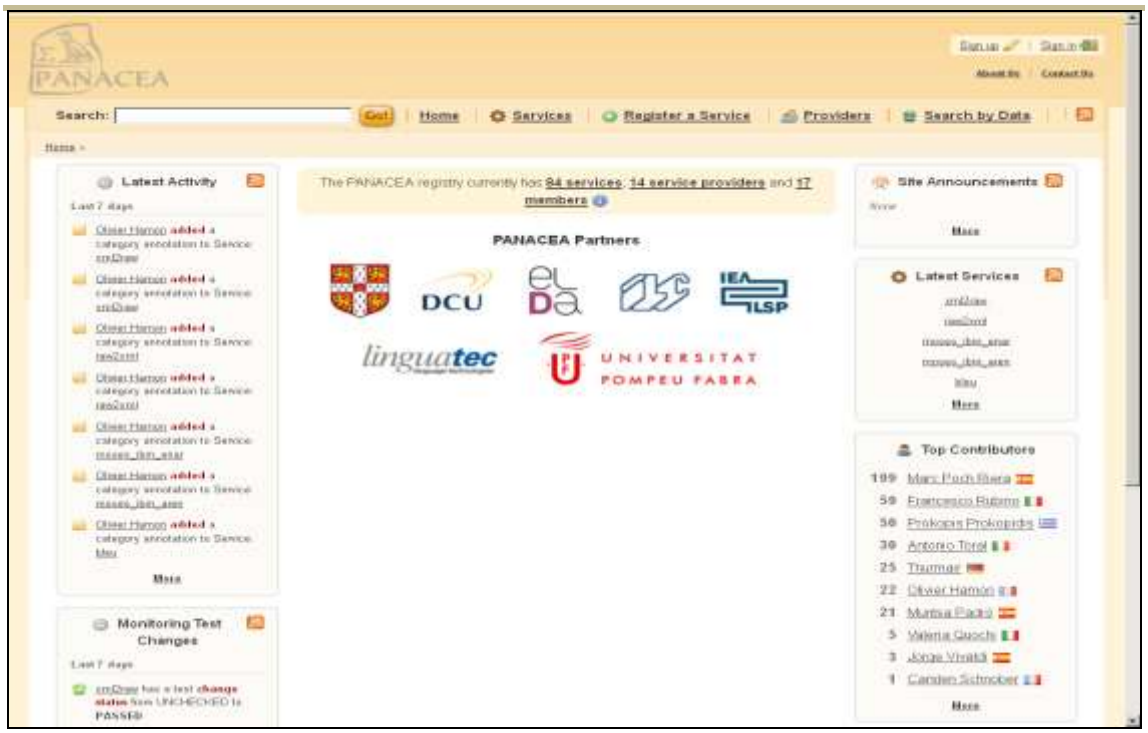

Figure 3: Registry to search and access PANACEA web services

In its first delivery, WP8 "Analysis of Industrial User Requirements" identified the typical use cases and operations that PANACEA web services will cover and that will include the following:

Corpus Tasks

- Build a corpus by web crawling
- Process a corpus by different services: sentence-segment it, tokenize / lemmatize / tag it
- Align two parallel texts: on document level, on paragraph level, on sentence level

Dictionary tasks

- Input a corpus for dictionary extraction (general purpose or domain specific)
- Submit a corpus for dictionary gap identification
- Acquire corpora for new / unknown words
- Merge corpus-extracted information (at entry level) possibly with existing computational dictionaries.

Extraction tasks

- Send a corpus to extract information items (named entities, or just key terms)

- Build an "Alerting System" (do texts match the alerting profile?) by intercalating a detecting dictionary gaps service
- Construct a workflow for "Topic Assignment" by using services for keyword extraction and training a classifier with pre-annotated data.

Translation Tasks

- Use a crawling system to collect / add corpus data for SMT creation
- Send a corpus to create a Language Model, for specific language, and / or for specific domain
- Send a parallel or aligned corpus to create your Translation Model (new language direction, new specific domain)
- Create / Adapt an (R)MT dictionary [with translations, with linguistic annotations (monolingual, transfer)]

5 Dissemination, Promotion and Awareness

The PANACEA user's workshop took place in Budapest on June 29, as a satellite event of the META-FORUM 2011.

