Platform for Automatic, Normalized Annotation and Cost-effective Language Technologies Acquisition of Language Resources

PANACOTA

A Small or medium-scale focused research project (STREP)

Challenge 2: Cognitive Systems, Interaction and Robotics,
Objective 2.2: Language Based Interaction

Coordinating Person: Bella X.
University of New Technologies (UNT)

Lists of Participants

<table>
<thead>
<tr>
<th>Nr</th>
<th>Participant name</th>
<th>Acronym</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Coord)</td>
<td>University of New Technologies</td>
<td>UNT</td>
<td>Spain (ES)</td>
</tr>
<tr>
<td>2</td>
<td>Institute of linguistic technology</td>
<td>ILT</td>
<td>Italy (IT)</td>
</tr>
<tr>
<td>3</td>
<td>Speech experts Dept, University of Greece</td>
<td>SE</td>
<td>Greece (GR)</td>
</tr>
<tr>
<td>4</td>
<td>University of UK</td>
<td>UUK</td>
<td>United Kingdom (UK)</td>
</tr>
<tr>
<td>5</td>
<td>German Linguistic Experts</td>
<td>GLE</td>
<td>Germany (DE)</td>
</tr>
<tr>
<td>6</td>
<td>University of Argentina, Linguistics Dept</td>
<td>UoA</td>
<td>Argentina (AR)</td>
</tr>
<tr>
<td>7</td>
<td>Association for Language Resources</td>
<td>ALR</td>
<td>France (FR)</td>
</tr>
</tbody>
</table>
# Table of Contents

B1. Section 1 ................................................................................................................. 5  
B1.1. Concept and objectives....................................................................................... 5  
B1.2. Progress beyond the state-of-the-art.................................................................. 9  
  B1.2.1. A web-services based platform for processing and producing LRs ............ 10  
B1.3. S/T methodology and associated work plan.................................................. 11  
  B1.3.1. Scientific and technological methodology................................................. 12  
  B1.3.2. PANACOTA Work Plan ............................................................................... 14 
  B1.3.3. GANTT chart/timetable: ............................................................................ 15  
  B1.3.4. Work Package List: .................................................................................... 15  
  B1.3.5. Deliverables List: ....................................................................................... 16  
B1.4. Work package descriptions................................................................................ 18  
  B1.4.1. WP1............................................................................................................ 18  
  B1.4.2. WP2............................................................................................................ 19  
  B1.4.3. WP3............................................................................................................ 21  
  B1.4.4. WP4............................................................................................................ 22  
  B1.4.5. WP5............................................................................................................ 23  
  B1.4.6. WP6............................................................................................................ 25  
  B1.4.7. WP7............................................................................................................ 26  
  B1.4.8. WP8............................................................................................................ 27  
B1.5. TOTAL EFFORTS............................................................................................... 28  
B1.6. Work Packages Dependencies: ....................................................................... 28  
B1.7. Significant risks and associated mitigation/contingency plans ....................... 28  

B2. Section 2 ................................................................................................................. 29  
B2.1. Management structure and procedures ............................................................ 29  
  B2.2. Individual participants .................................................................................... 32  
    B2.2.1. University of New Technologies (Spain) .................................................... 32  
    B2.2.2. ILT – Institute of Linguistic Technology (Italy) ........................................ 33  
    B2.2.3. SE – Speech Expert Dept., University of Greece (Greece) ....................... 33  
    B2.2.4. UUK - University of UK (Great Britain) ..................................................... 34
B2.2.5. GLE – German Linguistic Experts (Germany) ............................................................... 35
B2.2.6. UoA – University of Argentina, Linguistic Department (Argentina) .......................... 35
B2.2.7. Association for Language Resources (France) ............................................................. 36
B2.3. The PANACOTA Consortium as a whole ......................................................................... 36
B2.4. Resources to be committed ............................................................................................ 36
B3. Section 3 ............................................................................................................................ 37
B3.1. Expected impacts listed in the work programme ............................................................ 37
B3.2. Dissemination and/or exploitation of project results, and management of ................. 39
B4. Section 4 ............................................................................................................................ 42
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, 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.

PANACOTA is addressing the so-called language-resource bottleneck. In order to supply MT for every pair of European languages, for every domain, and for every text genre, appropriate language resources covering all these aspects should be found, processed and supplied to MT developers. What is needed is an automatic, dynamic and adaptive system for compiling, producing and validating language resources, a system conceived as integrated machinery for the production of LRs.

The objective of PANACOTA is thus to build a language resource factory that step by step covers the automation of all the stages involved in the acquisition, production, updating and maintenance of language resources required by MT systems and in the time required.

PANACOTA will work around four main pillars: 1) the creation of a language resource production platform; 2) the automatic production of massive amounts of LRs for MT; 3) the evaluation of this platform and the LR production chain within the framework of both R&D and industrial settings, and 4) the validation of the resources produced within the project for the selected languages and scenarios.
B1. Section 1

B1.1. Concept and objectives

Despite the complexity of handling its 23 languages, the European Union has decided that cultural and language differences need to be preserved in order not to lose people’s identity. Europe needs to find other means—such as technological ones—to overcome the language barriers to support citizens and industry in a globalised world. The large majority of marketed and industrial technological applications that handle natural language, i.e. Machine Translation (MT), Crosslingual Information Retrieval (CLIR), Multilingual Information Extraction (MLIE), Automatic Document Indexing (ADI), Question Answering (QA), Natural Language Interfaces (NLI), etc., include as critical components Language Resources (LR’s), that is repositories of language data like lexicons, language models extracted from corpora, dictionaries and bilingual dictionaries in the form of correspondences between words or sequences of words in different languages, etc. Besides, most of the currently near-market language technologies for Machine Translation make intensive use of data-driven models that are achieved after the compilation and processing of huge amounts of parallel language texts, which are also LR’s. LR’s, such as those mentioned, are repositories of information, accumulation of observations of each particular language covering different knowledge domains, each with its own and differentiating characteristics. If we compare language data with geographical data (a collection of observational data that are later interpreted by humans and machines), the current status of our LR’s would be similar to this map.

Figure 1: Agnese, 1544

Data from different territories, domains, had been covered but there were still big gaps. The map registered different levels of detail, and even a mixture of data and what is just possible interpretations of the data. Similarly, currently available LR’s do not cover in the same way all languages, and for each language there is a different level of coverage and accuracy depending on knowledge domains (e.g. medicine vs. new energies). Thus, lack of coverage constitutes a severe limitation for the development and proper functioning of language technologies and particularly for MT systems, that cannot work without appropriate LR’s. Moreover, LR’s could not be ever considered complete for any particular language because language change, domain tuning and neologisms are characteristic phenomena of human languages. Note for instance the emergence of new language varieties because of new
technologies: chats, sms messages and blogs are creating new language data that must be taken into account, too. Hence, when discussing the supply of Language Resources, we should not consider it as a single huge endeavour to compile all language resources, for all languages once and forever.

The current solution to the problem of language resources coverage in industrial applications is by means of a continuous supply of hand-made components. Thus, updating, tuning and maintenance of language resources components (annotated texts, lexica, language models, etc) is the most laborious and expensive part of NLP-based applications such as MT, in both Statistical and Rule-based paradigms. The acquisition of resources for each domain and each language is very expensive in terms of human resources because it has to be conducted mostly manually by highly skilled people; it is also expensive in terms of time, because production of the LR’s needed to cover a new market niche delays some times by more than one year the delivery of the complete application. And even worse, the simplicity of the problem, i.e. lack of the appropriate LR’s, affects end user’s confidence in the technology the application is based on.

Therefore, for a European company in the area of Machine Translation with the ambition of covering the enlarged Union market it means having and maintaining 23 different components and over 500 pairs of bilingual components. The cost of creating and maintaining these components is prohibitive, both because the cost of creating the required linguistic resources is too high to ever hope for a return of investment, except maybe for the most frequently spoken languages (from and into English), and because the market for most of these language pairs would be too small to allow a significant revenue to cover the development cost. Thus, this so called “resources bottleneck” is affecting the growth of companies willing to build upon these technologies and to deploy the applications that will be the only help for citizens to bridge language barriers in a digitalized society based on textual information.

What is needed is an automatic, dynamic and adaptive way of compiling, producing and evaluating the adequacy of LR’s, for any particular technology and application. The objective of our project PANACOTA (Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of LR’s for Human Language Technologies) is to cope with this need by developing a platform conceived as a LR’s factory that will step by step cover the automation of all the stages involved in the acquisition, production, updating, validation and maintenance of LR’s. The rationale behind this is:

1) Automation will cut down the current costs of handcrafting a continuous supply of data.
2) Automation will guarantee the production of appropriate LR’s required for a particular application in a much reduced period of time.
3) A dedicated platform that allows the interoperable, adaptable and flexible definition of workflows that produce particular types of LR’s will reduce the dependency on scarce highly skilled people broadening the types of possible users.

Our understanding is that PANACOTA addresses key challenges for the future of Machine Translation applications and of other Language Technologies. A quantitative objective that can provide a clear idea of the impact of PANACOTA’s results is to reduce, at the end of the project, the actual costs (both in human resources and time) of manually building a domain-tuned language resource required by MT providers by more than, on average, 25%. These improvements will be measured by an evaluation carried in an industrial scenario for a Rule-Based Machine Translation system and in a pre-marketing scenario for a Statistical Machine Translation system. This reduction in costs and time in finding, pre-processing and producing application suitable LR’s will be the only guarantee of massive supply of language data as required by current and future applications dealing with languages. Besides, specific research will be carried out in order to study the use of grid computing for the handling of massive data. In order to automatically and massively produce LRs, a number of components are needed, as schematically shown in Fig. 1. The PANACOTA platform will be incorporating these different components that will make possible a step-by-step automation of the whole process of producing LRs.

1. **Corpus creation.** Harvesting and crawling of texts in the web. Access to private archives and access to local files. **Text Pre-processing** for erasing non-linguistic code, indexing and segmenting the text.

3. **Text Annotation** integrating, at a minimum, part-of-speech tagging, lemmatization and
Named Entity recognition (for identification of proper nouns and other non-translatable elements). A further level of annotation that may be required by Parallel Corpus and Language Acquisition tools (cf items 4 and 5 below) will be syntactic analysis that will be performed by modules like chunkers, statistical parsers and dependency parsers.

4. What we have called Parallel Corpus Technologies are different technological components that take as input parallel aligned corpus, annotated or not, to induce phraseto-phrase and word-to-word correspondences (subsentential alignment and bilingual lexica) and the contexts (or conditions) where the correspondence holds and deliver as output a bilingual lexicon and/or a set of transfer rules or transfer grammar.

5. Lexical Acquisition Technologies in turn are also different components that take as input monolingual annotated corpus, annotated in different levels of analysis, to induce highly specialized characteristics of words such as subcategorization and selectional restrictions, lexico-semantic classes, etc.

For the functional combination of these components, PANACOTA will build a platform to functionally integrate the different components involved in the production of LRs. The platform is based on an especially dedicated workflow manager for the composition of different possible systems based on available although remote and distributed web services. This design is in line with the newest paradigms of software development that strongly rely on a future internet of services. According to the EIFFEL Think-Tank [http://www.future-internet.eu], the future internet is conceived as a global and open service delivery platform where all must contribute to the building of networks of services. Thus, our plans are that the components required (as in Fig. 2) for the production of LR’s for a number of particular languages and domains will be deployed as web services and will be integrated in our platform by means of the use of common interoperability standards. The use of these available standards will make PANACOTA an open, ready to grow and truly interoperable network of services where new components could be integrated by third parties. The participation of partners that have been involved in the development of interoperability standards (project LIRICS) is a guarantee of success.

PANACOTA and Topics in the Call 4 Language Based Interaction

Indeed our project handles directly the topics addressed by the Call 4 of the 7FP in the area of Language Based Interaction. We have addressed its proposed Research Aims and Focus for 2009-2010 concentrating on the problem of bridging language barriers by means of computerised methods that can help citizens to enhance communication, collaboration and information exchange, that is: supporting the development and increasing quality of Machine Translation. PANACOTA is addressing a specific solution for the key domain challenge b4: Methods for automated acquisition and annotation of LR’s. In the document Language Interaction Call 4 Guidance Notes the particular outcomes the call requires to cover in the research line b) and the PANACOTA contributions to them.

<table>
<thead>
<tr>
<th>Objectives of the Call</th>
<th>PANACOTA contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Reducing significantly the human effort involved in the collection and processing of language resources”</td>
<td>PANACOTA will significantly reduce human effort by proposing a step-by-step automation of the whole chain of producing language resources required by the two main types of MT systems. A reduction of the number of involved personnel with high skilled profiles (computer engineers, linguistics, domain specialists, etc.) is also aimed at with the delivery of the production platform.</td>
</tr>
<tr>
<td>“Assessing the quality, pertinence and coverage of the automatically extracted information/metadata”</td>
<td>PANACOTA provides components based on new technologies to ensure a quantifiable improvement of the quality and the variety of LRs needed by different types of MT systems and the use of the Rover approach to combine the performance of several aligners producing a better quality aligned corpus at sentence levels</td>
</tr>
<tr>
<td>“Addressing very large datasets, providing a broad coverage of languages, through”</td>
<td>PANACOTA will implement new versions of technologically proved components in order to</td>
</tr>
</tbody>
</table>
highly scalable, language independent methods." make them capable of dealing with all types of languages and massive data. Research on the use of grid technologies will also be carried out along scalability issues and on semi-supervised methods to guarantee language independence of the methods.

The concrete results of PANACOTA will be:

1. The PANACOTA platform: A portable, open, interoperable, workflow editor, with the required middleware and registry processes, for the (semi)automatic acquisition and derivative production of LR's.
2. Deployment of webservises and its metadata annotation data categories for a list of components that, integrated in workflows, will be used for the acquisition and production of LR's, in particular those listed below.
3. Evaluation in an industrial scenario as the proof of concept of the benefits offered by PANACOTA and its potential impact.
4. LR's for the 6 languages of the project English, Spanish, German, French, Italian and Greek (at least, in three for the languages of the project, depending on the final results of the different tools). These LR's will be the following ones and their particular characteristics will be determined by the industrial use cases defined by the project:
   - Sets of parallel texts cleaned and prepared for training/building bilingual translation models.
   - Large monolingual corpus, PoS tagged and lemmatized to be used as annotated resources for training and modelling language data.
   - Monolingual lexica tuned for a particular domain with morphosyntactic, syntactic and lexical-class semantic information.
   - Bilingual dictionaries and specific transfer grammars.

