SEVENTH FRAMEWORK PROGRAMME
THEME 3
Information and communication Technologies

PANACEA Project

Grant Agreement no.: 248064

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

2010 Annual Report

Dissemination Level:	Public
Delivery Date:	November, 2010
Status – Version:	Final
Author(s) and Affiliation:	UPF – Núria Bel

Table of Contents

1 Introduction

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 an MT system to a specific / specialized domain. This a very complex use case, however it does not cover all PANACEA tools, nor all 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 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

Thus, PANACEA reached the first Delivery Milestone in June with precise programs of work and objectives for the rest of the project for facing these questions.

We can summarize objectives and tasks for the technologies to be included in PANACEA as follows and according to the different Work Packages where they will be developed.

The objective of the WP4 are:

1. The creation of a Corpus Acquisition and Annotation (CAA) subsystem for the acquisition and processing of monolingual and bilingual language resources automatically. Therefore, the CAA subsystem includes: i) a Corpus Acquisition Component (CAC) for extracting monolingual and bilingual data from the web, ii) a component for cleanup and normalization (CNC) of these data and iii) a text processing component (TPC) which consists of NLP tools including modules for sentence splitting, POS tagging, lemmatization, parsing and named entity recognition.

2. The development of a Corpus Acquisition Component (CAC) for extracting monolingual and bilingual data from the web is one of the most innovative components of PANACEA. The CAC is the first stage in the PANACEA pipeline for building LR by crawling web documents with rich textual content. To implement the CAC, we use and adapt an efficient and distributed web crawling methodology that collects web pages with content belonging to specific languages and predefined domains. The CAC includes modules that examine if the relevant pages come from sites with content available in more than one language.

3. In addition to a “Corpus Clean-up and Normalization Component” that removes irrelevant parts of downloaded web pages, a Text Processing Component (TPC) deals with the processing of the automatically acquired and normalized corpora. Partners involved in this

task have started adapting and deploying existing NLP tools for the languages addressed by the project. Available lingware in the consortium and other open source tools for sentence splitting, POS tagging, lemmatization, and parsing/chunking are being integrated as web services in the first version of the PANACEA Factory. The aim is to test the scalability and efficiency of these tools in processing the large amounts of data expected. PANACEA partners will take care of developing web services for the tools they will support for each language.

The main objectives and tasks of WP5 "Parallel corpus and derivatives" are:

1. Developing word-aligned and chunk-aligned data from the parallel corpora induced in WP4 for training MT models. This task involves sentence alignment of parallel corpora, parallel sentence extraction from comparable corpora, and consequent sub-sentential alignment on world, chunk, and subtree level.

2. Using the produced sub-sentential aligned data for deriving bilingual dictionaries. This task includes filtering the bilingual dictionaries obtained from the alignments carried out in the previous task. This involves exploiting confidence measures provided by the alignment algorithms and frequency characteristics of the aligned terms in the corpora.

3. Using the produced sub-sentential aligned data and dictionaries for extracting transfer grammars. This task involves exploring several approaches to transfer selection: topic identification, definition of grammatical contexts (morphosyntactic and semantic tests), definition of conceptual contexts (conceptual clustering, co-occurrence interpretation)

Main objectives for WP6 "Lexical Acquisition" can be listed as follows:

1. Create domain-specific monolingual subcategorization frame resources for English, Spanish, and Italian verbs, as fully-integrated components in the PANACEA workflow
2. Explore the creation of domain-specific monolingual subcategorization frame resources for Greek verbs
3. Explore the creation of domain-specific monolingual selectional preference resources for English and Italian verbs
4. Explore the creation of domain-specific monolingual lexical-semantic class resources for English verbs and English and Spanish nouns
5. Explore the creation of domain-specific monolingual multi-word expression resources for Italian
6. Develop a method for inclusion of a confidence threshold for resource-building so that lexicons can be customized for increased precision
7. Develop a lexical merger component as part of the PANACEA platform which is capable of integrating the monolingual lexical resources produced by the PANACEA components into a single lexicon

Although not a technical WP, for PANACEA WP7 deserves also a particular attention. The main objective of WP7 is to take care of the internal quality control in terms of

1. validation of the platform/workflows and

2. evaluation of the components that produce resources as the proof of the smooth integration of the advanced technological components used in workflows. In section

After the summer break a new technical meeting was held in Dublin, in the DCU premises, 11th and 12 of October, the main objective being to organize the deployment of the first version of the PANACEA platform according to the objectives and the plans mentioned above.