Figure 4: User's workshop web page

The workshop proved interesting and new collaboration possibilities emerged as an end result. Participants expressed their interest in using such a platform, mostly as web services to apply to their work, but also as providers of the services themselves. As a matter of fact, some industrials communicated their intention to consider the sharing of some of their tools and data, both for

the project and for a wider audience. This had to be followed up at a later stage as participants had to define their strategy internally, at their organisations. In this regard, we can also say that this 1st users' workshop has helped to open a communication channel with the community. One of the challenges has been attracting the industrial audience, which is not a simple task for R&D projects. The fact of combining researchers and industry within PANACEA's Consortium itself has certainly been a plus to achieve.

Figure 5: Opening of the User's workshop

For further details on the presentations that took place during the workshop, the reader can refer to the workshop's programme page: http://workshops.elda.org/panacea2011/Programme, and to PANACEA web site where they are also available: http://www.panacea-lr.eu/en/news/project/2011/07/11/panacea-workshop-2011-held-on-29-june/

The following dissemination activities, mostly focusing on the engagement of target users have also been carried out so far.

As planned, once the first version of the platform was available, tutorial and explanatory videos were uploaded to the web site and a dissemination campaign inviting potential users to use and try the tools started by the end of March. Specifically, 20 selected users with different profiles were contacted and invited to visit the web site. A special mailing was prepared for more than 80 participants in the conference "Localization World" (http://www.localizationworld.com/lwbar2011/about.php) that took place in Barcelona, from 14 to 16 of June, targeting translators, terminologists and translation agency managers. Information about the PANACEA platform was also redistributed also via the Corpora List.

A total of 6 tutorial videos are accessible at http://panacea-lr.eu/en/info-for-professionals/tutorials/ regarding several introductory scenarios. At present, the following documents and tutorials are provided to the user:

· Documentation Index: PANACEA-Platform_documentation_index_v2.0.pdf
· General PANACEA tutorial: PANACEA-tutorial_v2.0.pdf
· Specific PANACEA tutorials: PANACEA-Soaplab-tutorial_v2.0.pdf and PANACEATaverna-tutorial_v2.0.pdf

Together with the following videos:

· PANACEA Building a workflow from scratch
· PANACEA Find and run a workflow
· PANACEA Registry (Learn to find web services at the PANACEA Registry)
· PANACEA myExperiment (Share your workflows with PANACEA myExperiment)
· PANACEA Part of Speech Tagging (Running a Part of Speech Tagger)
· PANACEA Bilingual Crawler (Running a Bilingual Crawler)

PANACEA has participated in the following events (2011). Information about previous activities can be found at PANACEA web site:

1. PANACEA was presented at the 15th Annual Conference of the European Association for Machine Translation which took place on 30-31 May, 2011 in Leuven, Belgium. Presentation, made by DCU, addressed specifically MT developers related to PANACEA developments.
2. A PANACEA poster was present at the ITC Industry Day, organized by the Spanish agency Centro de Desarrollo Tecnológico Industrial (CDTI) and held in UPF premises 15 June 2011.
3. PANACEA was present, with a poster and a demonstration stand, at META-FORUM 2011, which is the annual edition of the META-FORUM conference series, organized by META-NET. META-FORUM 2011, was held in Budapest on 27/28 June 2011.
4. The PANACEA platform has been presented at the Workshop on Language Resources, Technology and Services in the Sharing Paradigm – November 12, 2011 at IJCNLP 2011 (Chiang Mai, Thailand).
5. PANACEA was presented at the Localisation Innovation Showcase November 2011, 16th November 2011

Also, as envisaged in the Dissemination Plan, PANACEA has started to establish a liaison with other EU Projects with the objective of taking maximum advantage of collaboration between projects. This liaison has been developed via co-location of meetings/workshops, as the LREC workshop organized in LREC2010, a web cross-linking activity and direct communication between working groups. The following are the names of the projects relevant to PANACEA which has been contacted: META-NET, with which we have signed a collaboration agreement, TTC, ACCURAT, KYOTO, MONET, MEDAR and BOLOGNA.