PANACOTA will devise evaluation mechanisms to measure the achievement of these ambitious objectives and will additionally measure its achievements in terms of the reduction of costs in the production of these LR's in a workpackage specifically devoted to the Validation and Evaluation of PANACOTA in industrial environments. In order to cover the production of LR as a global problem, another workpackage will be devoted to issues related to IPR's and possible business models required by this new scenario for the production of LR's. These topics as well as those related to dissemination are detailed in section 3 of this proposal. The consortium expects that the availability of PANACOTA will have a high impact as it will be explained also in section 3. PANACOTA will be a production-oriented facility that will allow reducing the costs, speeding up and validating the design of massive resource acquisition exercises.

PANACOTA will produce resources as proof-of-concept of the proposed components and their integration. The resources to be produced by PANACOTA are basically those recommended as Basic Language Resources Kits (BLARKs) and moreover recommended by the latest reports concerning Machine Translation, the ones produced by Euromatrix (Uszkoreit 2009). The BLARK concept was coined (Krauwer, 1998) in order to define a (minimal) set of LR's to be made available for every language as to meet the needs for this language to be handled by Language Technologies.

According to the ALR's study, the essential resources for each single language are: Annotated and Unannotated monolingual corpora, Monolingual Lexicon, Multi-Bilingual Lexicon, Parallel Multilingual Corpora, Proper Nouns gazetteer, Thesauri, Ontologies and Wordnets, as well as Multimodal corpora specifically for Optical Character Recognition (which will not be addressed in our proposal as they are not crucial for the Xlingual applications the call is addressed to). For most of these resources a generic automation of their production is currently possible with enough guarantees. PANACOTA's partners are the developers of generic automation technologies for the production of LR's, technologies capable of reducing the costs of producing these LR's for new languages and, for each language, for new domains and genres, thus aiming at a full coverage of each language.

To confirm ALR's studies particularly for the area of Machine Translation, the central topic of this Call on Language Based Interaction, we have taken into account the EU Euromatrix project report (Uszkoreit 2009) because it clearly points at the importance that the supply of
LR’s has for the future of Machine Translation. According to the Euro-MATRIX report the situation for MT is the following:

1. Statistical MT systems have shown a considerable progress in the last decade, but they still have problems for achieving the same level of quality than rule-based systems. SMT systems are considered interesting because they are seen as demanding a faster and cheaper development. However, their quality and performance are threatened by the following requirements that concern exclusively the availability of appropriate LRs:
   - Lack of large quantities of parallel language data for all possible pairs of languages.
   - Lack of large parallel language data for all possible text genres and domains.
   - Classification of language data w.r.t. sorts, domains, registers, etc.

In addition to these reports, we have also reviewed the work reported on by the Treble-Clef project (http://www.trebleclef.eu/) both in their deliverable 3.1 “Bringing Multilingual Information Access to Operational Systems, – TrebleCLEF System Developers Workshop REPORT” and deliverable D5.2 on “BEST PRACTICES IN LANGUAGE RESOURCES FOR MLIA”, to confirm our analysis.

To provide the means to ensure the acquisition of massive quantities of texts by proposing a step-by-step automation of the compilation and alignment for all possible pairs of languages and for all possible genres and domains is the objective of PANACOTA. The platform should be able to harvest data from all sources according to predefined quality criteria (possible pairs of languages, possible genres, and domains) and be exploited by the necessary tools for cleaning, annotating, formatting and normalizing. For instance, PANACOTA will incorporate mechanisms for the acquisition of massive quantities of texts which will be compiled and whose degree of comparability and alignment, in particular for data coming from different sources, will be assessed (an alignment/comparability score will be computed).

Another innovative asset of PANACOTA is that it will explore means for the identification and compilation of LRs from private and/or hidden web sites/pages. Thus, while technically we will be solving the access to and easy retrieval of texts from these private or local archives, a business model will explore new legal ways of exploiting these materials. Finding information about private multilingual archives and the formal contacts with their representatives, as well as proposing solutions for the legal issues that could arise will be some of the tasks carried out by PANACOTA.

2. Rule-based MT systems are currently the most marketed and used ones, but the bottleneck that threatens their progress is the lack of a stable and adequate supply of LRs and specifically of the following ones.
   - Lack of transfer grammars for a large variety of different language pairs and of particular domains and genres
   - Lack of comprehensive bilingual lexica
   - Lack of sufficient information in monolingual lexica

The use of highly innovative technologies for the automation of the production of such urgently demanded transfer grammars, bilingual lexica and monolingual lexica is another objective of PANACOTA.

Another of the most innovative assets of PANACOTA is that our project can address the automatic production of rich lexica (including subcategorization information, selectional restrictions, lexico-semantic classes), both monolingual and bilingual, that, as can be seen in BLARK matrices and in the EuroMatrix report, are among the most demanded resources.

Lexical production components will be based on known and proven techniques on which the partners have been working for several years (Korhonen 2002 and Korhonen et al. 2006, Sun et al. 2008; Bel et al. 2007 and 2008, Quochi et al. 2008, Piperidis et al. 2007) and which are ready to be transferred for the industrial production of LR’s.

B1.2. Progress beyond the state-of-the-art

Besides the innovative concept and implementation of a factory of LR’s, PANACOTA is to improve significantly the integrated components that will be processing and producing resources. We describe below all these innovative aspects.
B1.2.1. A web-services based platform for processing and producing LRs

The automatic production of LRs is crucially based on text processing components. Thus, while the status of the current technologies for automatic language resource acquisition and tuning are mature enough to benefit industrial applications, there are still many aspects related to their integration into particular application development processes that need to be considered. The objective is to propose them in an industrial scenario for dedicated, massive, high-volume and high-performance production of LRs on which the industry can rely.

Current NLP platforms for the integration of text processing components (such as GATE or UIMA) require that users spend time and resources installing, integrating different tools, and defining processes and object types being processed. They may also need to locally install different components and re-purposing each installation when new components are developed or required. In order to improve the current situation, PANACOTA will make use of the recent developments in the definition and supply of processing components as web services, i.e. software systems designed to support interoperable machine-to-machine interaction over a network (Service Oriented Architecture, SOA). It will be chained in workflows defined by users by means of a workflow editor and the required middleware.

Workflows will be the assembling of several distributed application components that together can handle the acquisition and processing of large LRs. With the loose chaining of services by means of established standards, SOA enables dynamic, flexible applications which can always change rapidly by integrating new services and, in principle, old legacy systems even across technological platforms. Thus the task of gathering text according to a domain, and language, as well as for its cleaning, preprocessing and annotation can be seen as a workflow that handles the synchronization of the different tasks required, and that minimizes the impact of working with different sources and from different components. A platform that assists in the creation of a flexible workflow composition and execution is the instrument required for a rapid development and validation of LR’s reducing the production costs.

Thus, PANACOTA, supports the view of the future internet initiative (www.future-internet.eu), and wants to deliver an especially dedicated workflow manager for the composition of LRs production systems based on webservices. A number of webservices will be specified and integrated during the life-time of the project in order to prove the platform by producing a number of resources, as required by the work program and mentioned above. However more components can be integrated to build different workflow chains by making an actual use of the same standards (for describing input/output requisites) creating, as required by the current work program, an open and interoperable network of services that the future internet is advocating for. The addition of new services will be tested and validated within the project with the production of LR’s.

A web services-based system is based on and fosters interoperability because it proposes the publication of input/output descriptions required and makes de facto standards emerge. In the field of Language Technologies, the emergence of standards is also supported by the current efforts made by a number of standardisation bodies such as ISO, W3C, ETSI, etc. For instance, the ISO TC37/SC4 is regulating the formal and semantic characteristics of the processes and objects which are most frequently handled in annotation and acquisition exercises (the data category repository ISOcat, www.isocat.org). PANACOTA will take advantage of the use of these proposals for standards (several partners have actively participated in its definition and are still involved in ISO TC37/SC4 subcommittee) to support the development of web services.

PANACOTA proposes a way into a future where new systems and new resources will be integrated in different workflows as soon as they are deployed as web services by asking these providers to use a set of particular standards.

A large number of components are involved in such a platform, in addition to those already mentioned below. The links between components are shown in Figure 3.

- The work flow engine, working as a centralized server that handles the processing chain;
- The web services, which are plugged onto the platform through the network and used to process data, and also to interact with registered software or linguistic tools;
- The user interface(s) and work flow editor, used to configure, run the processing chains and access the results;

The web services registry and the tools registry, which store the available services and technologies, respectively;

- The transfer mechanisms, debugging tools, etc.
Furthermore, the platform will be able to use web services in parallel (such as Task C and Task D in Figure 3) and process data with several applications of the same technology, so as to produce ROVER data. Another interest of PANACOTA is to provide a dynamic platform that would allow the use of several chains producing the same type of LRs, but with different tools, so as to compare them.

Research regarding acquisition mostly focuses on extraction itself and disregards the next step: what to do with the results of the acquisition procedure and how to take advantage of them. Differently, in this project we will go beyond acquisition by posing ourselves the question: What can we do with the existing material? How can we take the biggest advantage of it for further exploitation? We will look for solutions to these questions, thus progressing from the current state-of-the-art. In order to tackle this issue, PANACOTA platform will also foresee the creation of repositories which can store different pieces of information acquired (Toral et al. 2008) and its merging with new or legacy data.

A web services-based architecture will also allow us to study and to implement the use of grid computing for the gathering of data from different users and a common processing chain to be distributed as to reduce the impact of working with massive data.

Our project is specifically aimed at achieving a significant progress along the following challenges:

Accurate, large-scale and portable acquisition techniques. One of the biggest current research challenges is to improve the accuracy of existing techniques further and to replace small-scale techniques with more powerful and portable techniques. Without this leap, the technologies will always be limited in what they can achieve. This challenge, which remains to be tackled in most areas of automatic lexical acquisition, requires developing more accurate technology capable of (i) large-scale acquisition from very large corpora, (ii) domain-specific acquisition from smaller (and possibly sparse) corpora, and (iii) reliable acquisition of both high and low frequency lexical items despite the size of the corpora. We propose the use of sophisticated filtering and back-off modules, and investigating realistic approaches for domain-tuning of acquisition techniques (Korhonen et al. 2008 and Bel et al. 2008).

Multi-lingual lexical acquisition. Many of the currently available techniques have been developed for English and other resource-rich languages only. Evaluating the applicability of the techniques to other languages would be critical for both theoretical and practical reasons; for 1) improving the accuracy, scalability and robustness of the techniques, 2) advancing work in other languages, 3) gaining a better understanding of the language-specific / cross-linguistic components of lexical information, and 4) improving the performance of (multilingual) NLP applications. We propose acquiring resources for Spanish, Italian and Greek as a validation of the multilinguality of the approaches (Bel et al. 2007, Quochi et al. 2008).

Large-scale application to build and tune existing lexical resources and aid important real-world application tasks. Although automatically acquired lexical information and associated frequency data carries potential for further improvement in performance in many NLP application tasks, much of the research has focussed on small-scale experiments. The usability of automatically acquired lexical data for lexical resource building and tuning, and enhancing important applications is largely yet to be demonstrated. Some important under-explored tasks and applications include e.g. machine translation as applied to both general and domain-specific data within and across different languages.

B1.3. S/T methodology and associated work plan

The technical workpackages of PANACOTA (From WP3 to WP6) will be developed along two main phases: analysis, and development. The evaluation has been considered as a separate WP because of the need of evaluating the integration of the components and the produced resources. The main phase of each technical WP, development, will be organized in three iterative cycles of development of components and their integration into the platform, i.e. deployment as a web service and the creation of middleware required for having working components. But the success of PANACOTA will be based on “keep seeing the big picture”, i.e. a platform for the creation of workflows that will be factories of different types of LRs. While each particular component can be substituted by another doing the same function, the way of integrating the components must be stable as soon as possible.
12

B1.3.1. Scientific and technological methodology

The general methodology of the work plan will be further explained below. Now, we will first address the most challenging issues of the Scientific and Technical Methodology for each of the main technological components of PANACOTA. After we give details of all the work packages, and finally we will analyse the main risks and we will afford contingency plans. The methodology concerning Management is addressed at length in section 2, and the methodology for WP2 having to do with dissemination and legal issues about exploitation will be covered in section 3.

B1.3.1.1 The PANACOTA Platform

In PANACOTA, the platform can be seen as a LRs Factory where the raw material is captured by usual corpus creation methods (crawling the web, accessing to archived texts, or local documents of the user computer) then this material is cleaned, and processed in order to create derivatives: annotated corpus, parallel corpus and annotated parallel corpus. These first derivatives are later used to produce, by means of induction tools, a second order or synthesized derivatives: rich lexica (with morphological, syntactic and lexico-semantic information) and massive bilingual dictionaries (word and multiword based) and transfer grammars. The platform should also make available repositories to merge new with old resources as well as validation tools for each of the LR’s derivatives.

The platform is, in sum, the functional integration of a variety of components. To minimize installation and interoperability problems, we have opted, in line with current proposals for the future internet, for proposing a web services based scenario (SOA) where each of the components will be wrapped as an independent web service that can be chained into dedicated workflows that have as objectives the production of the different resources. Web services are function based systems where given an input they deliver an output. In order to guarantee interoperability, input sent to the web service must comply with the specifications declared in the description of the web service itself. These descriptions are declared as metadata in xml format that inform possible users about the characteristics both of the input and of the output of every particular web service. PANACOTA will provide a specific registry and standards for declaring such requirements as well as for allowing a quick and user friendly search of web services and its requirements (based on WSDL and UDDI). The platform must face the “managing legacy systems”, especially for requests for both joining and contributing to the platform by providing proprietary resources for specific tasks. Legacy systems have usually been developed neither without following interoperability guidelines nor with specific access interfaces. In addition, these systems may have been developed using old languages or self-made techniques, often designed for specific in-house purposes. Frequently, these programs offer only a simple command line interface to the users. In order to easily integrate these tools and facilitate the construction of web services on top of such resources, the factory needs to take into account these difficulties. In order to tackle these issues, the factory will define a think-tank-force especially focused on the managing of these legacy systems, whose main aim is to support institutes to set up their resources to achieve compatibility with the platform. Briefly, this force will help users map their tag sets into shared standard ones providing both consultancy on standards and developing some mapping software.