3 The PANACEA platform

PANACEA's platform is an interoperability space where to join 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 LR as required by MT and other Language Technologies.

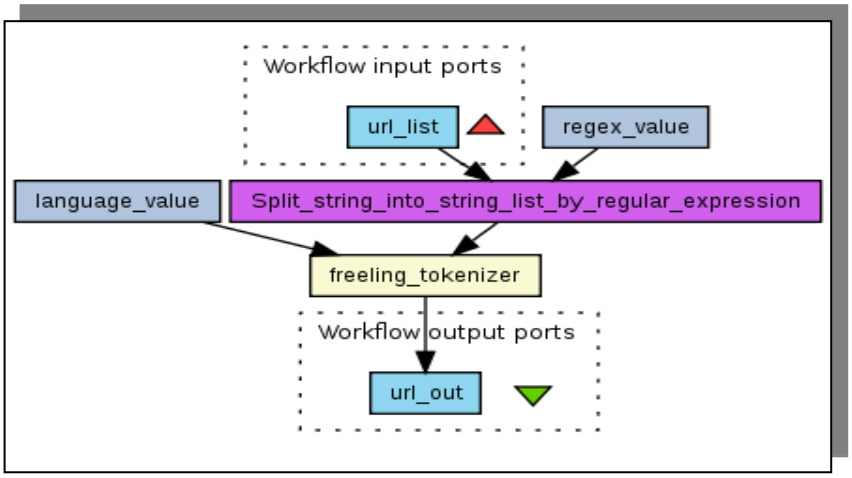

Figure 1: Graphical representation of a workflow drawn by the WF editor TAVERNA

The list of tools to be deployed as web services for the platform are the following:

Functionality	Tool to be deployed as a WS	Host of the WS	Month of delivery as a WS
Bilingual Crawler	Bilingual Crawler	ILSP	January 2011
Monolingual Crawler	Monolingual Crawler	ILSP	January 2011
Sentence aligners	Hunalign	DCU	January 2011

Word aligners	GIZA++,	DCU	January 2011
Word aligners	BerkeleyAligner	DCU	January 2011
Chunk aligners	OpenMaTrEx	DCU	January 2011
Tree aligners	Subtree aligner	DCU	January 2011
Boilerplate removal	Boilerplate removal tool	ILSP/DCU	March 2011
Decomposer for DE	LT-Lemmatiser	LT	March 2011
Document and sentence segmentation for IT	Syn SG	CNR-ILC	March 2011
Duplicate detection	Duplicate detection tool	ILSP/DCU	March 2011
Lemmatiser for DE	LT-Lemmatiser	LT	March 2011
Lemmatiser for EN	LT-Lemmatiser	LT	March 2011
POS Tagger and Lemmatizer for EL	ILSP FBT Tagger & ILSP Lemmatizer	ILSP	March 2011
POS Tagger and Lemmatizer for ES	IULA POS Tagger	UPF	March 2011
POS Tagger and Lemmatizer for IT	Syn SG **	CNR-ILC	March 2011
Sentence Splitter for DE	LT-SentenceSegmentiser	LT	March 2011
Sentence Splitter for EL	ILSP Sentence Splitter and Tokenizer	ILSP	March 2011
Sentence Splitter for EN	LT-SentenceSegmentiser	LT	March 2011