PANACEA Papers in international conferences (2011)

The following is a list of PANACEA related papers from PANACEA partners that have been accepted in major scientific conferences in the period of reference. Information about previous publications can be found at PANACEA web site: http://panacea-lr.eu/en/project/publications/

1. Bel, N., Padró, M. and Necsulescu, S. **A Method Towards the Fully Automatic Merging of Lexical Resources.** this article will be presented on the Workshop on Language Resources, Technology and Services in the Sharing Paradigm – November 12, 2011 at IJCNLP 2011 (Chiang Mai, Thailand).

2. Mastropavlos, Nikos; Papavassiliou, Vassilis. (2011). **Automatic Acquisition of Bilingual Language Resources.** *Proceedings of the 10th International Conference on Greek Linguistics*. Komotini, Greece: 1-4 September 2011.

3. Necsulescu, S.; Bel, N.; Padró, M.; Marimon, M.; Revilla, E. (2011). **Towards the Automatic Merging of Language Resources, First international Workshop on Lexical Resources**. Woler 2011. Ljubljana, Slovenia: 1-5 August 2011.

4. Padró, M.; Bel, N.; Necsulescu, S. (2011) **Towards the Automatic Merging of Lexical Resources** in Angelova, G; Bontcheva, K., Mitkov, R., Kikolov, N. Proceedings of the International Conference Recent Advances in Natural Language Processing, Hisar, Bulgaria, ISSN 1313-8502

5. Pecina, Pavel; Toral, Antonio; Way, Andy; Papavassiliou, Vassilis; Prokopidis, Prokopis; Giagkou, Maria . (2011). **Towards using web-crawled data for domain adaptation in statistical machine translation.** In Forcada, Mikel L.; Depraetere, Heidi and Vandeghinste (Eds.) *Proceedings of the 15th conference of the European Association for Machine Translation (EAMT 2011)*. Leuven, Belgium: 30-31 May 2011, pp.297-304.

6. Poch, M.; Bel, N. **"Interoperability and technology for a language resources factory"** this article will be presented on the Workshop on Language Resources, Technology and Services in the Sharing Paradigm – November 12, 2011 at IJCNLP 2011 (Chiang Mai, Thailand).

7. Prokopidis, Prokopis; Georgantopoulos, Byron; Papageorgiou, Haris. (2011). **A suite of NLP tools for Greek.** *Proceedings of the 10th International Conference on Greek Linguistics*. Komotini, Greece: 1-4 September 2011.

8. Russo, Irene; Caselli, Tommaso; Rubino Francesco. (2011). **Recognizing deverbal events in context.** International Journal of Computational Linguistics and Applications – IJCLA Vol.2 (1-2).

9. Toral, Antonio, Pavel Pecina, Andy Way, & Marc Poch: **Towards a user-friendly webservice architecture for statistical machine translation in the PANACEA project.** In Forcada, Mikel L.; Depraetere, Heidi and Vandeghinste (Eds.) *Proceedings of the 15th conference of the European Association for Machine Translation (EAMT 2011)*. Leuven, Belgium: 30-31 May 2011, pp.63-70.

10. Tim Van de Cruys, Thierry Poibeau and Anna Korhonen. 2011. Latent Vector Weighting for Word Meaning in Context. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP). Edinburgh, UK.
11. Lin Sun and Anna Korhonen. Hierarchical Verb Clustering Using Graph Factorization. 2011. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP). Edinburgh, UK.