Thus the platform will provide with a dedicated work flow editor (inspired in myexperiment.org) and a registry (where information about all possible components will be stored) that will assist in the creation of flexible workflows defined by jobs templates that can be edited and partially replicated and thus that can manage the following possible scenarios:

- The same job but with a new data set, when no specific modification because of the data (language, for instance) has to be done.
- A variant of a type of job. When one variable can determine the selection of one of the modules (language and pos tagger, for instance).
- Specifying only critical parts of the job. When the user don’t need to specify jobs that are common and can be predefined as templates.
- Composition of jobs. Where the output of a job can be the input of another job.
- New types of analysis. Free selection of components, that can be new and even including human intervention.

This webservice-based scenario where inputs and outputs must be interoperable is crucially dependent on the use of standards for defining allowed input/output formats. We will be using
already defined standards for web service declaration and specifications: W3C XML Schema, SOAP, REST, being the most used. For linguistic materials will be using ISO TC37/SC4 formats: Lexical Markup Framework (LMF, Francopoulo et al. 2008); Morpho-syntactic Annotation Framework (MAF, Clement and de la Clergerie, 2005) and we will work according to ISOcat enterprise for the cooperative construction of a repository of DataCategories for NLP.

B1.3.1.2 Corpus Acquisition and annotation

Specific services will be incorporated within the platform and would consist of the required tools for semi-automatic harvesting, cleaning, annotating, and packaging textual corpora. Such tools should allow crawling the web at different levels and with different types of data (textual documents, pdfs, html, etc.) and identifying relevant data according to pre-defined criteria (language, domain, genre, size, timespan, etc.).

One of the ideas of the project is to support and benefit from self-improving data-driven tools that are an acceptable performance level. It will exploit them to bootstrap an acquisition process with reasonable manual work and once the amount of data collected and (semi)automatically annotated is large enough it will inject that in the training process to improve the performance and thus the quality of the processed data through this automatic, dynamic and self-organising acquisition, and processing of the data. Some technologies and annotation tools may lead to semi-automatic processes from which it would benefit while others may implement fully automatic processes from which high level techniques would benefit (e.g. text tagging versus text alignment/lexica extraction for MT).

We plan that among the tools that will be incorporated within the platform we will primarily have a number of text aligners (at different levels: words, sentences, named entities, etc.) The aligners will allow us to compute a comparability score as well as an alignment score. The comparability score will indicate how comparable two corpora are and if one could extract aligned pieces from the data or not; If this is doable the alignment score will help assess the amount of supervision and human work that will be required to align the texts at various levels.

Another experiment will be the combination of several aligners using the Rover method exploited in speech recognition (Recogniser Output Voting Error Reduction, Fiscus, 1997) and in textual Pos and syntactic tagging (references: Grace/Multitag/Easy projects & Passage). The experimentation e.g. within TC-STAR and Easy shown that Rover method that combines three different systems based on different approaches lead to better performance than the best system. These techniques will be extended to other types of data..

Sources of information: the need for data

The web is considered to be one of the main sources of text, which is the raw material of our factory. The web is offering texts organized in clouds of links that can determine the domain, language of its contents. To keep track of this information is crucial and a special crawler will be designed to keep a continuous stream of data being processed although keeping specific information about genre, domain, languages. PANACOTA also counts on the imminent deployment of metadata as part of the W3.0 or Semantic Web which will contribute to the feasibility of an easy structured capture of materials from the web by means of harvesting methods. Apart from "the web as a corpus" we will consider exploiting also the so-called "New Text" (resources that have emerged from the Web 2.0 collaborative paradigm, e.g. Wikipedia, Wiktionary, Omega Wiki). Compared to corpora, New Text presents two additional advantages: corpora presents a high degree of subjectivity (Wiebe et al 2004) and information is extracted in a non-canonical form (Suchanek et al 2007). Research has given the first proofs of the usefulness of New Text in order to create LRs. First attempts to exploit New Text in order to build LRs include proposals to build bilingual dictionaries (Jones 2008), ontologies (Pedro et al 2008) or thesaurus (Milne et al 2006). In addition, the availability of APIs (Zesch et al 2008) facilitates developing acquisition procedures for this kind of resources.

But most of the content is nowadays stored in private archives or databases. Agreements with the owners of these archives must be devised by means of a generic collaboration contract. Special attention deserves proliferating digital libraries and archives which are a homogeneous source of data. The project will negotiate rights will all regional and international organizations that use languages in a multilingual setting (UN, UNESCO, UNICEF, FAO, WHO, European institutions, Santé Canada, Media group (e.g. Euronews, Arte, France-24, BBC, AFP, Reuters, etc.). It is crucial to consider a number of dimensions when talking to these sources in particular
the typology of the language used (journalistic, novels, health reports, etc.) but also the volumes. A monthly magazine like "Le-monde diplomatique", with whom ELRA has several agreements, produces over 73 versions in 26 languages; part of this is translations that would constitute a useful dataset for SMT.

B1.3.2. PANACOTA Work Plan

PANACOTA’s duration is 36 months and it is a complex research project because of multiple interdependencies among work packages and the need for each group to contribute to a single system/platform. In order to structure the technical work in PANACOTA, it is going to be structured into three main phases. Analysis, Development and the Evaluation of the integration and of the produced resources. 

The **Analysis** phase will cover the first 3 months of the project and will collect information from all workpackages in order to get an overall picture of the requirements to be faced and the availability of components to be integrated. The outcomes of this Analysis phase will be:

- The Analysis of Requirements Report. While technical WPs (WP4-WP6) will concentrate on listing available components and input/output requirements (DX.1), the non-technical ones will concentrate on questions to be taken into account from: WP2 Dissemination and Licencing of resources, tools and business models to be explored (D2.1) and WP8 in Requirements from the MT systems production: type of information required, formats, domain characteristics to be evaluated, etc. (D8.1). The final compilation of the Analysis Requirement Report will be done by WP7 that will use it to establish criteria of validation (D7.1).
  - The Basic Design of the architecture of the whole platform (D3.1)
  - Selection of Work Flow editor and engine (D3.2).

However, given the duration of this research project, partial revisions of the analysis made will be considered after each development round. That is, if necessary, more deliverables or revised versions of existing ones can come out along the project. The objective of this first analysis is a quick start of the development phase.

The **Development** phase is the central phase of the project, and aims to deploy all the webservice middleware, deployment of web services covering all the existing components and the implementation of new software tools to produce the final PANACOTA outcomes. It will last from month 6 to month 30. The development is going to be organised in three consecutive cycles of 8 months each, mainly involving Work Packages WP3 to WP7. WP3 is devoted to the standards and metadata required by the interoperability aimed at in PANACOTA, as well as the WorkFlow machinery. Three work packages, from WP4 to WP6, will be devoted to the adjustment and development of different components for automatically producing different LR’s. Development in cycles of 8 months are meant to ensure an available platform as soon as possible with a first set of web services of already available components (NER, Crawler, PreProcessing, POS annotation, and simple Aligner), while in cycle 2 at t21 in addition to the integration of more components such as those integrated in cycle1 for possible choices and comparison of results, we will deal with new developed components covering new technologies for improving quality, sub-sentential alignment, lexical acquisition and bilingual lexicon automatic production. Cycle 3 will deliver new versions, of industrial impact, of these new components in time for the industrial evaluation that should start at t30.

**Deliveries** for each of these components WP’s are structured according to internal critic points: a first version for integration (milestones M2, M5 and M9) and a second version with documentation and possible changes due to the integration (milestones M4, M7 and M10). In turn, the WP3 has a first delivery in order to be evaluated by WP7 (milestones M3, M6 and M9), and the final delivery of each phase together with documentation.

WP7 is devoted to evaluate the integration of the tools, the functioning of the Work Flow and the good quality of the LR’s results obtained after each cycle of development. Its aim is to verify that intermediate results meet the requirements and to provide feedback to take the necessary corrective measures if required. After each cycle, the verification will result in LR’s produced and evaluation reports that must clearly state the resources involved in the creation of new resources and the inherent quality of the results achieved according to the requirements identified in the analysis phase. The Industrial validation in real scenarios and against real tasks is the object of WP8. During this validation the usefulness, i.e. the benefits of using automated
tools in a dedicated platform will be identified and quantified.

B1.3.3. GANTT chart/timetable:

![GANTT chart](image)

B1.3.4. Work Package List:

<table>
<thead>
<tr>
<th>Work-package No.</th>
<th>Work-package title</th>
<th>Type of activity</th>
<th>Lead part no.</th>
<th>Lead part. short name</th>
<th>Person-story</th>
<th>Start month</th>
<th>End month</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP1</td>
<td>Coordination &amp; Management</td>
<td>MGT</td>
<td>1</td>
<td>UPF</td>
<td>30</td>
<td>T1</td>
<td>T6</td>
</tr>
<tr>
<td>WP2</td>
<td>Dissemination and Exploitation</td>
<td>RTD</td>
<td>7</td>
<td>ELDA</td>
<td>29</td>
<td>T1</td>
<td>T6</td>
</tr>
<tr>
<td>WP3</td>
<td>The Platform</td>
<td>RTD</td>
<td>1</td>
<td>UPF</td>
<td>68</td>
<td>T1</td>
<td>T2</td>
</tr>
<tr>
<td>WP4</td>
<td>Corpus Acquisition and annotation</td>
<td>RTD</td>
<td>3</td>
<td>ILSP</td>
<td>49</td>
<td>T1</td>
<td>T2</td>
</tr>
<tr>
<td>WP5</td>
<td>Parallel corpus and derivatives</td>
<td>RTD</td>
<td>6</td>
<td>UD</td>
<td>78</td>
<td>T1</td>
<td>T2</td>
</tr>
<tr>
<td>WP6</td>
<td>Lexical Acquisition</td>
<td>RTD</td>
<td>4</td>
<td>UC</td>
<td>76</td>
<td>T1</td>
<td>T2</td>
</tr>
<tr>
<td>WP7</td>
<td>Evaluation of integration and resources</td>
<td>RTD</td>
<td>2</td>
<td>CNR-ILC</td>
<td>46</td>
<td>T1</td>
<td>T6</td>
</tr>
<tr>
<td>WP8</td>
<td>Evaluation in industrial environment</td>
<td>RTD</td>
<td>5</td>
<td>Linguate</td>
<td>22</td>
<td>T1</td>
<td>T6</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>398</td>
</tr>
</tbody>
</table>
## B1.3.5. Deliverables List:

<table>
<thead>
<tr>
<th>Del. no.</th>
<th>Deliverable name</th>
<th>WP no.</th>
<th>Nature</th>
<th>Dissemination level</th>
<th>Delivery date</th>
<th>Delivery month</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1.2.6.</td>
<td>Contractual Periodic Progress Report</td>
<td>WP1</td>
<td>R</td>
<td>PP</td>
<td>7, 13, 19, 25, 31, 36</td>
<td></td>
</tr>
<tr>
<td>D1.3, 9 and D0.9</td>
<td>Annual Public Reports</td>
<td>WP1</td>
<td>R</td>
<td>PP</td>
<td>12, 25, 29</td>
<td></td>
</tr>
<tr>
<td>D2.3.m</td>
<td>Project brochures and dissemination materials</td>
<td>WP2</td>
<td>O</td>
<td>PP</td>
<td>12, 24, 30, 36</td>
<td></td>
</tr>
<tr>
<td>D2.1.</td>
<td>Dissemination Plan</td>
<td>WP2</td>
<td>R</td>
<td>PP</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>D2.2.m</td>
<td>Work-packs (plus monthly updates: content, titles, etc)</td>
<td>WP2</td>
<td>O</td>
<td>PP</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>D3.1</td>
<td>Requirement analysis of the platform architecture and design, W1</td>
<td>WP3</td>
<td>R</td>
<td>PP</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>D4.1</td>
<td>Report describing the inventory of corpus compilation tools to be developed and integrated in UAC, W1</td>
<td>WP4</td>
<td>R</td>
<td>PP</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>D4.2</td>
<td>Report describing the inventory of corpus cleanup and normalization tools to be developed and integrated in CNC</td>
<td>WP4</td>
<td>R</td>
<td>PP</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>D4.3</td>
<td>Report describing the inventory of annotation tools to be developed and integrated in TPC</td>
<td>WP4</td>
<td>R</td>
<td>PP</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>D6.1.</td>
<td>Report describing the inventory of parallel technologies tools to be developed and integrated in FCT</td>
<td>WP5</td>
<td>R</td>
<td>PP</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>D6.2.</td>
<td>Integrated report describing the inventory of technologies and tools to be developed and integrated in PANACEA and criteria for evaluating its results</td>
<td>WP6</td>
<td>R</td>
<td>PP</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>D7.1</td>
<td>Analysis: Requirement Summary and Criteria for the evaluation of resources, technology and integration</td>
<td>WP7</td>
<td>R</td>
<td>PP</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Del. no.</td>
<td>Deliverable name</td>
<td>WP no.</td>
<td>Nature</td>
<td>Dissemination level</td>
<td>Delivery date month</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>------------------</td>
<td>--------</td>
<td>--------</td>
<td>---------------------</td>
<td>---------------------</td>
<td></td>
</tr>
<tr>
<td>D3.1</td>
<td>Analysis of user requirements</td>
<td>WP3</td>
<td>R</td>
<td>FU</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>D4.1</td>
<td>制订函数原型的CTA subsystem和其组件(CAC, CNC and TPC)</td>
<td>WP4</td>
<td>P</td>
<td>CO</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>D5.1.1</td>
<td>需求在平台上的集成,第一版</td>
<td>WP5</td>
<td>P</td>
<td>CO</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>D5.2.1</td>
<td>第一次版本的集成平台</td>
<td>WP3</td>
<td>O</td>
<td>CO</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>D5.2.2</td>
<td>第一次版本的集成平台</td>
<td>WP3</td>
<td>P</td>
<td>CO</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>D6.2.2</td>
<td>需求在平台上的集成和文档,以及用于训练/构建/翻译的工具和模型</td>
<td>WP3</td>
<td>P</td>
<td>CO</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>D7.2</td>
<td>最终评估报告</td>
<td>WP3</td>
<td>R</td>
<td>CO</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>D4.3</td>
<td>修订的函数原型的CTA subsystem和其组件的集成</td>
<td>WP7</td>
<td>R</td>
<td>CO</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>D5.3.1</td>
<td>第一次版本的组件</td>
<td>WP5</td>
<td>P</td>
<td>CO</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>D5.4.1</td>
<td>第一次版本的组件</td>
<td>WP5</td>
<td>P</td>
<td>CO</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>D6.2.1</td>
<td>第一次版本的组件</td>
<td>WP5</td>
<td>P</td>
<td>CO</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>D6.3.1</td>
<td>第一次版本的组件</td>
<td>WP5</td>
<td>P</td>
<td>CO</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>D3.1.1</td>
<td>第二次版本的集成平台</td>
<td>WP3</td>
<td>O</td>
<td>CO</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>D3.1.2</td>
<td>第二次版本的集成平台</td>
<td>WP3</td>
<td>O</td>
<td>FU</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>D4.4.4</td>
<td>修订的函数原型的CTA subsystem和其组件和文档</td>
<td>WP4</td>
<td>P</td>
<td>FU</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>D5.3.2</td>
<td>集成版本的CTA Dictionary Extractor和文档</td>
<td>WP5</td>
<td>P</td>
<td>FU</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>D6.4.2</td>
<td>集成版本的组件</td>
<td>WP5</td>
<td>P</td>
<td>FU</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>D5.2.2</td>
<td>集成版本的组件</td>
<td>WP5</td>
<td>P</td>
<td>FU</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>D6.2.2</td>
<td>集成版本的组件</td>
<td>WP5</td>
<td>P</td>
<td>FU</td>
<td>22</td>
<td></td>
</tr>
</tbody>
</table>
B1.4. Work package descriptions

Work packages are described in separate pages in the following sections.