Sentence Splitter for ES	IULA Preprocessing tool	UPF	March 2011
Tokeniser for DE	LT-Tokeniser	LT	March 2011
Tokeniser for EN	LT-Tokeniser	LT	March 2011
Tokenizer for EL	ILSP Sentence Splitter and Tokenizer	ILSP	March 2011
Tokenizer for IT	Syn SG **	CNR-ILC	March 2011
TopicIdentifier for DE	LT-TopicIdentifier	LT	March 2011
Lemmatiser for EN	RASP	UCAM	July 2011
Parser for EN	RASP	UCAM	July 2011
PoS tagger and parser for EN, FR, DE	Berkeley	DCU	July 2011
POS tagger for EN	RASP	UCAM	July 2011
Sentence splitter and tokeniser for EN, FR, DE, ES	Europarl tools	DCU	July 2011
Tokenizer for EN	RASP	UCAM	July 2011
Chunker for EL	ILSP Chunker	ILSP	September 2011
Chunker for IT	Syn SG **	CNR-ILC	September 2011
Dependency Parser for IT	Syn SG **	CNR-ILC	September 2011
Term Extraction (mono) DE	LT-TermExtract	LT	November 2011

Term Extraction (mono) EN	LT-TermExtract	LT	November 2011
NE recognition DE	LT-Namer	LT	June 2011
NE recognition EN	LT-Namer	LT	June 2011
Term Extraction (biling.) EN/DE	LT-BiExtract	LT	June 2011

Table 1: Tentative list of web services to be included in the PANACEA platform

The PANACEA platform typical and main users 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 need not 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 will be announced and described in a Registry that will allow the search and access to them.

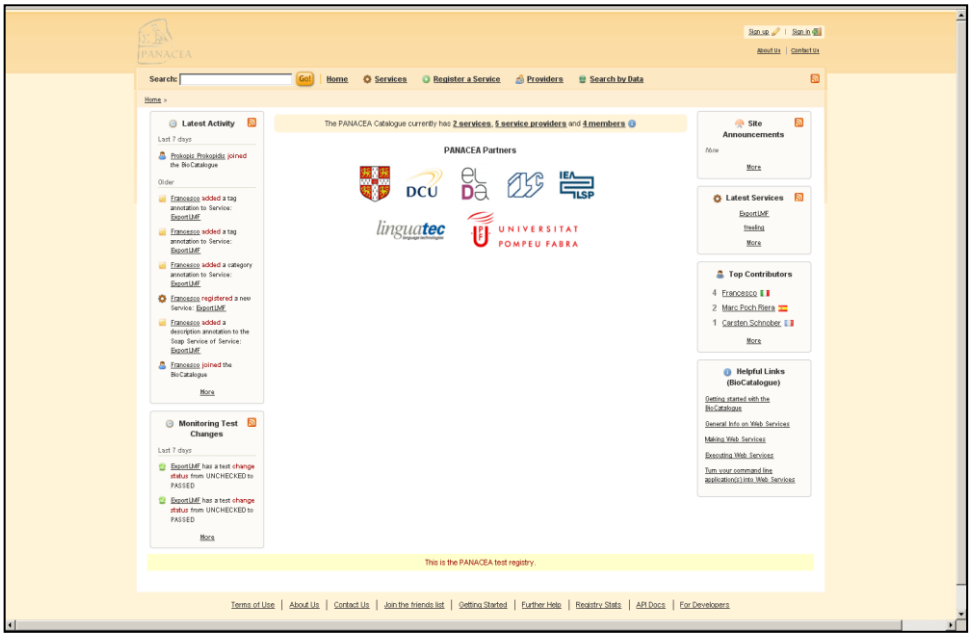

Figure 2: First version of the Registry containing PANACEA web service descriptions

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)]

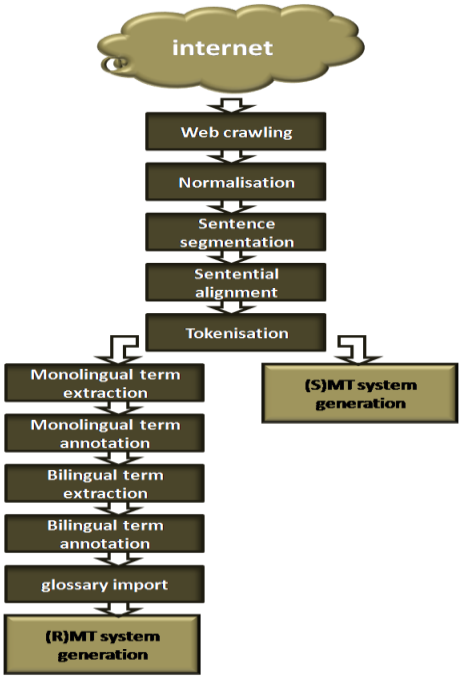

Figure 4: Representation of a production line for MT resource production

4 Dissemination, Promotion and Awareness

PANACEA has participated in the following events:

1. Language Technology Days organized by the EC on March 22-23 2010 in Luxembourg. There has been a project specific presentation together with other research projects, inside Session 7: *Presentation of newly started Language Resources projects*. PANACEA's coordinator Núria Bel approached other project representatives in order to collaborate and find similarities that could forge synergies.