6 Consortium and Contact Persons

Núria Bel & Marc Poch **Universitat Pompeu Fabra – UPF** info@panacea-lr.eu	ES	
Nicoletta Calzolari &Valeria Quochi **Consiglio Nazionale delle Ricerche - Istituto de Linguistica Computazionale – CNR-ILC**	IT	
Stelios Piperidis & Prokopis Prokopidis **Institute for Language & Speech Processing - ILSP**	GR	
Anna Korhonen & Thierry Poibeau **University of Cambridge – UCAM**	UK	
Gregor Thurmair **Linguatec -- LT**	DE	
Antonio Toral & Pavel Pecina **Dublin City University -- DCU**	IR	
Victoria Arranz & Olivier Hamon **Evaluations and Language Resources Distribution Agency – ELDA**	FR	

7 Further Information

Please, visit PANACEA web site at www.panacea-lr.eu to keep informed about the project and its progress.

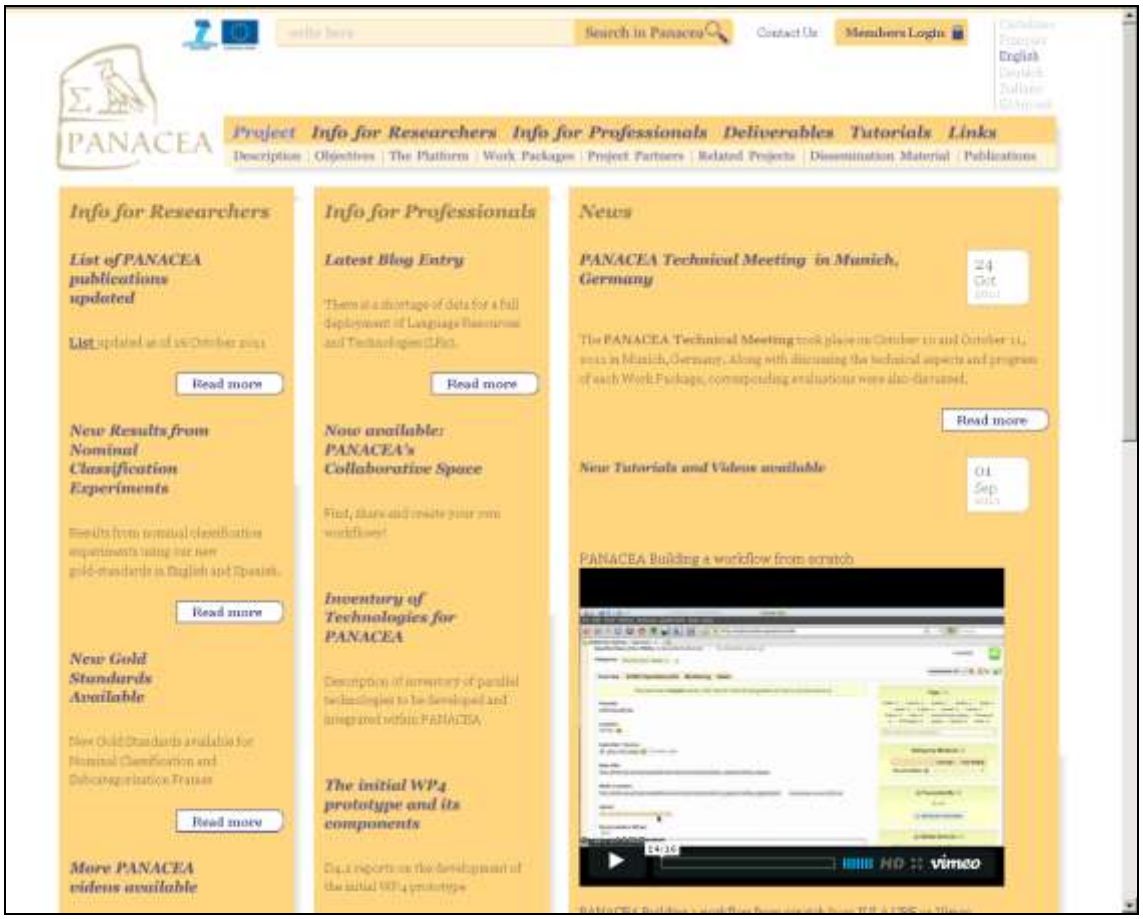

Figure 6. PANACEA web site (14-11-2011)