B1.4.1. WP1

<table>
<thead>
<tr>
<th>Work package number</th>
<th>WP1</th>
<th>Start date or starting event:</th>
<th>M1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>Management and Coordination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity type ¹</td>
<td>MGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant number</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Participant short name</td>
<td>UNT</td>
<td>ILT</td>
<td>SE</td>
</tr>
<tr>
<td>Person-months per participant</td>
<td>12</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Objectives

Set up and implement an efficient coordination and management of the overall project. Key aspects to be addressed are: project planning, monitor and stimulate project progress in accordance with objectives, overall coordination of WPS, internal communication and project external representation, risk management, administrative, financial and legal procedures, drive the Project’s decision making mechanisms, feed them with timely information and suggest, support and monitor corrective actions if required.

Specific key objectives are:

- Ensure achievement of general and technical objectives within the constraints of time, budget, quality and compliance with related regulations
- Efficient communication with EC, including timely and complete reporting on project status and compliance with programme procedures

Description of work (possibly broken down into tasks) and role of partners

This WP is in charge of the overall coordination and management of the project so that all stakeholders regard it as a success. The WP will set up and implement the suitable structure and procedures to guarantee the fulfilment of the objectives. It covers high-level issues and daily management, as well as the coordination of the integration of daily matters into the project’s bigger picture.

This WP monitors that deliverables and milestones are achieved in time, quality and budget in close cooperation with relevant WP leaders. This WP carefully analyses and keeps track of the evolution of risks and its palliative solutions, and drives the process for implementing corrective actions if required across the project defined decision mechanisms. This WP pays special attention to optimal communication mechanisms that guarantee that relevant information is available at any moment to required stakeholders, including the European Commission.

In order to design, implement, facilitate and monitor the execution of the actions described in the work plan, this WP will be in charge of the preparation of project meetings, where project goals, plans, procedures and relevant information on progress will be discussed. This WP will be also responsible for project administration within partners’ organizations.

This WP will be executed mainly by the Project Manager and supervised by the Coordinator and the Project Board. The Coordinator will be responsible to ensure that correct procedures are carried out and all deadlines and obligations met. It will be her responsibility to seek consensus on project steering by clearly separating the basic functions to be performed.

Coordination is structured into two levels:

High level decision making: Consensus (whenever this is unreachable a majority voting

¹Please indicate one activity per work package: SUPP = Support; MGT = Management of the consortium.
would be enough) will be required for both strategic & technical decisions. A senior committee decides about overall project strategy: the Project Board Committee (PB) and will be chaired by the Project Coordinator (PC). The PC will monitor project progress according to objectives maintaining the wider view and always considering longer term aspects going beyond the immediate achievement of technical results. The PB represents the interests of all partners and each partner has one voting representative. A critical aspect for the success of this aspect will be the availability and presentation of relevant information on progress and risks, executed by the Project Manager.

**Day-to-day management:** This level of Management will be driven by a Project Manager and the leaders of each of the WPs and Tasks following plans described in the Work Program and decisions at project level.

There are three major distinct functions to be covered in the daily activity of the RTD project (and extended to every level): administrative, technical and strategic. The Project Manager will act as the link between every responsible figure (WP Leader, partner representative, etc) for carrying out the co-ordination activities in these day-to-day levels.

There is a number of Progress Reports to be submitted to the EC by the PM and approved by the Project Board, after compilation and processing of relevant information from WP leaders. Management Meetings, at which overall project status, both technical and administrative, is reviewed, are crucial for the success of RTD projects for maintaining solid relationships, promote information exchange and make agreements and major decisions.

**Deliverables (brief description) and month of delivery**

<table>
<thead>
<tr>
<th>Deliverable</th>
<th>Month of Delivery</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1.x. Contractual Periodic Progress Report</td>
<td>t7, t13, t19, t25, t31, t37</td>
</tr>
<tr>
<td>D1.7, D1.8 and D1.9. Annual Public Reports</td>
<td>t13, t25, t37</td>
</tr>
<tr>
<td>D1.10. Final Report</td>
<td>t36</td>
</tr>
</tbody>
</table>

**B1.4.2. WP2**

<table>
<thead>
<tr>
<th>Work package number</th>
<th>WP2</th>
<th>Start date or starting event:</th>
<th>M1_____</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>Dissemination and Exploitation: Legal Issues and Licensing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity type²</td>
<td>MGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant number</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Participant short name</td>
<td>UNT</td>
<td>ILT</td>
<td>SE</td>
</tr>
<tr>
<td>Person-months per participant</td>
<td>6</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

**Objectives**

The objectives of this workpackage are the following:

- **Licensing and Legal Issues:** this aims to sort out the legal and IPR issues before making LR available.

This requires working out contractual relationships preferably through some simple and generic contracts (distribution agreements as well as user agreements) Particular agreements will be sought with international organizations that could supply multilingual resources.

Bearing all this in mind, this workpackage covers the definition of requirements related to the

²Please indicate one activity per work package: SUPP = Support; MGT = Management of the consortium.
resource production and technology integration aimed at within the project, as well as the establishment of "standard" licenses and model agreements, adapted to the LR providers and the LRs users. This will be well defined within the establishment of clear legal frameworks and a data/resource sharing/reuse procedure.

- **Exploitation**: Exploitation of the platform (open source distribution of the platform) and the services based on the platform productions of LRs (resources on demand).
- **Dissemination**: This will be a key activity within PANACOTA as the project aims at sharing the experiences that the project will bring about, promoting the need for research and development in the field of Automatic LR acquisition and production and demonstrating the existence and benefits of using these technologies in the framework of PANACOTA.

**Description of work (possibly broken down into tasks) and role of partners**

WP1.1 Exploitation: Definition of a business model for the platform, as well as of a legal framework to cover negotiation of rights and IPR issues regarding both the data and tools to be used within the platform: In order to secure the long-term usability of the platform it is of major importance to establish an exploitation plan according to the new type of LR production procedure envisaged. This should focus on aspects such as long-term maintainability, updating, derived costs to be covered, etc.

WP1.2 Legal frameworks: In order to establish an adequate legal framework, the requirements related to data sources (concerning both data and tools) for the different components and their integration should be identified. The legal framework should be then defined accordingly, considering issues such as licensing, LR production within the terms of the platform, using of available data as well as sharing of produced resources. All copyright, IPR issues should be determined and cleared so as to allow a transparent use of resources. For that purpose, standard licenses will be defined for the different scenarios.

WP1.3 Dissemination: A dissemination plan will be established within PANACOTA comprising a series of dissemination materials and activities that will support the sharing and exchange of the acquired know-how with the rest of the community, both industrials and academics. For that purpose, the following actions will be taken:
- Production, maintenance and updating of TIC dissemination material: Web page and newsletters. The updating and correct maintenance of the webpage will be considered crucial as to become a reference for the topic of automatic acquisition of LRs. Periodic press-releases, and presentations in conferences and congresses. After each integration cycle, and in order to highlight new components and new features of PANACOTA, specific papers will be produced.
- Organization of Dissemination Workshops. Its organization will be detailed in the dissemination plan, but we plan to have at least 1 workshop collocated with congresses or conferences to maximize its impact, as well as a Technology Transfer Day, so as to introduce the platform to a wide audience of HLT experts and industrials.

**Deliverables (brief description) and month of delivery**

D2.1. (t6) Dissemination Plan, M1
D2.2.n (t6) Web page (plus monthly updates: content, news, etc) , M1
D2.3.n (t13-t24-t32-t36) Project brochures and dissemination materials, M4, M7, M10 and M11.
D2.4 (t30 ) Platform software licensing (open source) + Project Tools
D2.5 (t32 ) Mini-catalogue of LR produced under the project (with specifications, identification, production + LR usage and distribution licensing issues)
B1.4.3. WP3

<table>
<thead>
<tr>
<th>Work package number</th>
<th>WP3</th>
<th>Start date or starting event:</th>
<th>M1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>The Platform</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity type ³</td>
<td>RTD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant number</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Participant short name</td>
<td>UNT</td>
<td>ILT</td>
<td>SE</td>
</tr>
<tr>
<td>Person-months per participant</td>
<td>18</td>
<td>18</td>
<td>-</td>
</tr>
</tbody>
</table>

Objectives
Development of a platform for the easy integration of a variety of software, tools and methodologies to support the automation of acquisition, processing and annotation of language resources.

Description of work (possibly broken down into tasks) and role of partners
The platform will be the means of functionally integrating a variety of components that must interchange data according to defined standards and protocols and the managing of the platform will be done with a workflow editor. The tasks to be done are related to the wrapping of different components as web services and their effective use for the production of LRs. The work to be done will be divided into the following Tasks:

WP3.1 Deployment of web services as required by WP4 to WP6.
The integration work will be done by first wrapping components to be integrated as web services, ensuring that these web services are properly described and declared in the PANACOTA registry of services and testing their functionalities.

WP3.2 Dedicated Work Flow editor, engine and associated processes and machinery: Registry, Standards for input/output definition, temporal files, journaling, etc.
In order to give the users a user-friendly managing tool for the composition of their preferred combinations of components, a dedicated Workflow editor will be developed and delivered.

WP3.3 Studies on grid structure for the handling of massive data.
Study the capability of the platform as a distributed network

Deliverables (brief description) and month of delivery
D3.1 (t6) Requirement analysis of the platform: architecture and design, M1
D3.2.1. (t12) First version (v1) of the integrated platform ready for evaluation (First Web service integration, feasibility), M3
D3.2.2 (t14) First version (v1) of the integrated platform and documentation (First Web service integration, feasibility), M4
D3.3.1 (t20) Second version (v2) of the integrated platform ready for evaluation (v1 + WF description and big data management), M6
D3.3.2 (t22) Second version (v2) of the integrated platform and documentation (v1 + WF description and big data management), M7
D3.4.1 (t30) Final version of the integrated platform ready for evaluation (v2 + new components + WF editor), M9
D3.4.2 (t32) Third version of the integrated platform and documentation (v2 + new

³Please indicate one activity per work package: SUPP = Support; MGT = Management of the consortium.
components + WF editor), M10

B1.4.4. WP4

<table>
<thead>
<tr>
<th>Work package number</th>
<th>WP4</th>
<th>Start date or starting event:</th>
<th>M1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>Corpus Acquisition and Annotation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity type(^4)</td>
<td>RTD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant number</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Participant short name</td>
<td>UNT</td>
<td>ILT</td>
<td>SE</td>
</tr>
<tr>
<td>Person-months per participant</td>
<td>6</td>
<td>6</td>
<td>15</td>
</tr>
</tbody>
</table>

**Objectives**

Development of a Corpus Acquisition and Annotation (CAA) subsystem that will include:
- A component for corpus acquisition (CAC) from a variety of sources
- A component for cleanup and normalization (CNC) of corpus data
- A text processing component (TPC) comprising adapted annotation technologies for the automatic shallow processing of the acquired textual data in which languages.

**Description of work (possibly broken down into tasks) and role of partners**

**WP4.1. Corpus Acquisition Component**

As our first source for the compilation of the lexical resources targeted by the project, we plan to use corpora already available to consortium. Nevertheless, the largest part of the data needed for building the LRs will be extracted from crawled web pages with rich textual content. In the development of the Corpus Acquisition Component, we will use and adapt an efficient and distributed web crawling methodology that will collect html pages from selected sites known to contain texts belonging to specific languages, geographical regions and/or predefined domains and genres (e.g. multilingual versions of wikis, blogs, EU announcements etc.). Sites with content available in more than one language will be used for building parallel corpora collections. The schedule for the crawler runs will be based on the site renewal rate and availability of past pages. The webpage metadata (if available) and site structure will be used to index html pages according to language, domain, date or other available data. The development of means to query private archives will also be carried out.

**WP4.2. Clean-up and normalization Component**

Web pages typically contain “noise” including site navigation links, advertisements, disclaimers repeated on several pages etc. The Clean-up and Normalization Component will deal with removing such irrelevant information, as well as filtering out duplicates and too small or too large documents. The text resulting from html pages in different languages will also be filtered for extreme differences in size from one language to another. Shallow automatic corrections will be done for common html text errors like erroneous use of different scripts in the same word, unclean encoding etc. Texts will be converted to UTF-8, while minimal structure will be automatically inserted in each text, including tags representing header, paragraph and sentence limits.