2. Workshop "*Methods for the automatic acquisition of Language Resources and their evaluation methods*", organized by FLaReNet WG6 in collaboration with EU projects: PANACEA, TTC and ACCURAT. The workshop was collocated with LREC 2010 - Language Resources and Evaluation Conference and was held May 23rd in La Valetta, Malta. Approximately 40 participants attended. The objective of the workshop was to agree on a common strategy for evaluating the resources automatically created. The most salient result was the agreement that ACCURAT, TTC and PANACEA will organize a second workshop in LREC 2012 for presenting the results of their projects and the evaluation methods followed.

3. A Workshop on Language Technology issues for International Cooperation was organized by COCOSDA/WRITE and FLaReNet. It included a special session on *Infrastructural initiatives for sharing Language Resources (Data, Tools, and Services).* The workshop was attended by 40 participants.

4. EC Projects Village that took place during the conference days in LREC's 7th edition (from May19th to May 21st). With an expected participation of 1000 attendees, the EC Projects Village gave visibility to PANACEA and offered the opportunity to interact with conference participants. A Project poster was displayed to promote the challenges and goals of the project and brochures were distributed to the visitors.

5. 14th Annual Conference of the EAMT 2010 on May 27th-28th in Saint Raphaël – France. A special session on *"European Community supported projects - plenary presentations"* was dedicated to EU-funded projects. A PANACEA representative made a project presentation together with other EU projects, iTRANSLATE4, META-NET, MOLTO, PLUTO, TTC.

6. PANACEA has participated in the BERLIN THEME TANK meeting organized by META-NET during June 4th-5th 2010.

7. Translingual Europe 2010, on 7th June 2010 in Berlin, Germany. Project Brochures were distributed.

8. Antonio Toral, Pavel Pecina, Andy Way. OpenMaTrEx in Panacea. In the meeting "The future of MaTrEx". Dublin City University. June 2010.

9. Jornada CONNECT-EU, 22-23 of September. The Generalitat de Catalunya organizes a dissemination event for presenting next call of 7FP at Catalan Universities and Companies. PANACEA has been invited to participate in the panel "Results of European Projects".

10. Annual Meeting of the Sociedad Española de Procesamiento de Lingüística Aplicada, 8-10 September 2010, Valencia. A presentation of PANACEA has been accepted at the poster session: Projects.

11. META-FORUM 2010, which is the first edition of the annual META-FORUM conference series, organized by META-NET. META-FORUM 2010, on November 17 and 18 in Brussels. PANACEA presents a poster.

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 mentioned above, 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, MONNET, MEDAR and BOLOGNA.

PANACEA Papers in international conferences

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.

1. Mohammed Attia, Antonio Toral, Lamia Tounsi, Monica Monachini and Josef van Genabith. **An automatically built Named Entity lexicon for Arabic**. In Proceedings of the 7th Conference on Language Resources and Evaluation (LREC 2010). Valletta (Malta). May 2010

2. Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina and Josef van Genabith. **Automatic Extraction of Arabic Multiword Expressions**. Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010), COLING 2010.

3. Bel, Nuria and Coll, Maria and Resnik, Gabriela (2010). **Automatic Detection of Non-deverbal Event Nouns for Quick Lexicon Production.** dins Huang; Chu-Ren and Jurafsky, Dan Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Beijing, China: Coling 2010 Organizing Committee. Pàg. 46-52.

4. Jinhua Du, Pavel Pecina, Andy Way: **An Augmented Three-Pass System Combination Framework: DCU Combination System for WMT 2010.** In Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR (WMT 2010), ACL workshop. Uppsala, Sweden, 2010.