\(^4\)Please indicate one activity per work package: SUPP = Support; MGT = Management of the consortium.
WP4.3. Text Processing Component
This task involves developing a Text Processing Component (TPC) that, following CAC and CNC, will deal with the processing of the automatically acquired and normalized corpora. Partners will at first adapt and prepare existing tools and training resources for the languages addressed by the project. Available lingware in the consortium and other open source tools will be used for this subtask, including state-of-the-art tools performing, at a minimum, sentence splitting, POS tagging, lemmatization, chunking and named entity recognition. We will use scalable tools that will be able to efficiently process the large amounts of data expected from CAC. We will explore availability and usefulness of a) open source to tools performing the functionalities mentioned above, and b) other analyzers such as dependency parsers and tools for co-reference resolution.

Deliverables (brief description) and month of delivery
D4.1 (t6): Report describing the inventory of corpus compilation tools to be developed and integrated in CAC, M1
D4.2 (t6): Report describing the inventory of corpus cleanup and normalization tools to be developed and integrated in CNC, M1
D4.3 (t6): Report describing the inventory of annotation tools to be developed and integrated in TPC, M1
D4.4.1 (T10): Initial functional prototype of the CAA subsystem and its components ready for integration (CAC, CNC and TPC), M2
D4.4.2 (T14): Initial functional prototype and documentation describing the initial CAA subsystem and its components, M4
D4.4.3 (t18): Revised prototype of the CAA subsystem and its components ready for integration, M5
D4.4.4 (t24): Revised prototype of the CAA subsystem and its components and documentation, M7
D4.4.5 (t28): Final prototype of the CAA subsystem and its components for integration, M8
D4.4.6 (t32): Final prototype and documentation describing the CAA Subsystem and its components, M10

B1.4.5. WP5

<table>
<thead>
<tr>
<th>Work package number</th>
<th>WP5</th>
<th>Start date or starting event:</th>
<th>M1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>Parallel corpus and derivatives</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity type</td>
<td>RTD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant number</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Participant short name</td>
<td>UNT</td>
<td>ILT</td>
<td>SE</td>
</tr>
<tr>
<td>Person-months per participant</td>
<td>-</td>
<td>-</td>
<td>18</td>
</tr>
</tbody>
</table>

Objectives
- developing word- and chunk-aligned data from the parallel corpora induced in WP4
- using this sub-sentential data to seed a range of MT systems (SMT, EBMT, Hybrid) developed at DCU
- evaluating the suitability of the induced parallel corpora as training material for a range of MT systems

5Please indicate one activity per work package: SUPP = Support; MGT = Management of the consortium.
Description of work (possibly broken down into tasks) and role of partners

In this WP the development of new components for the production of aligned corpus, annotated aligned corpus, bilingual dictionaries and transfer grammars will be carried out. The Tasks will be the followings:

**WP5.1. Aligners**

Once the parallel data has been found, we will sententially align the data in preparation for the bilingual dictionary and translation induction phases in WP5.2 and WP5.3 respectively to derive the required subsentential data.

**WP5.2. Bilingual Dictionary Induction Technologies**

From the sententially aligned data in WP5.1, we will use open-source (GiZAs++, Och and Ney, 2003) and inhouse tools (e.g. (Ma et al., 2007)) to develop a range of word-aligned data for input into the chuncking phase of the translation process.

**WP5.3. Transfer Grammar Induction Technologies**

From the word-aligned data in WP5.2, we will use open-source (GiZAs++, Och and Ney, 2003) and in-house tools (e.g. (Gough and Way, 2004); (Groves and Way, 2005)) to develop a range of phrase (or chunk) aligned data for input into the training phase of the translation process. Transfer rules from bilingual dictionaries will also be produced to feed transfer grammars of rule based MT.

**WP5.4 Seeding MT Systems with Derived Sub-sentential Data**

From the word-aligned data in WP5.2 and the phrase-aligned data in WP5.3, we will seed the MaTrEx system (Tinsley et al. 2007) with this automatically induced training data to derive output translations for a baseline SMT system (MaTrEx offers a wrapper around Moses for this task), an EBMT system (using the marker and edit-distance-style chunks cited above) and a hybrid system containing a number of variants of these subsentential resources.

Deliverables (brief description) and month of delivery

D5.1. (T6): Report describing the inventory of parallel technologies tools to be developed and integrated in PCT, M1

D5.2.1. (t10) Aligners to be integrated into the platform, first version, M2

D5.2.2. (t14) Aligners integrated into the platform and documentation, and set of parallel texts (49Mw) cleaned for training/building translation models, M4

D5.3.1. (t18) First version of the Component for the Bilingual Dictionary Extractor for integration, M5

D5.3.2. (t22) Integrated version of the Bilingual Dictionary Extractor and documentation, M7

D5.4.1 (t18) First version of the Component for the Transfer Rules producer for integration, M5

D5.4.2. (t22) Integrated version of the Transfer Rules producer and documentation, M7

D5.5.1. (t28) Final version of the Bilingual Dictionary Extractor for integration, M8

D5.5.2 (t32) Final version of the Bilingual Dictionary Extractor integrated and documentation, and sample of bilingual dictionaries produced, M10

D5.6.1 (t28) Final version of the Transfer Rules producer for integration, M7

D5.6.2 (t32) Final version of the Transfer Rules producer integrated and documentation and sample of transfer rules produced M10

D5.7. (t36) Validated bilingual dictionary for 1 language pair M11
B1.4.6. WP6

<table>
<thead>
<tr>
<th>Work package number</th>
<th>WP6</th>
<th>Start date or starting event:</th>
<th>M1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>Lexical Acquisition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity type</td>
<td>RTD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant number</td>
<td>1 2 3 4 5 6 7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant short name</td>
<td>UNT ILT SE UUK GLE UoA ALR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person-months per participant</td>
<td>30 16 6 24 - - -</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Objectives**
Development of techniques for automatic acquisition of subcategorization frames, selectional preferences, multiword expressions and lexical-semantic classes. Building on existing techniques, the objective is to improve their accuracy, scalability and portability between different domains, and adapt and apply them to the full range of languages and language-specific corpora which we propose to deal with.

**Description of work (possibly broken down into tasks) and role of partners**
This WP relies on the availability of annotated corpora which will be processed by different means that induce the information required. The tasks are the following:

**WP6.1. Induction mechanisms.** Subcategorization, selectional preference and multiword expression acquisition will be based on statistical induction methods that will have to focus on filtering and back-off methods. Research will be carried out on adapting existing tools into scalable, language independent and domain/genre/variation adaptive methods to be made available as stand-alone components to be deployed as a web service.

**WP6.2. Lexical-semantic classification methods.** Classification methods based on the identification of cues and properties related to different classes will be implemented and adjusted to maximise precision. Efforts will also be devoted to ensure scalable, language independent and domain/genre/variation adaptive methods to be made available as a stand-alone components to be deployed as web services.

**WP6.3. Merging of dictionaries.** Development of merging techniques for enlarging existing dictionaries based on lexical standards such as LMF, and methods to create updatable repositories with automatic techniques for merging/bootstrapping new resources.

**Deliverables (brief description) and month of delivery**
D6.1. (T6): Integrated report describing the inventory of technologies and tools to be developed and integrated in PANACOTA and criteria for evaluating its results, M1
D6.2.1. (T18): First version of the Components for Lexical Acquisition for integration, M5
D6.2.2. (T22): Integrated version of the Components for Lexical Acquisition and documentation, M7
D6.2.3. (T28): Final version of the Components for Lexical Acquisition for integration, M8.
D6.2.4. (T32): Integrated Final version of the Components for Lexical Acquisition and documentation, M10
D6.3.1. (T18): First version of the Components for Lexical Merging for integration, M5
D6.3.2. (T22): Integrated version of the Components for Lexical Merging and documentation.

---

Please indicate one activity per work package: SUPP = Support; MGT = Management of the consortium.
M7
D6.3.3. (T28): Final version of the Components for Lexical Merging for integration, M8.
D6.3.4. (T30): Integrated Final version of the Components for Lexical Merging and documentation, M10
D6.4. (T36): Monolingual lexicon for three languages tuned for a particular domain with morphosyntactic, syntactic and lexical-class semantic information, M11

B1.4.7. WP7

<table>
<thead>
<tr>
<th>Work package number</th>
<th>WP7</th>
<th>Start date or starting event:</th>
<th>M12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>Evaluation of integration and resources</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity type 7</td>
<td>RTD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant number</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Participant short name</td>
<td>UNT</td>
<td>ILT</td>
<td>SE</td>
</tr>
<tr>
<td>Person-months per participant</td>
<td>6</td>
<td>10</td>
<td>6</td>
</tr>
</tbody>
</table>

Objectives
In this Work Package we start from the information acquired in WP4, WP5 and WP6. The objective is twofold: on the one hand, it regards the validation of the integration of component as to form the desired platform; while, on the other, it will provide means to evaluate the acquired information.

Description of work (possibly broken down into tasks) and role of partners
This WP will be responsible for internal evaluation. The work package will face two main tasks:

**WP7.1. Evaluation of the Integration of components.**
Evaluation of the proper functioning of the integrated platform, the editor and processing of the different components together will be performed just after each integration cycle, and will verify improvements from version to version.

**WP7.2. Evaluation of the Resources produced**
It will involve evaluating the results obtained from WP4, WP5 and WP6 after each development cycle. The evaluation will be performed both from the specific task (by using gold standards, etc.) and final application (by incorporating the acquired information into a state-of-the-art MT system) points of view. Storing and structuring the acquired information into a unique repository will facilitate the evaluation of the different pieces by providing unified interfaces to the platform. As for evaluating Induced Parallel Corpora as Suitable Training Material for MT. The translations output by the MT systems built in WP5.4 using the sub-sententially aligned data derived in WP 5.2—5.3 will then be evaluated using a range of automatic MT evaluation metrics, as well as a human evaluation necessarily more limited in scope. These evaluations will then validate the training corpora that have been inferred across each metric and MT system type, providing important insights into the effectiveness of the techniques proposed in this project for automatic resource creation for MT.

---

Please indicate one activity per work package: SUPP = Support; MGT = Management of the consortium.
**Deliverables** (brief description) and month of delivery

D7.1 (t6) Analysis Requirements Integrated Report and Criteria for the evaluation of resources, technology and integration, M1
D7.2 (t14) First evaluation report: first components. M4
D7.3 (t22) Second evaluation report: component integration and intermediate LR’s. M7
D7.4 (t32) Third evaluation report: evaluation of final components, M10
D7.5 (t36) Final evaluation report on the evaluation of automatically produced LRs, M11

**B1.4.8. WP8**

<table>
<thead>
<tr>
<th>Work package number</th>
<th>WP8</th>
<th>Start date or starting event:</th>
<th>M18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>Evaluation in industrial environments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity type</td>
<td>RTD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant number</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Participant short name</td>
<td>UNT</td>
<td>ILT</td>
<td>SE</td>
</tr>
<tr>
<td>Person-months per participant</td>
<td>3</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**Objectives**

- Task oriented evaluation of the usefulness (e.g. cost reduction) of PANACOTA results judging by the adequacy of language resources and their multilingual interaction in the most dynamic scenarios.

**Description of work** (possibly broken down into tasks) and role of partners

The task consists in evaluating and demonstrating the capabilities of PANACOTA techniques directly and in the context of practical tasks. The evaluation will be done utilizing public domain tools which are available, and by manual efforts. The proceeding is based on the linguistic resources extended with information obtained via manual analysis of corpus data / system output by linguists and domain experts. The practical (task-based) evaluation will be done by using the automatically acquired lexical information (in the format and the granularity considered the most suitable) to see the improvement of practical MT application results using enriched and tuned lexical resources and by evaluating the added benefit for the application.

This will be carried out in two different aspects and subtasks:

- **8.1 Tool-oriented evaluation** of the PANACOTA functionality, using mixed automatic and manual evaluation procedures. Here, the single tools, and their combination in the platform processing chain will be evaluated for large scale industrial use.

- **8.2 Task-based evaluation** performed in real scenarios and against real tasks: commercial and industrial application measurable by evaluation in a commercial scenario for a Rule Based-Machine Translation system as well as in a pre-marketing scenario in a Statistical Machine Translation system.

**Deliverables** (brief description) and month of delivery

D8.1. (t6) Analysis of user requirements, M1
D8.2. (t36) Evaluation of PANACOTA tools and production chain, M11
D8.3. (t36) Evaluation of PANACOTA integrated into a commercial chain (MT), M11
B1.5. TOTAL EFFORTS

<table>
<thead>
<tr>
<th>Partic. no.</th>
<th>Partic. short name</th>
<th>WP1</th>
<th>WP2</th>
<th>WP3</th>
<th>WP4</th>
<th>WP5</th>
<th>WP6</th>
<th>WP7</th>
<th>Total person months</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UPI</td>
<td>12</td>
<td>6</td>
<td>18</td>
<td>6</td>
<td>0</td>
<td>30</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>CNR-ILC</td>
<td>3</td>
<td>3</td>
<td>18</td>
<td>6</td>
<td>0</td>
<td>16</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>ILSP</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>24</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>UC</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>12</td>
<td>6</td>
<td>0</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>Linguatec</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>12</td>
<td>6</td>
<td>0</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>DCU</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>10</td>
<td>6</td>
<td>3</td>
<td>48</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>ELDA</td>
<td>3</td>
<td>10</td>
<td>14</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>30</td>
<td>29</td>
<td>68</td>
<td>49</td>
<td>78</td>
<td>76</td>
<td>46</td>
<td>22</td>
</tr>
</tbody>
</table>

B1.6. Work Packages Dependencies:

B1.7. Significant risks and associated mitigation/contingency plans

The risks identified at global level have been listed in the table below and the response put in place by the PANACOTA Consortium. A further detailed risk analysis and contingency plans will be done during the first analysis phase of the project.
### B2. Section 2

#### B2.1. Management structure and procedures

Considering PANACOTA’s size, length, life cycle and funding scheme, a lean management organisation is considered to be adequate to guarantee a fluid and efficient communication between partners that continuously feeds relevant information into the decision making process and allows all members to fulfill their objectives. It basically consists of two levels: workpackage level, and project level as projected in the two dimensions of the management: technical and research work and day-to-day management. A schema of the envisaged organization is the following:

**Project Board**

The project will be managed by the Project Board. The Board has the managing control, and can decide about the project content and funding, in the framework provided by the EC contract. The board consists of one representative of each partner; each partner nominates a representative for the board, and a deputy. The representatives will be fixed at the kick-off meeting to take place within the first month of the project. The Project Board will be chaired by the Project Coordinator who will perform as a link between the EC and the Consortium. The Project Coordinator will be supported by the Project Manager, who will also be in charge of recording, distributing and implementing the resolutions of this committee.

**Workpackage organisation**

WPs in the proposal aim at being synergistic, with the outcomes of each informing the activities in others (as described in the detailed WP descriptions). Each workpackage has a workpackage leader. This person is responsible to design and implement the workpackage in the time (as in the schedule approved by the Project Board), cost (overseeing the appropriate use of finance budgets associated with each WP) and content frame (integration of activities and ensuring quality of assigned deliverables) as given in the Technical Annex or in the framework provided.

<table>
<thead>
<tr>
<th>Risk</th>
<th>Probability</th>
<th>Impact</th>
<th>Mitigation/Contingency Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web-service based workflows cannot handle the level of required details for the complete interoperability as an open platform</td>
<td>Medium</td>
<td>Medium</td>
<td>As a fall back position, UIMA can be also used as the object modeling platform for ensuring adequacy of input/output requirement. UIMA is an open source resource that is currently being used by a large number of developers of Language Technologies.</td>
</tr>
<tr>
<td>Difficult integration of legacy systems</td>
<td>High</td>
<td>Low</td>
<td>The project will create a think-tank force to apply the difficulties and propose general solutions.</td>
</tr>
<tr>
<td>Technologies that have been experimentally proven do not scale as expected</td>
<td>Medium</td>
<td>Medium</td>
<td>PANACEA is proposing a range of possible technologies in order to come up with more than one possible solution. Thus, the possible failure of one technology will not compromise the results of the whole project.</td>
</tr>
<tr>
<td>Semantics of different tagging sets make comparison and integration fall for some resources</td>
<td>Low</td>
<td>High</td>
<td>Some partners of PANACEA are directly involved in standardization activities and will be able to identify semantic problems and the solution proposed at the ISOGAT level.</td>
</tr>
<tr>
<td>Technology changes required redesign</td>
<td>Medium</td>
<td>Medium</td>
<td>The Project Board will monitor external developments and use them to re-plan if necessary.</td>
</tr>
<tr>
<td>Time for development is underestimated</td>
<td>Medium</td>
<td>Low</td>
<td>Project checkpoints will monitor, detect problems early and take corrective action. The PANACEA factory builds on existing language modules developed by the partners themselves or open source, so the development overhead for creating the PANACEA web services will be lower than developing everything from scratch.</td>
</tr>
<tr>
<td>PANACEA Methodology and standards are not accepted by potential users</td>
<td>Medium</td>
<td>High</td>
<td>Key efforts are being set up to define an industry driven strategy for the factory development. The PANACEA consortium already includes the world-leading sites carrying out distribution of LFs, so take-up by the industry will be ensured through wide dissemination via publications and web sites. There will be workshops aimed at ensuring take-up.</td>
</tr>
<tr>
<td>Partner leaves consortium</td>
<td>Low</td>
<td>Medium</td>
<td>Consortium is of sufficient strength and diversity for other partners to replace if required.</td>
</tr>
<tr>
<td>Staffing &amp; Recruitment Problems</td>
<td>Medium</td>
<td>Low</td>
<td>PANACEA technical excellence in given disciplines is shared by different key persons which can be seen as a risk balancing to avoid excessive dependency on any person.</td>
</tr>
</tbody>
</table>
WP meetings during the project and returning reports to the Project Manager as required. Of special relevance is their role to build a sense of common purpose among the WP partners which means, building the consortium into a coherent group oriented to seek overall benefits above the WP. WP leader also serves as a contact point for the other workpackage leaders in case multilateral issues need to be discussed. The workpackage leaders report to, and are coordinated by, the Project Coordinator that will act in this respect as a technical coordinator, by collecting, collating and assessing information for periodic reports from contributing partners according to guidelines provided by the Project Manager.

Management
The day-to-day activities are under the responsibility of the Project Manager, both on administrative and technical level. For optimizing the management, PANACOTA has foreseen that the administrative management will count on the support of the UNT central university research department, which is a professional structure made up of highly skilled experienced personnel. These specialists assist university coordinators with the administration of contractual, financial and administrative matters of the projects coordinated by UNT and count on relevant expertise on international collaboration and a solid technical background. The role of the Project Manager in this STREP will be providing a management structure that both helps and monitors the project development and administrative duties, thus becoming a crucial tool for its success.

Project Manager will thus take the day-to-day management, which implies: Monitoring the development process, coordinating the workpackages and supervising the workpackage leaders, and ensuring that the project meets its technical objectives. There will be one person nominated for each of the tasks.

The strategic Management, which is related with the final objectives of the project, will also be under the responsibility of the Project Board. The management will assure the success of the whole project looking forward the consecution of the expected results both in terms of quality, costs and time, and further dissemination and exploitation of these. This function also includes a basic Quality Assurance System for the project itself and their results. Strategic issues are the management of conflicts or possible changes or project re-orientation, as well. The Project Manager will inform the Project Board on the monitoring of the strategic issues, and will propose decisions to be taken by majority in case of necessity.

Partner representatives
Each partner will name persons responsible for organisational, financial and management issues.

Management Instruments
Management Instruments refer to communication and reporting, and to decision making.

Project Communication
Official internal project communication will be done by means of meetings, video-conferencing and reporting.

Meetings
The organisation of meetings will follow the management structure and comprise meetings on two levels:

- Board meetings will take place every semester. They will consist of reports of the coordinators and workpackage leaders, and decide on all relevant project issues: release of workpackage deliverables, preparation and follow-up of reviews, discussions with the EC and/or external partners, cost issues, reaction to latest developments of the field, participation / organisation of workshops and conferences, etc. Board meetings will be organised by the project manager, with a written agenda as part of the invitation, and minutes. Board meetings should be attended by the Board members, the coordinators and the workpackage leaders (or their deputies). Voting right is only with the Board members.

- Workpackage meetings and special workshops will be held if the need arises. They will be
organised by workpackage leaders (or their deputies). The results of these meetings will be documented in minutes.

**Reporting**
There will be strict reporting lines:
- Workpackage members will send **bimonthly** reports to the workpackage leaders, consolidated there and sent to the coordinators.
- Workpackage leaders will provide input for the **six-monthly report**, to be composed by the Project Manager as input for the Board review.
- There will be the **annual** report documenting the project progress to the interested parties outside of the project and the public.
Additional reports, for reviews and other occasions, will be provided on request.

**Decision Making**

**Form of decisions**
The usual form of decision making is during **meetings**. Issues to be decided must be announced in the invitation to such meetings, and the decision must be stated in written form in the minutes of the meeting, in order to be verifiable by anyone interested. In special cases, or in urgent cases, decision by **email** is also permitted if no partner objects. In such a case it is the task of the workpackage leader (on workpackage level) or the coordinator (on Board level) to organise this decision and count the votes. Like other decisions, email decisions should be documented in the minutes of the meeting following such decisions.

**Decision lines**
The line of decision making reflects the different levels of the project organisation.

- **Decisions on workpackage level**: Technical decisions within the scope of a workpackage are taken by the workpackage leader, after consulting all workpackage members. If partners of the workpackage do not agree, the decision is brought to the board. Also, if the scope of the decision extends beyond the scope of the workpackage, the board must be involved.

- **Decisions on Board level**: Decisions which concern the project as a whole are taken on Board level. This refers to priorities, content of the development, funding issues, formal contacts to other partners, and the like. Decisions are taken by the Board members by majority; in case a project member is involved (for failure to deliver etc.) with a majority of two thirds of the votes.

The Board can entrust the coordinators with the implementation of its decisions.

**Special Management Tasks**

There are special tasks which need to be followed on by the Board, assisted by the Project Manager:

**Management of IPR**
With the Intellectual Property Rights policy, the management will track the IP questions, and the foreground/background components emerging during the project. The list of IP-relevant components will be kept up-to-date by the leader of WP2, especially devoted to Licensing and Legal Issues.

**Risk Management**
It is the responsibility of the workpackage leaders to inform the Board on delays, problems in reaching the functionality, problems with availability of resources, and other factors jeopardising the project success and the proposal to overcome them. The Project Manager will register and oversee this process and will follow up key risks within integration of activities and will propose the strategies to minimize them.
It is the responsibility of the Board to take appropriate measures to keep the project on track. If this is not possible, the Board must negotiate a contract amendment with the EC.

**Special Task Forces**
In case unforeseen events occur (like an invitation to a conference presentation, or some development problem not foreseen so far), the Board can instantiate a special task force to
solve these problems. The board will decide on the composition and the funding of such task forces.

**Dissemination Management**
The Board will support the coordinator in dissemination activities, by identifying opportunities and releasing project resources for such efforts. Making the project results visible is considered to be a task of all Board members.

**Standardisation Management**
The Board will monitor and support standardisation activities, both in the area of linguistic standards and in the area of technologies, and entitle project members to act on behalf of the project as a whole.

**B2.2. Individual participants**
[Relevant publications by participants are specifically signalled at the References section]

**B2.2.1. University of New Technologies (Spain)**
The Institute for Applied Linguistics (IULA) of UNT, created in 1994, is a research and graduate training center that gets together 60 researchers. The members of IULA are professors of the Departments of Translation and Language Sciences, Information Technologies and Communication, PhD. students, research scholarship holders and researchers contracted for specific projects. The Language Resources Technologies group (Tecnologies dels Recursos Lingüístics, TRL) directed by Prof. X is especially devoted to the hybridization of techniques for the automatic development of linguistic resources for applications of Natural Language Processing and the standardization of data and processes. The TRL group is currently formed by 6 researchers (4 Computational Linguists and 2 Computer Engineers) and 2 granted Master students. The group participated in the project LIRICS: Linguistic Infrastructure for Interoperable Resources and Systems. The group is currently involved in the projects Common Language Resources and Technologies Infrastructure (CLARIN, co-funded by the 7FP of the EU (FP7-INFRASTRUCTURES-2007-1-212230) where they are leading Working Groups related to the provision of a webservice-based infrastructure and the Spanish Ministerio de Educacion y Ciencia (CAC-2007-23) and Ministerio de Ciencia e Innovacion (ICTS-2008-11), and Flarenet: Fostering Language Resources Network, funded by the econtentplus EU program: ECP-2007-LANG-617001.

UNT will be coordinating PANACOTA and mainly involved in the development of the WP3, The Platform and WP6 Lexical Acquisition. For this last topic, the group developed two national projects:

Key personnel

**Bella X.** (PhD. Universitat de Barcelona, 1993, joined the IULA in November 2003 where she currently is Head of Research for the group Tecnologies dels Recursos Lingüístics (Language Resources Technologies). Before, and since 1993, she was technical director of the Grup d'investigacio en linguistica computacional at the Universitat de Barcelona (gilcUB). Her current lines of research are related to Natural Language Processing tools for the automatic acquisition of lexical information, area in which she has several international publications. She has actively participated in more than 10 EU and national funded research projects related to LRs such as: MULTEXT (LRE62050), LSGRAM (LRE61029), PAROLE (LE2-4017), as Technical Manager in the project SIMPLE (LE4-8346), and LIRICS (econtent 22236) where she co-authored the Lexical Markup Model (LMF). Dr. X has also been involved in the standardization of LR's participating in ISO TC37/SC4 activities in LRs, group in which she is now national representative and designated expert by the Spanish AENOR. She has also directed the Spanish project AAILE HUM2004-05111-C02-01/FILO and HUM2007- 61067/FILO, on automatic lexical acquisition for the last 4 years.

**Maria V.** (1990, BA in English Philology, Universitat de Barcelona. 1992, Master in Computational Linguistics, Universitat de Barcelona. 1998. PhD in the program "Formalizacion del lenguaje natural", Universitat Politecnica de Catalunya). She is Senior Researcher at IULA since 2004. Before she was researcher at the Grup d’investigacio en Linguistica Computacional, Universitat de Barcelona (gilcub), where she started as researcher in 1993. She
has been involved in different European projects related to LRs such as LE-PAROLE, LESIMPLE, IST-PEKING and in standardization projects: EAGLES, ISLE.

B2.2.2. ILT – Institute of Linguistic Technology (Italy)

The ILT is active in the field of Computational Linguistics since many years and has a distinguished international and national leadership in the area of language resources for NLP whereby it has operated under the aegis of EC, European Council, ESF and United Nations. Its mission is to improve and foster language technologies through new methods and techniques for managing digital content and understanding human language, in view of promoting Italian in the multilingual information society. It promotes basic research in areas where the need of significant innovations emerges, fostering the synergies among different disciplinary competences, and ensuring synergies between basic and applied research. It has been often the leader of the international community towards innovative objectives, by promoting new "paradigms" in the field. In synergy with European initiatives, it is the pioneer of the vision of an open and distributed infrastructure of resources and tools, to be integrated in various services and systems, and has launched the notion of language resources as the central component of the linguistic infrastructure. ILT has designed and built several innovative language resources: corpora, treebanks, computational lexicons (WordNet, SIMPLE) with the respective ontologies, semantic networks, grammars, and has defined international standards for language resources. It has coordinated the major initiatives relating to language resources and standardisation in Europe and worldwide. It has developed a broad range of language technologies: a complete suite of tools for robust processing of Italian; tools for: knowledge acquisition from corpora, machine learning, word-sense disambiguation, mono- and multi-lingual terminology extraction, ontology acquisition and structuring, named entity recognition; multimedia tools for language learning. Such language technologies are widely used to tackle content management, to transform textual documents into structured digital knowledge, and are applied in several application domains, ranging from filtering of Web documents, knowledge management, e-government and eparticipation, e-learning, machine translation, intelligence, humanities, e-commerce, semantic web, data mining, e-health, domotics, biomedical informatics. The development of standards is another of the missions of CNR-ILC, both for research in multilingual content processing and for supporting the production of linguistic resources for commercial/industrial use. CNR-ILC has participated to important EU projects on standardisation (e.g. LIRICS and INTERA) and is actively involved within the International Standardisation Organization, ISO TC37 SC4, whose aim is the definition of standards for linguistic resources for language engineering.