5. Santanu Pal, Sudip Kumar Naskar, Pavel Pecina, and Sivaji Bandyopadhyay: **Handling Named Entities and Compound Verbs in Phrase-Based Statistical Machine Translation.** In Proceedings of Multiword Expressions: from Theory to Applications (MWE 2010), COLING workshop. Beijing, China, 2010.

6. Sergio Penkale, Rejwanul Haque, Sandipan Dandapat, Pratyush Banerjee, Ankit K. Srivastava, Jinhua Du, Pavel Pecina, Sudip Kumar Naskar, Mikel L. Forcada, Andy Way: **MATREX: The DCU MT System for WMT 2010.** In Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR (WMT 2010), ACL workshop. Uppsala, Sweden, 2010.

5 Future Work

PANACEA work plan addresses the improvement of the technologies that will be automatically producing LR. As PANACEA aims at creating a factory for the production of large scale resources, internal evaluation becomes a critical and challenging issue. Critical because it is important to assess the quality of the results that should be delivered to users. Challenging because PANACEA deals with large quantities of data, new domains and aims at performing processing through a technical distributed platform of web services. As briefly mentioned above, R&D will be carried out in the next phase for technical packages and also for evaluation. Evaluation, addressed in WP7, has as a main goal to take care of the internal quality control in terms of the validation of the technical platform and of the evaluation of the components that produce LR. Within WP7, the focus is therefore on technology evaluation, which tries to assess the performance and appropriateness of a technology for solving a certain problem that is well defined, simplified and abstracted. Advances in the technological components will be evaluated against the resources they produce and the way they are presented for promoting their use, i.e. the PANACEA platform. These will be in a variety of languages, since the components are multilingual. Evaluation will be carried out on some languages, as a proof of concept the efficacy of the platform. Given the extent of functionalities in the Platform, it is impracticable to aim at evaluating every single technology and resource integrated and produced. Evaluation will therefore concentrate on the key components of the PANACEA platform.

The first evaluation round will take place in January-February 2011 and will assess the first version of the platform and the components that will be integrated at this stage, namely: the platform, monolingual corpus crawlers for specific domains, and sentence aligners. Thus it becomes our more close future work.

The Platform, that is the integration of web service components, will be validated at all integration and evaluation cycles. The goal of this validation is to check the proper functioning of all the technical components related to the PANACEA platform according to 4 general requirements: *reliability* (the platform outputs are those expected), *robustness* (the platform results are stable and secure), *scalability* (the platform can evolve, relatively to the workload, in adding other components and/or data without restriction) and *usability* (the platform usage is easy). This task will be split into three main parts: technical validation (i.e. some binary criteria to check the architecture), functional validation (i.e. whether the platform requirements are made available or not), and quality validation (i.e. the integration of components keeps the same quality regarding that of the separate components). There will be no validation scores: a requirement is either validated or not. The validation of the PANACEA architecture will be made in a generic environment, without using any specific language and/or domain. In fact, whatever the technical, functional or quality validation is, this must be language and domain independent: a component *technically* working for a given language must work for another.

The monolingual corpus crawler will be evaluated intrinsically. The main tasks in evaluating a focused crawler performance are to formulate and measure its capability to grade a Web page relevance to the topic and guide the process through the most "important" external links of the already visited pages. The main purpose is to provide "accurate" results (i.e. extract text from crawled pages which are highly relevant to these topics). In order to estimate precision, a subset of crawled pages will be selected randomly and each page will be assessed on a four point scale. The experiments will be carried out for every language of PANACEA and the results will be used as feedback to improve the functionality of the Corpus Acquisition Component.

Finally, due to the weak correlations between alignment quality and MT performance and the lack of references for alignment in the domains tackled by PANACEA, aligners will be evaluated extrinsically. The impact that alignment has on the MT performance will be measured using well-known scores employed in MT and comparing against a baseline, which is set to be phrase alignment in Moses (Koehn et al., 2007).

For the first cycle the MT system will incorporate domain monolingual corpora for English, Greek and French in the work legislation and environment domains. These will be used to build the Language Models. As parallel corpora, since the PANACEA ones will not be ready for the first version, existing data will be used, from the Europarl corpus. For this task a test set of parallel sentences is being created that will be used for the evaluations.

6 Further Information

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