Key Personnel

N.N.Z., Director of Research at ILT, has a long-lasting experience in the coordination of international, European and national projects as well as strategic initiatives. She has been Vice-President of the ALR Board and chair of the ALR, acts as member and General Secretary of ICCL and is founding member of the Italian Forum for HLT at the Ministry of Communications, Convenor of the ISO/TC 37/SC4WG4, member of many International Committees and Advisory Boards, Conference Chair of LREC, General Conference Chair for COLING-ACL2006, chief co-editor of the International Journal Language Resources and Evaluation and has more than 300 publications.

B2.2.3. SE – Speech Expert Dept., University of Greece (Greece)

The SE is a scientific and technological organisation, a legal entity governed by private law, which is under the auspices of the General Secretariat for Research and Technology (Ministry of Development). It comprises five research Institutes, one of which is the SE. Its activities are organized in 5 scientific departments: Speech Technology, Machine Translation, Electronic Lexicography, Educational Technology, and Language Technology Applications. The experience of its researchers, the close relations it holds with key research centres in other European countries and its industrial orientation are basic elements in the profile of SE; it has successfully participated in more than 150 R&D projects. The work undertaken by SE in this project is to be conducted by the Natural Language and Knowledge Engineering (NLKE) Department. The NLKE Dept conducts basic and applied research in the fields of Natural Language Processing and Knowledge Technologies. Its main aim is to design computational models for natural
language recognition and "understanding" as well as knowledge extraction and representation. In particular, the department designs, implements and integrates human language technologies in systems and applications dealing with structured data as well as with unstructured data processing. Applications include language-aware information retrieval and extraction, natural language interfaces, text mining, knowledge management, etc. The core of the approach adopted consists in developing, improving and valorising methods and techniques in the areas of artificial intelligence, machine learning, statistical and rule-based processing as well as hybrid combinations in an attempt to propose solutions that are flexible and adaptive to a range of linguistic phenomena and applications. Multilinguality is catered for by resorting to processing of parallel and comparable text corpora and elicitation of multilevel translation equivalents.

Key personnel:

P. S. is the Head of SE. He was the Project Scientific Coordinator of the Translearn/LRE, Musa/IST and and Revealing This/IST projects, and has been leading SE’s participation in more than 20 R&D projects in the areas of mono/multilingual and multimedia information processing. He is the President of the European Language Resources Association, member of the LREC Programme Committee, national scientific coordinator for the Clarin Research Infrastructure. His research interests include statistical and deductive methods in natural language processing and understanding, language resources and automatic linguistic knowledge elicitation, machine translation and philosophy of language. He also teaches postgraduate courses on Language Technology and Logic at the National Technical University of Athens. He has published more than 100 articles in international scientific books, journals and conference proceedings, while he has given several invited tutorial and keynote speeches.

G.M. is a senior researcher at the SE. She is responsible for the Language and Lexical Resources Unit of the Department. A researcher on computational linguistics since 1988, she has been coordinating SE’s work in several EU and national projects on corpora and lexica. She was also the scientific coordinator of the construction of the “state-stamped” dictionary for Greek as L1 used at the country’s secondary schools and she has participated in the development of a course book for the teaching of Greek as L1. Her research interests are language resources (written corpora and lexica) and metadata.

B2.2.4. UUK - University of UK (Great Britain)

The language department of UUK is an interdisciplinary research centre which acts as a links between different language-based departments at the University of the UK: Linguistics, Engineering, Computer Laboratory and Experimental Psychology. Its 10 permanent staff run a master’s course in English and Applied Linguistics and two dedicated PhD tracks: English and Applied Linguistics, and Computation, Cognition and Language. Assisted by over 40 (post-)doctoral researchers at UUK, the staff conduct research in computational linguistics, linguistics, language acquisition, learning theory, language processing, typology, and pragmatics in the context of a number of UK (e.g. AHRC, the British Academy, ESRC, EPSRC) and EU-funded projects, many of which are joint with other departments in and outside of the UK.

The Computation, Cognition and Language (CCL) group has close links with the Computer Laboratory, with two of its computational linguists holding a joint appointment in the two departments. The Computer Laboratory has been a centre for work on NLP and information retrieval/extraction for 40 years. Its Natural Language and Information Processing (NLIP) group co-teaches a specialized graduate course on speech and language processing and has developed a range of state-of-the-art, widely-distributed text processing tools which are reliably usable for large-scale processing while incorporating insights from theoretical linguistics. The tools for parsing and lexical acquisition are developed and maintained in close collaboration with the CCL group. The CCL and NLIP groups also share teaching and supervision of students and post-docs, and conduct joint research on a wide range of areas (e.g. parsing, lexical acquisition, word sense disambiguation, text classification, information extraction, summarisation, NLPI for e-Science) in the context of several research projects.

Key Personnel

Dr. J.K. holds a PhD in Computer Science from the University of UK, Computer Laboratory. She has seven years of post-doctoral experience in Natural Language Processing (NLP) at research institutions in the UK, USA and Japan, and is currently holding a
Royal Society University Research Fellowship. She has worked on syntactic and semantic analysis of texts, and has made major contributions to automatic acquisition of lexical information from corpora.

**L.L.** is a senior research associate who holds a PhD in Computer Science from the University of UK. Her research focuses on (psycho-)computational models of language acquisition and in automated corpus analysis (grammatical and lexical acquisition from corpora). She is collaborating with the Computer Laboratory, Cambridge University Press, University of UK Local Examinations and iLexIR (a language processing technology provider and consultancy) in the context of e.g. EPSRC-funded joint projects which she is helping to comanage. Buttery is a London Technology Network Business fellow and has worked for Dragon Systems, Google and Intel, obtaining industrial experience as a research engineer.

**B2.2.5. GLE – German Linguistic Experts (Germany)**

GLE is a private company specialised in the development and marketing of language technology solutions. The company was founded in 1992 and is based in Germany. GLE is an SME with about 25 employees. The company has built up significant technological expertise in the area of machine translation, speech recognition and speech synthesis as well as in content technology. GLE markets machine translation systems and Speech synthesis programs. The products support several languages, including French, Spanish, Italian, Czech, Polish and Chinese.

In the area of research and innovation, GLE is in a leading position as well. The translation product includes innovative solutions for transfer selection (called ‘neural transfer’), for error correction, and for automatic language and topic detection, and is heading towards a hybrid architecture of rule-based backbone and statistical / phrase-based enhancements. The company participated also in several research projects both on European and national level.

Main Researchers:

**Dr. H.I.T.** Computational Linguist. Member of GLE development team since 2005. Main research areas: machine translation, terminology, multilingual search. Project manager of several MT and terminology products. Member and coordinator of several EU research projects in this area. Member of ALR Board.

**Dr. W.T.** Computational Scientist. Member of GLE development team since 2004. Member of several EU research projects. Main research areas: Integrating statistical components into language technology applications: Topic classification, correction tools, word sense disambiguation.

**B2.2.6. UoA – University of Argentina, Linguistic Department (Argentina)**

The proposed research project will look into developments carried out on the other side of the Atlantic. The Linguistic Department of UoA houses, amongst others, Centre for Next Generation Localisation. The UoA (LD) develops applications in Natural Language Processing (NLP), Computational Linguistics (CL) and Human Language Technology (HLT). The Centre involves 10 affiliated academics from the School of Computing, the School of Electronic Engineering and the School of Applied Languages and Intercultural Studies as well as 20 research post-graduate students.

UoA (LD)‘s mission is to revolutionize localisation via breakthroughs in automation, composition and integration, focusing on: Integrated machine translation technology, Speech-based interfaces and more personalised speech output, Multilingual digital content management for personalised multilingual content access and delivery, Localisation workflows and system integration.

The UoA provides a stimulating, world-class NLP research environment for the proposed research project.

Key personnel

**Prof. A-M S.** obtained his BSc (Hons) in 1986, MSc in 1989, and Ph.D. in 2001 from the University of Stanford. From 1988 to 1991 he worked at the University of Essex, UK, on the Eurotra MT project. He was promoted to Senior Lecturer in
2001 and Associate Professor in 2006. He is co-editor of the recently published Recent Advances in Example-Based Machine Translation, and editor of the 'Machine Translation'.

B2.2.7. Association for Language Resources (France)

The ALR was created in 1995 with the aim of identifying, collecting, classifying, validating and distributing LRs as well as collecting and disseminating general information related to the field of Human Language Technologies and evaluation. Anticipating the evolutions in HLT field, ALR has broadened its activities to cover multimedia/multimodal resources as well as evaluation activities, distributing the language resources needed for evaluation purposes, and coordinating evaluation campaigns. ALR will bring in its expertise in a number of activities planned within PANACOTA, such as the handling of licensing, ownership and copyright issues and the definition of a dissemination plan.

Key personnel

Dr. C.C. obtained an Electrical Engineering degree from Ecole Nationale de l'aviation civile, and Masters Degree and doctoral degrees in Computer sciences and Signal processing at the Ecole Nationale Superieure des Telecommunications (ENST) in Paris. He was a research scientist at the Signal Department of ENST, involved in Man-Machine Interaction. He has also consulted for several French companies. In 2001, he joined CAP GEMINI INNOVATION, R&D center of CAP SOGETI to work as the team leader on speech processing. He then moved to ACSYS in 2009 to take on the position of Speech technologies manager. Since then, he has been the Chief Executive Officer of ALR.

B2.3. The PANACOTA Consortium as a whole

PANACOTA has been conceived as a STREP because of the focus of its objectives, and the aim of addressing the problem of massive acquisition and annotation of language resources as required under the Call 4 Work Program. In order to tackle this problem with a reduced number of participants, a combination of expert research teams which can afford last generation technologies as well as specific expertise in the other areas as required by the workplan, i.e. standards, integration of components, etc. has been required. Besides, we have sought the participation of the industry, because the project is meant to be the proof of concept that the technical proposals and prototypes generated in PANACOTA are suitable for real environments. It is a SME because PANACOTA wants to demonstrate its tangible potential for contributing to the European industry in multilingualism which must count with a critical mass of SME’s to achieve success. An additional asset of the PANACOTA consortium is to have incorporated the European LR evaluation and distribution agency (ALR) with proven experience in the handling and distribution of LRs. Thus, our consortium has been able to assemble the partners for the success of PANACOTA.

PANACOTA partners have been selected among the best technological providers for delivering not only already available components for the production of LRs, but crucially to implement prototypes of new components that until this project were only proven experimental systems. Each Partner excels at least in one area of the ones addressed by the project and its specific skills in what respects PANACOTA are summarized in Table 4 at the end of this section. For the technical WP’s, SE is leading WP4 Corpus Acquisition and Annotation, UoA is leading WP5 Parallel corpus technologies and derivatives and UNT is leading WP6 on Lexical Acquisition Technologies. Nevertheless, all the partners have a proven experience in the different areas covered by the project, all of them with important publications in the most recognized conferences of the LRs and HLT areas.

B2.4. Resources to be committed

PANACOTA major costs will be on labour, with no other major costs budgeted. The only issue that could be significant is the use of the specific grid computing facilities that are however afforded by partners with no costs because their use is for research purposes.
PANACOTA project counts on a human labour force of 398 persons/month, which amounts approximately to 33 persons/year.

As it can be seen in the figure above, most of the resources, about a 70% are devoted to Development tasks, the main concern of PANACOTA. Management and Dissemination and the dissemination task including the creation of Licences and a Business model amount for the 13% of the resources in order to guarantee an effective coordination and a maximization of the results and impacts of the project. Last but not least, a 16% of the resources of the project will be devoted to Validation and Evaluation of results.

PANACOTA counts in addition to the EU required funding, with the investment of resources of its partners which will be affording experienced researchers and professors of well known curriculum that will lead the project to success.

Subcontracting foreseen in the project is in order to cover audits in most of cases and to be assisted for legal advice in the questions related to IPRs, legal issues and privacy considerations in WP2, this is the reason for being allocated in partner ALR, leader of WP2.

UNT will bring to the project its well-equipped laboratories.

B3. Section 3

B3.1. Expected impacts listed in the work programme

PANACOTA will address the wider objectives and desired impact of the FP7 Work Programme and general Challenge 2, as well as those aspects related to the more technical impact expressed in the Language Interaction Call 4 Guidance Notes in the line b4) Methods for automated acquisition and annotation of language resources as we will now explain.

Despite the complexity of handling its 23 languages, the European Union has decided that cultural and language differences need to be preserved not to lose people’s identity. Europe needs to find other means – such as technological ones – to overcome the language barriers to support citizens and industry in a globalised world. Thus, the ICT Work-Programme for 2009-2010 remarks under its Challenges that “With 23 official languages, the EU is at the forefront of multilingualism and it would be unrealistic to assume that the lingua franca in machine translation is, or will remain, English”. Machine Translation systems are thus expected to have a high impact in the managing of multilingualism in Europe. Machine Translation has to reduce language diversity shortcomings and difficulties in interaction among citizens at all levels: commercial, political and personal, and freeing the access and dissemination of information by translating it into citizens own language. Note that the problems to be solved by Machine Translation in these areas are not only economic, i.e. that Machine Translation is to make translation cheaper (even free, as might be the case), but to make materially possible the translation of the huge quantity of texts (written or oral) that are produced in the Union and to cover the necessities of hundreds of millions of citizens in the required timeframe. However, the state-of-the art concerning MT is said to be (cf. Call 4 Guidance Notes. Language based Interaction, pag .1):
- The overall quality of MT is still not adequate for many practical purposes
- Many Languages still lack MT facilities of sufficient quality and coverage

We will see now how PANACOTA will contribute to transform these crucial problems and to get the desired impact of Machine Translation in Europe and what of its actions will bring about these contributions.

As justified in Section 1, the “resources bottleneck” is compromising the quality, and then the full
coverage in terms of languages of MT systems. Thus, PANACOTA has to contribute to improve the quality of MT systems by provoking a breakthrough for the current situation with respect to the acquisition and production of LR's. The objective of our project is to develop a platform of interoperable services that is used like a Language Resource Factory. Our factory covers, step by step, the automation of all the stages involved in the acquisition, production, updating and maintenance of language resources required by MT systems, and other Xlingual applications, and in the time required. Automation will dramatically cut the economic and temporal costs of keeping a continuous production of LR and will deliver both quantitatively and qualitatively better resources thus directly contributing to improve the quality of the MT.

In order to maximize the impact of interoperability related issues, we will profit the possibility of a mutual contribution and provision of resources and infrastructures with other projects of the 7FP currently developing networks of web services and related elements. Among them, Incentives for semantics (INSEMTIVES), Responsive open learning environments (ROLE), Sharing open source software middleware to improve industry competitiveness in the embedded systems domain (SHARE); Fast and advanced storyboard tools (FAST); Service oriented architectures for All (SOA4ALL); Semantically-enabled heterogeneous service architecture and platforms engineering (SHAPE); and very specially with the project Common Language Resources and Technologies Infrastructure (CLARIN), because of its related object. A quantitative objective of PANACOTA that can afford a clear idea of the impact of the project main results is that at the end of the project, users of PANACOTA will be able to reduce the actual costs of its formerly manually building of a domain tuned language resource by, at least, 25% as the outcome of WP8, the evaluation carried in an industrial scenario for the two most representative types of MT system, will confirm. This reduction of costs and time in the supply of LR's is the only way to guarantee sustainability for continuous updated versions of particular MT systems, particular domain adaptation and creation of new pairs of languages systems. Thus the impact of PANACOTA will also be quantifiable in terms of the time of production of new languages, or new domains for new languages also contributing to the actual achievement of the Call impact related to the lack of coverage of MT systems in terms of languages and domains.

But PANACOTA aims at impacting not only within the specific commercial and academic goals of the partners, but also within the European research and industrial community. As such, an emphasis will be placed on the development of clear and consistent lines of communication beyond the project itself, crucial to the ongoing success of automatic translation as a whole. The first step to contribute to bring about these impacts will be done during the life of the project. We plan to achieve these goals by having particular dissemination actions which will be set up as to make the community aware of the feasibility and advantages of the automatic production of LR's for MT producers. This will be done in WP1, specifically focused on the definition of the strategic dissemination and exploitation plan and its realization. As a second step, we will set up the effort in dissemination according to the importance of reaching not only developers of MT systems, which is the main target of the current call, but also other meaningful providers of Language Technology-based applications that also require Language Resources thus maximizing the impact of our project.

Here follows a list of other highly innovative and technologically demanding commercial systems that could benefit of automatic mono and bilingual language resources acquisition capabilities as the ones provided by our project maximizing its impact:

**Crosslingual Information Extraction.** It typically works with declarative patterns that look for entities –typically nouns- related by an activity or event which is expressed by a verb. To work properly, these verbs must be described in terms of their valency or subcategorization frame. But, the current commercial products using this technology are, either tied to a particular domain in which they are pretty performant, but with high costs for adaptation to other domains (for instance Spirit Collector, [http://www.spiritengine.com/](http://www.spiritengine.com/)), or offers specific tools for customers to build the required resources adapted to their own domain (for instance SRA’s NETOwl, [http://www.netowl.com](http://www.netowl.com)).

**Multilingual Automatic document indexing.** Although the most basic samples of this application have been working with language independent technologies, the demand for more
and more sophisticated tools for information search has promoted the introduction of linguistic technologies that need lexica. One of these technologies is Term Extraction algorithms that need to process text, and thus require information on the words. The most innovative application of automatic document indexing techniques is the automatic creation of metadata for the Semantic Web. Commercial tools such as Connexor Machinese Metadata (http://www.conexor.com) include means to use lexical data, as well as ontological data, of the final user to permit the correct use of its linguistic technological components.

Question Answering. Technologically related with Information Extraction technologies, also rely on the identification of the content by means of sophisticated tools that rely on lexical information. One of the recently put into the market applications for Question Answering that directly tackles the problem of quick adaptation to a new domain is Hakia http://www.hakia.com/. It acknowledges the use of an ontological information acquisition module that considerably reduces the amount of work related to the adaptation to new domains. This experience supports our decision of working on automatic acquisition modules for lexical information acquisition that can be of help to a broad range of systems.

Natural Language Interfaces The development of natural languages interfaces has been dominated by speech recognition tools that work with previously computed language models. With the success of call-centre applications, the market has been demanding more powerful systems that face, however the problem of finding language resources, i.e. large monolingual corpora, for an accurate modelling of the language used.

B3.2. Dissemination and/or exploitation of project results, and management of intellectual property

Dissemination will be a key activity within PANACOTA, as the project aims at sharing the experiences that the project development will bring about, promoting the need for research and development in the field of automatic language resources acquisition and production and demonstrating the existence and usefulness of technologies for the automatic acquisition of language resources.

PANACOTA will therefore count on a project-wide dissemination strategy which complements the specific actions that each of the partners will take as part of their internal strategies and interests and makes use of their positioning within relevant target groups. This strategy starts with an initial step of identification, where the following key issues will need to be defined:

- The major points to be shared
- The major target audiences to be reached,
- The design of specific internal procedures, activities and materials to meet the targets
And its follow-up and evaluation during the project execution.

Target public for the dissemination of PANACOTA are:
- Major target audience are industrial stakeholders in the fields described in the previous section, with special attention but not limited to:
- MT providers
- Language Technology based applications developers
- The scientific community, with special attention but not limited to:
- Academic LT community
- Data owners, with special attention to proliferating digital libraries and archives.
- Standardisation committees, with special attention to those related to innovation emerging from ICT
- General European audience
- International audience

Specific actions are to be planned and executed for the abovementioned stakeholders in WP6. As an example of the actions, a broad information campaign will be carried out so that providers of MT but also other Language Technology based application providers to know of the existence of the offered web services and derivative products. On the other hand, dissemination among technology developers, also academic ones, is of paramount importance in order to
invite them to integrate their own tools as components of the web services factory. Information about these web services, the standards used and other technical details will be maintained in the project webpage. Direct contacts will be also carried out with data owners as described in Section 1.

The **main tools for dissemination** will be:

**An initial step will be the generation of a project identity for all public communications, including logo, presentation templates and generic project information.** The logo of the project will use the idea of the Rosetta Stone which was the first aligned corpora produced and kept by humanity. The logo is shown on the cover page.

**Project brochures** will be produced for publicising the demonstration workshops and individual invitations sent out to the target industry representatives. Moreover, brochures will be also done to disseminate the platform at public events attended by the project members as well as to distribute among the community, as sometimes done with communication Newsletters.

Creation and maintenance (with monthly updates) of a **project web page**. To support the above mentioned dissemination events the partners will also set up a public project website which will update the target public on the results of the PANACOTA project and will also provide information and links to the project factory once it is available from their own corporative areas.

**Publication of project results** throughout the duration of the project in widely accessible and, where available, openly accessible science and technology journals, as well as through conferences and through other channels, including the Web, that reach audiences beyond the academic community. This dissemination through internet will take into account copyright restrictions with regard to the origin of the publications. This aims to respect and take into account the fact that some publishing houses do not allow internet publication of the articles published under their supervision. Specific permissions will be sought from these organisms. Publication of periodic **press releases**, leaflets and other informative materials, with special attention to project kick-off, release of main demonstrators and project’s end and conclusions. These press-releases will be customized for specialized media and generic ones in order to increase general public’s awareness of progresses in EU funded research within the field.

Organization of **demonstration workshops** to show the viability of the results and its potential economic advantage. A minimum of one workshop and one Technology Transfer Day will be organised during the project lifetime, supported by individual partner event and conference participation. The focus of one of these events will be more on technology developers, and the other on industrials.

Participation in **EU Commission** or other **FP7 projects events** is also crucial in order to place PANACOTA’s aims and achievements in the international research picture and, thus, make the community aware of its achievements and approaches. In addition to information days, we have

**Presentation in conferences and congresses**

The project is expected to come out with new modules and technological advanced components for the gathering, annotation and induction of language resources which will be presented in the highest level conferences. Conferences addressed will be both generic and sectorial ones, within the sectors previously identified, as well as any other opportunity which appears during the project’s execution and for which PANACOTA will design internal procedures for its identification and evaluation. Here follows a list of the conferences and congresses where the consortium partners will be presenting the results of the project. Should these participations be accepted, this will be also taken as an indicator of the success of the project.

Besides, and as explained in preceding sections, PANACOTA will actively contribute to the creation and dissemination of **standards** for the field of Human Technologies by active participation in the creation of data categories (isocat.org) and its promotion. The groups and persons participating in the project have a proven experience in the creation and promotion of standards and are specially well fitted for this purpose.
Management of knowledge, of intellectual property, and of other innovation related activities arising from the project.

Finally, a business model for PANACOTA will be worked out during the life of the project that will both approach the use by a producer of Language Resources and the use as a further component in a MT system production chain. The maintenance of the laboratory will be guaranteed by its use by the partners of the consortium, that will regulate rights and duties in the Consortium Agreement. All these aspects together with the important problem of finding non publicly available texts and the agreements with that providers will be handled out by WP1 which is leaded by the European Language Resources Distribution Agency, currently the organism with more experience in the world in the licensing and handling of rights related to LRs. An important aspect that ALR has been promoting during the last decade (and that will continue) is to increase public awareness on all legal concerns in the academic centers, who, very often, "forget" to address such issues before embarking in huge LR production/customization efforts (e.g. request prior authorization from some data sources). The failure to take into account the different "layers" of intellectual property (e.g. production of new resources, derivation of resources that incorporate other resources covered by intellectual property, etc.) makes sharing/distribution efforts useless. It is also important to understand the diversity of legal protections in Europe and the rest of the world (e.g. "Fair use act" in the USA versus exclusive author's exploitation rights in Europe, Database directive 96/9 in Europe, necessity to deposit any IP within a copyright body in the USA, not required in France, etc.). It is also important to promote the use of "standard" licenses (e.g. ALR, LDC, Creative Commons licences, etc.) instead of in-house specific contract models, often inspired by models of software distribution, not necessarily well suited to the Language resources and requires, each time, a review by legal experts. The purpose of the license has to be clearly understood by the licensor and the licensee. If the licensor agreed to go for an ALR VAR (Value added Reseller) License, he will be given the details about the rights that will be granted to such user in particular the rights to "derive products and services that could be brought to the market". The VAR will have no rights to redistribute the original LRs or any revised/augmented version of it. If the licensor agreed to go for an ALR end-user license, then the user is granted the rights to use the data for internal language engineering research activities and the user is not permitted to distribute and market any derivative product or service. Other ALR licenses but also licenses like the Creative Commons ones allow the users to legally build upon the acquired resources and to share them with other parties (under the same terms which could be for Non commercial only or for both).

It is also important to increase awareness about all the open licenses that are being used. Many of them are "open" and could grant the license a freedom to adapt, improve, make changes, derive products, etc; and a large number of additional rights but not necessarily free of charge, see for instance GNU-GPL licenses.

In addition to the legal protection of LRs, one has also to consider privacy, ethical, and other human rights when it comes to the sharing of LRs. For instance medical files can not be shared unless strictly anonymised (not only deleting names and personal data but also any information or feature that may help identify the person). With respect to IPRs of the different components of PANACOTA there will be different cases:

- Third party already developed software under the GNU-GPL license (FreeLing, Weka, etc.).
- Background software owned by the partners that will keep the ownership.
- Components and software developed by the partners during the life of the project, either by one or more partners, will be open source under the GNU-GPL license.

All these aspects will be studied in detail by ELDA and appropriate licenses will be established to cover the different scenarios. Further to the legal issues, it is also our objective to design a business model that would help the project consortium and its associated players understand the platform’s core business from all points of views (feasibility to keep the platform running, financial solvency to maintain it, offers (customers/providers), strategies, practices, policies but also daily operations).

The challenge in this case is to find the best compromise between the background activities that
are conducted by the platform developers, the maintenance team and its potential users, both academics and industrials. The objective is to consolidate/establish a viable entity (organization) capable to fulfill such activities.

A crucial aspect will be to position the new platform and its whole production chain within a value-chain that would fulfill requirements of the users of the HLT community and attract new ones that are not optimally serviced today. A major challenge will be to approach this platform positioning from a point of view which is both appealing and financially viable for all types of potential users, as we should consider that such a production chain may be very beneficial for them. This may call for different varieties of the business model to be implemented.

A first phase of the project will help us to draw a business model that describes the future activities of the platform in terms of core capabilities, functionalities, applications and depicting the offers and the targeted customers. The group needs to focus on the instruments to establish in order to add value to the current distribution mechanisms (new instruments, new approaches, but also new offers). It is crucial to clearly understand the expectations so as to set up the right instruments that would make it beneficial for the users (but also for the platform “management” body). In particular it will be very important to understand the new trends in the current language resource scenario, considering, for instance, competing approaches. A number of factors are important in the construction of sustainable competitive advantage such as trust and reliability, reputation, but also efficiency, cost-effectiveness and innovation that have to be conveyed by the full production chain to be implemented by means of the platform.

Such exercise will help understand the revenue streams (having identified the different “market” segments) but also the cost structures (cost of different operations). The aim of this is to define the strategy we should adopt in order to become self-sustaining within a foreseeable period of time. The notion of self-sustaining has also to be analyzed with respect to the expected role. Alliances/agreements between the core members and the players of the different layers described in this section will also condition the success of the whole business and will be considered as a key element within the business model to draw.

During the project, we will clearly define the mission(s) of the production platform and the mechanisms to be implemented. This will help us draw a sound economic/business model and the corresponding Business Plan at least for the three years of activities that will follow the end of the PANACOTA project. When it comes to practical issues, such document will elaborate on the management of the daily operation, “supply” chain, finances, human resources, technical infrastructure, etc. The basic ideas will be structured according to a pyramid of goals: Mission statement, objectives (short and long term), tasks, and action plans for the next 3 years.

B4. Section 4

Ethical issues

PANACOTA does not raise any ethical issues: