# INFORMATION SOCIETY TECHNOLOGIES (IST) PROGRAMME


**Contract for:**

## Shared-cost RTD

# *Annex 1 - "Description of Work"*

Project acronym: MEANING
Project full title: Developing Multilingual Web-scale Language Technologies
Proposal/Contract no.: IST-2001-34460
Related to other Contract no.: *(to be completed by Commission)*

Date of preparation of Annex 1: December 3rd 2001

Operative commencement date of contract: *(to be completed by Commission)*

# CONTENTS

**CONTENTS** ..... **2**

**1. PROJECT SUMMARY** ..... **3**

**2. PROJECT OBJECTIVES** ..... **4**

**3. PARTICIPANT LIST** ..... **6**

**4. CONTRIBUTION TO PROGRAMME/KEY ACTION OBJECTIVES** ..... **6**

**5. INNOVATION** ..... **7**

**6. COMMUNITY ADDED VALUE AND CONTRIBUTION TO EC POLICIES** ..... **12**

**7. CONTRIBUTION TO COMMUNITY SOCIAL OBJECTIVES** ..... **13**

7.1 INTRODUCTION ..... 13
7.2 IMPROVING QUALITY OF LIFE, WORKING CONDITIONS AND CONTRIBUTION TO IMPROVING EMPLOYMENT ..... 14
7.3 PROJECT'S COMPLIANCE WITH ETHICAL REQUIREMENTS ..... 14

**8. ECONOMIC DEVELOPMENT AND SCIENTIFIC AND TECHNOLOGICAL PROSPECTS.** ..... **14**

**9. WORKPLAN** ..... **15**

9.1 GENERAL DESCRIPTION ..... 15
9.2 WORKPACKAGE LIST ..... 26
9.3 WORKPACKAGE DESCRIPTIONS ..... 27
9.4 DERIVERABLES LIST ..... 37
9.5 PROJECT PLANNING AND TIMETABLE ..... 39
9.6 GRAPHICAL PRESENTATION OF PROJECT COMPONENTS ..... 40
9.7 PROJECT MANAGEMENT ..... 41

**10. CLUSTERING** ..... **42**

**11. OTHER CONTRACTUAL CONDITIONS** ..... **43**

11.1 SUBCONTRACTING ..... 43
11.2 TRAVEL OUTSIDE THE EU MEMBER STATES AND ASSOCIATED STATES ..... 43
11.2 OTHER SPECIFIC PROJECT COSTS ..... 44

**APPENDIX A – CONSORTIUM DESCRIPTION** ..... **45**

# 1. PROJECT SUMMARY

## 1.1 Objectives

To be able to build the next generation of intelligent open domain HLT application systems we need to solve two complementary and intermediate tasks: Word Sense Disambiguation (WSD) and large-scale enrichment of Lexical Knowledge Bases.

WSD is the task of assigning the appropriate meaning (sense) to a given word in a text or discourse. And this is one of the most important open problems in NLP. However, progress is difficult due to the following paradox:

- In order to achieve accurate WSD, we need far more linguistic and semantic knowledge than is available in current lexical knowledge bases (e.g. current wordnets[1]).

- In order to enrich Lexical Knowledge Bases we need to acquire information from corpora, which have been accurately tagged with word senses.

The major objective of MEANING is to provide innovative technology to solve this problem. As a by-product, we will be able to index large portions of the web based on concepts rather than terms. In the long term, the results of MEANING will be also useful for the purposes of the semantic web.

## 1.2 Description of work

MEANING will develop concept-based technologies and resources through large-scale processing over the web, robust and fast machine learning algorithms, very large lexical resources and new strategies for combining them.

MEANING will treat the web as a (huge) corpus to learn information from, since even the largest conventional corpora available (e.g. the British National Corpus) are not large enough to be able to acquire reliable information in sufficient detail about language behaviour. Moreover, most European languages do not have large or diverse enough corpora available. We will use a combination of Machine Learning and novel Knowledge-Based techniques in order to enrich the structure of the WordNets in different domains (subsets of the web) in five European languages: English, Italian, Spanish, Catalan and Basque.

MEANING will produce: a) A Tool Set that using the semantic knowledge of EuroWordNet will obtain automatically from the web large collections of examples for each particular word sense. b) A Tool Set for enriching EuroWordNet using the knowledge acquired automatically from the Web. c) A Tool Set for selecting accurately the senses of the open-class words for the languages involved in the project.

MEANING will also develop a Multilingual Central Repository to maintain compatibility between

[1] A wordnet is a conceptually structured knowledge base of word senses. The synonym set, or synset, which represents a lexicalised concept is the basic unit of wordnets. The English WordNet (Miller 90, Fellbaum 95) has been developed at Princeton University over the past 14 years. EuroWordNet (Vossen 1998) is a multilingual database with wordnets for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). The EuroWordNet project was completed in the summer of 1999.

WordNets of different languages and versions, past and new. The acquired knowledge from each language will be consistently uploaded to the Multilingual Central Repository and ported over to the local WordNets involved in the project. MEANING will also produce a semantically annotated corpus for each WordNet word sense, that is, a multilingual web corpus with semantically annotated corpora containing concept and domain labels.

### 1.3 Milestones and expected results

The Multilingual Central Repository provided by MEANING will be directly used in any multilingual Internet application. MEANING will release a Showcase for evaluating the products of the project. The Showcase will include test beds and demonstrations of the enhanced WordNets in WSD, automatic acquisition of lexical knowledge, concept based Cross-lingual Information Retrieval and multilingual Q&A (Question and Answer) Systems.

## 2. PROJECT OBJECTIVES

This project aims to provide meaning to the web. MEANING will enhance current web applications by automatically increasing the linguistic depth and breath of existing multilingual resources and by devising improved concept-based Natural Language Processing (NLP) technologies using those resources.

Current web access applications are based on words; MEANING will open the way for access to the Multilingual Web based on concepts, providing applications with capabilities that significantly exceed those currently available. MEANING will facilitate development of concept-based open domain Internet applications (such as Question/Answering, Cross Lingual Information Retrieval, Summarisation, Text Categorisation, Event Tracking, Information Extraction, Machine Translation, etc.). Furthermore, MEANING will supply a common conceptual structure to Internet documents, thus facilitating knowledge management of web content. This common conceptual structure is a decisive enabling technology for allowing the semantic web.

Progress is being made in Human Language Technology (HLT) but there is still a long way towards Natural Language Understanding (NLU). An important step towards this goal is the development of technologies and resources that deal with concepts rather than words. MEANING will develop concept-based technologies and resources through large-scale knowledge processing over the web, robust and fast machine learning algorithms, very large lexical resources and novel strategies for combining them. Small-scale, isolated experiments with limited infrastructure (such as Internet access, processing power, and storage space) have no chance of bridging the gap to understanding. Advances in this area can only be expected in the context of large-scale long-term research projects.

MEANING will treat the web as a (huge) corpus to learn information from, since even the largest conventional corpora available (e.g. the Reuters corpus, the British National Corpus) are not large enough to be able to acquire reliable information in sufficient detail about language behaviour. Moreover, most European languages do not have large or diverse enough corpora available.

Even now, building large and rich knowledge bases takes a great deal of expensive manual effort; this has severely hampered HLT application development. For example, dozens of person-years have been invest into the development of wordnets for various languages, but the data in these resources is still not sufficiently rich to support advanced concept-based HLT applications directly. Furthermore, resources produced by introspection usually fail to register what really occurs in texts.

Applications will not scale up to working in the open domain without more detailed and rich general-purpose and also domain-specific linguistic knowledge. To be able to build the next generation of intelligent open domain HLT application systems we need to solve two complementary intermediate tasks: Word Sense Disambiguation (WSD) and large-scale enrichment of Lexical Knowledge Bases. However, progress is difficult due to the following paradox:

- In order to enrich Lexical Knowledge Bases we need to acquire information from corpora, which have been accurately tagged with word senses.

- In order to achieve accurate WSD, we need far more linguistic and semantic knowledge than is available in current lexical knowledge bases.

The major objective of MEANING is to innovate technology to solve this problem. MEANING will use state of the art NLP techniques pioneered by the consortium to enhance EuroWordNet with mainly language-independent lexico-semantic (concept) information. We will use a combination of Machine Learning and Knowledge-Based techniques in order to enrich the structure of the wordnets in different domains (subsets of the web) in five European languages: English, Italian, Spanish, Catalan and Basque. The core technology used by MEANING will include tools to perform language identification, morphological analysis, part-of-speech tagging, named-entity recognition and classification, sentence boundary detection, shallow parsing and text categorization. MEANING will produce:

- A Tool Set for obtaining automatically from the web large collections of concept-based data sets. This Tool Set will use the semantic knowledge of EuroWordNet to obtain automatically from the web large collections of examples for each particular word sense.

- A Tool Set for enriching automatically EuroWordNet. The knowledge acquired using these tools will support the interface between the syntactic and the semantic layers. This Tool Set will include a set of specific tools for acquiring information including domain terminology, new senses, clusters of related senses, topic signatures, Diathesis Alternations, Subcategorization Frames (including prepositional constraints), Selectional Preferences (i.e. typical objects, subjects, etc.), and specific lexico-semantic relations (i.e. purpose, location etc.).

- A Tool Set for selecting accurately the senses of the open-class words for the languages involved in the project. This WSD system will rely on robust, advanced Machine Learning algorithms able to model the behaviour of each word sense from labelled and unlabelled text.

MEANING will also develop a Multilingual Central Repository to maintain compatibility between wordnets of different languages and versions, past and new. The acquired knowledge from each language will be consistently uploaded to the Multilingual Central Repository and ported over to the other wordnets involved in the project. MEANING will also produce a semantically annotated corpus for each wordnet word sense, that is, a Multilingual Web corpus with semantically annotated corpora containing concept and domain labels.

All of these tools and data will be readily usable by users of different wordnets (including EuroWordNet and future versions of the WordNet financed by the NSF), using automatic tools for mapping the concepts between the different versions. Enriching EuroWordNet with mostly language-independent information will allow us to port newly acquired semantic information from one language to the others. This will be possible because a large portion of EuroWordNet's conceptual structure is language independent.

Research in MEANING will also cover new methods for terminology acquisition, keyword identification, topic detection, domain classification, text classification and wordnet adaptation (including identification of new senses and clustering of concept sets).

The results provided by MEANING will be directly used by any multilingual Internet applications. MEANING will release a Showcase for evaluating the products of the project. The Showcase will include test beds and demonstrations of the enhanced wordnets in WSD, concept based Cross-lingual Information Retrieval and multilingual Q&A (Question and Answer) Systems that will try to show improvement over a baseline state-of-the-art traditional word-based system.

# 3. PARTICIPANT LIST

| Partic. Role | Partic. No. | Participant name | Participant short name | Country | Date enter project | Date exit project |
|---|---|---|---|---|---|---|
| C-F | 1 | Universitat Politècnica de Catalunya | UPC | E | Start of project | End of project |
| P | 2 | Instituto Trentino di Cultura | ITC-IRST | I | Start of project | End of project |
| C-S | 3 | Universidad del País Vasco / Euskal Herriko Unibertsitatea | UPV/EHU | E | Start of project | End of project |
| P | 4 | University of Sussex | UoS | UK | Start of project | End of project |

# 4. CONTRIBUTION TO PROGRAMME/KEY ACTION OBJECTIVES

MEANING relates to a number of Key Action III objectives, but addresses most centrally action line III-3.1 (Multilingual Web) and III-4.1 (Semantic Web). The project will develop innovative enabling technologies that support the delivery of high quality information services to European individuals and companies in a **multilingual environment**, all these being among the main goals of the IST programme. Moreover, it will provide keystone enabling technologies for the **semantic web**.

Europe's language diversity is at the same time a valuable cultural heritage worth preserving, and an obstacle to achieving a more cohesive social and economic development. This situation is reflected in many official EU documents, and has been further stressed as a major challenge in the accompanying document for the HLT research lines. Improving language communication capabilities is a prerequisite for increasing European industrial competitiveness, this way leading to a sound growth in key economic sectors.

However, this obstacle will be helpful for MEANING. The idiosyncratic way the meaning is realised in a particular language will be captured and ported to the rest of languages involved in the project. MEANING will work with three major European languages (English, Spanish and Italian) and two minority languages (Catalan and Basque). All of them realise the meaning in different ways and MEANING will benefit from that because wordnets for all these languages have been constructed following the model proposed by the LE-EuroWordNet projects (LE-2 4003 & LE-4

8328). That is, the wordnets are linked to an Inter-Lingual-Index (ILI). Via this index, the languages are interconnected so that it is possible to go from the words in one language to similar words in any other language. MEANING will reuse and enrich these wordnets.

MEANING emphasises **content-based** access to information in the web, which is also a relevant topic of key action III.4.1 (Semantic Web technologies). The project will provide basic multilingual technology for language based applications. In particular, the Multilingual Central Repository produced by MEANING is going to constitute the natural linguistic resource for a number of semantic processes that need large amounts of linguistic data to be effective tools (e.g. Web ontologies). NLP tools and software of the next generation will benefit from the MEANING outcomes. The acquisition of knowledge information from large-scale document collections is one of the major challenge for the next generation of text processing applications.

At the same time, the presence of **scalability** and cross-domain **portability** as two of the main technological goals shows the consortium commitment to device solutions which have a great impact far beyond any particular economic sector, proving beneficial for the whole of e-content area.

# 5. INNOVATION

MEANING will extend the state of the art in human language technologies (HLT) in four important, related areas. It will: (1) devise innovative processes and tools for automatic acquisition of lexical information from large-scale document collections; (2) develop novel techniques for accurately selecting the sense of open-class words in a number of languages; (3) enrich existing multilingual linguistic knowledge resources with new kinds of lexical information by automatically mapping information across languages; and (4) build test-beds (i.e. Cross-lingual Information retrieval, Question Answering systems) for evaluating the use of the new information in the emerging class of advanced multilingual applications for information access from the Web. We address each of these areas in next sections.

One of the main requirements for continued progress in HLT in general, and the development of improved language processing application systems in particular, is the ability to detect the domain and topic of a text reliably, and to accurately disambiguate the senses of the words in it. The EuroWordNet multilingual knowledge base contains information that can be used as a starting point in these tasks. MEANING focuses on using, extending and enriching this resource across different languages, working on automated discovery of new senses, clustering of sense groupings ("synsets"), induction of domain information and topic signatures, amongst others. MEANING plans to structure the documents in four levels of information, one level feeding the other:

- keywords: open list of relevant terms for a document
- topic: open list of relevant concepts for a document
- domain: close list of domains
- genre: close list of genres

The partners in the consortium are uniquely qualified to work in this area. For instance, ITC-IRST has a large experience studying the linguistic relations between these levels of information and their representation into WordNet (Magnini and Cavaglià, 2000). Furthermore, UPC, ITC-IRST and UPV/EHU have been involved in building European wordnets from its inception (Atserias et al., 1997; Benítez et al., 1998; Roventini et al., 2000, Agirre et al., 2002). In addition, out of a strong

international field, the partners were also in the top-scoring groups in the recent SENSEVAL-2[2] word sense disambiguation exercise.

### 5.1 Innovative processes and tools for automatic acquisition of linguistic knowledge from large-scale document collections

The acquisition of linguistic knowledge from corpora has been a very successful line of research. Research in the acquisition of subcategorization information, selectional preferences, terminology, domain information, topic signatures, etc. has obtained remarkable results. The acquisition process usually involves large bodies of text which have been previously processed with shallow language processors.

Much of the use of the acquired information has been hampered by the fact that the texts are not sense-disambiguated, and therefore, only knowledge for words can be acquired, that is, subcategorization for words, selectional preferences for words, etc. It is a well established fact that much of the linguistic behavior of words can be better explained if put into reference to word senses.

MEANING plans to make use of texts that have been automatically sense-tagged with high accuracy in order to produce significantly better acquired knowledge at a sense level, including subcategorization frequencies, domain information, topic signatures, selectional preferences, specific lexico-semantic relations, thematic role assignments and diathesis alternations. Furthermore, MEANING plans to investigate automatic methods for dealing with new senses not present in the EuroWordNet and clustering of word senses.

All partners have expertise concerning several knowledge acquisition algorithms. UPC has been involved in thematic role assignments and diathesis alternations (Ribas, 1995). ITC-IRST has experience in the acquisition of domain information for WSD (Magnini et al., 2001). UPV/EHU is involved in the acquisition of subcategorization frequencies, topic signatures, selectional preferences, specific lexico-semantic relations as well as diathesis alternations (Agirre and Martínez 2001a, 2002; Agirre et al. 2001a). UoS has a large experience in acquisition of subcategorization, selectional preferences, thematic roles and diathesis alternations (McCarthy and Korhonen, 1998; Korhonen et al., 2000; McCarthy 2001).

The fact that the word senses will be linked to concepts in Multilingual Central Repository will allow for the appropriate representation and storage of the acquired knowledge. Section 5.3 mentions also the advantages of having a common multilingual concept inventory linked to the word senses in each language.

### 5.2 Novel techniques for accurately selecting the senses of open-class words

Word Sense Disambiguation (WSD) is the task of assigning the appropriate meaning (sense) to a given word in a text or discourse. Ide and Veronis (1998) argue that word sense ambiguity is a central problem for many established HLT applications (for example Machine Translation, Information Extraction and Information Retrieval). This is also the case for associated sub-tasks (for example reference resolution and parsing). For this reason many international research groups are working on WSD, using a wide range of approaches. However, no large-scale broad-coverage accurate WSD system has been built up to date (Kilgarriff and Rosenzweig 2000). With current

[2] http://www.sle.sharp.co.uk/senseval2/

state-of-the-art accuracy in the range 60-70%, for systems trained on a small number of words (due to the large effort needed to manually annotate examples from running text), WSD is one of the most important open problems in NLP.

This promising current line of research uses semantically annotated corpora to train Machine Learning (ML) algorithms to decide which word sense to choose in which contexts. The words in these annotated corpora are tagged manually with semantic classes taken from a particular lexical semantic resource (most commonly WordNet). Many standard ML techniques have been tried, such as Bayesian learning, Exemplar based learning, Decision Lists, Neural Networks, and recently margin-based classifiers like Boosting and Support Vector Machines. These approaches are termed "supervised" because they learn from previously sense annotated data and therefore they require a large amount of human intervention to annotate the training data.

Supervised WSD systems are data hungry. They suffer from the "knowledge acquisition bottleneck" -it takes them mere seconds to digest all of the processed corpus contained in training materials that take months to annotate manually. So, although Machine Learning classifiers are undeniably effective, they are not feasible until we can obtain reliable unsupervised training data.

Ng (1997) estimates that the manual annotation effort necessary to build a broad coverage word-sense annotated English corpus is about 16 person-years; and this effort would have to be replicated for each different language. He estimates this based on his experience when producing the DSO corpus with around 1000 training examples for 103 words (Ng, 1996). Unfortunately, many people think that Ng's estimate might fell short, as the annotated corpus thus produced is not guaranteed to enable high accuracy WSD. In fact, recent studies that use the corpus produced by Ng for 103 words have shown that: 1) the performance for state of the art supervised WSD systems continues to be in the 60%-70% for this corpus, 2) some highly polysemous words get very low performance, and 3) domain and genre shifts degrade seriously the performance. Effects 1) and 2) can be explained by the fact that the number of examples is still low, specially in the case of highly polysemous words, where the average amount of examples per word sense can be as low as 20.

Apart from the DSO corpus, there is another major sense-tagged corpora available for English, SemCor (Miller et al., 1993), and a few comparable resources for other languages resulting from SENSEVAL competitions. UPC and UPV/EHU were involved in the production of the data for Spanish and Basque languages (Rigau et al. 2001; Agirre et al., 2001b). All these corpora provide similar or less examples per word than DSO.

Some recent work is focusing on reducing the acquisition cost and the need for supervision in corpus-based methods for WSD. Leacock et al. (1998) and Mihalcea and Moldovan (1999) automatically generate arbitrarily large corpora for unsupervised WSD training, using the synonyms or definitions of word senses provided in WordNet to formulate search engine queries over the Web. On another independent research area (Yarowsky, 1995) and (Blum and Mitchell, 1998) have shown that it is possible to reduce the need for supervision with the help of large amounts of unannotated data. Following this ideas, UPV/EHU has developed knowledge-based prototypes for obtaining accurate examples from the web for specific WordNet synsets, as well as, large quantities of unannotated examples (Agirre and Martínez, 2000).

In order to make significant advances in WSD system accuracy, systems need to be able to use types of lexical knowledge that are not currently available in wide-coverage lexical knowledge bases: for example subcategorisation frequencies for predicates (particularly verbs) rely on word senses, selectional preferences of predicates for classes of arguments, amongst others (Agirre and

Martínez, 2001). UoS has carried out pioneering work on the acquisition of such information and their use in WSD (Carroll and McCarthy, 2000; McCarthy et al., 2001). Unfortunately, a prerequisite for acquiring this information is the existence of large quantities of sense-tagged data.

In short, the different ways forward to tackle the acquisition bottleneck for WSD that have been proposed so far are the following:

a. Applying domain information in WSD algorithms.
b. Using sophisticated linguistic knowledge like e.g. syntactic structure, selectional preferences, domain information etc. (knowledge acquired as described in section 5.1) to induce a richer set of features.
c. Acquiring large numbers of automatically sense tagged examples from the web.
d. Combining annotated and unannotated data with transductive models of ML techniques.

Moreover, the fact that the acquisition processes in section 5.1 gets more accurate data whenever the source texts have been conveniently tagged with word senses, produces an inter-dependency: one of the components of a WSD system relies on the knowledge acquired in section 5.1. MEANING proposes an innovative bootstrapping to deal with this inter-dependency:

1. Train accurate WSD systems and apply them to very large corpora by coupling knowledge-based techniques on the existing EuroWordNet (e.g. to populate it with domain labels, to induce automatically training examples) with ML techniques that combine very large amounts of labeled and unlabeled data. When ready, use also the knowledge acquired in 2.
2. Use the obtained accurate WSD data in conjunction with shallow parsing techniques and domain tagging to extract new linguistic knowledge to incorporate into EuroWordNet.

This method will be able to break this interdependency in a series of cycles thanks to the fact that the WSD system will be based on all domain information, sophisticated linguistic knowledge, large numbers of automatically tagged examples from the web, and a combination of annotated and unannotated data (points a. through d. above). The first WSD system will have weaker linguistic knowledge, but the sole combination of the rest of the factors will produce significant performance gains. Besides, some of the required linguistic knowledge can be acquired from unnanotated data, and can therefore be acquired without using any WSD system. Once acceptable WSD is available, the acquired knowledge will be of a higher quality, and will allow for better WSD performance.

Another important factor to overcome the acquisition bottleneck is the huge amount of training data that MEANING is intending to use. The web will be used to get large numbers of automatically acquired annotated data and very large numbers of unannotated data. This large numbers will warrant that the ML algorithms attain high levels of performance, as the ML algorithms will have a large enough number of training examples per word sense in each possible domain.

The improvements in points a-d above have been explored separately with relative success, but the mixture of them all requires the combination of the expertise in all of them. Moreover, the inter-dependency of the acquired knowledge and the WSD system calls for in-depth knowledge on the processes involved. The expertise required to come up with a WSD algorithm that is able to combine factors a-d above, and that takes into account the inter-dependencies involved is not to be found in a single research group. In fact, no research group in isolation has tried to combine all this aforementioned factors. But we are convinced that only a combination of all relevant knowledge and resources will be able to produce significant advances in this crucial research area.

UPC and UPV/EHU have large experience in this field studying the performance and developing efficient Knowledge-Based and ML algorithms for WSD (Escudero et al., 2000a, 2000b, 2000c, 2000d, 2001; Martínez and Agirre, 2000). Partner UPV/EHU has prototypes to produce large numbers of automatically annotated and unannotated data from the web (Agirre and Martínez, 2000). ITC-IRST has a running system doing domain-based word sense disambiguation (Magnini et al., 2001). UoS has also working systems on word sense disambiguation based on subcategorization and selectional preference information (Carroll and McCarthy 2000; McCarthy et al., 2001). All partners have experience on the acquisition of linguistic knowledge from corpora, as mentioned in section 5.1.

Most of the partners have been participating either in joint research projects, or have been doing research together, and have the required knowledge of each other to work in such a technical project together. We think that the ambitious goals for this project can only be met by a balanced combination of expert research teams which is not too small to deal with all required resource and algorithms, neither too large to be too difficult to be effectively coordinated. We think that the number and expertise of the partners involved offers the best conditions to get successful results.

MEANING plans to validate their technology against the state of the art in future SENSEVAL lexical and all-words disambiguation tasks (whichever available). In fact, the consortium will be involved in the organization of the next SENSEVAL competitions.

### 5.3 Enriching existing multilingual lexical resources with new kinds of linguistic knowledge by automatically mapping data across languages

All languages involved in MEANING realise the meaning in different ways. MEANING will benefit from that using a novel multilingual mapping process. The project will use the existing EuroWordNet knowledge base, concentrating on the component wordnets for English, Italian, Spanish, Basque and Catalan. The wordnets are currently linked via an Inter-Lingual-Index (ILI) allowing the connection from words in one language to translation equivalent words in any of the other languages.

MEANING technology will help one language to each other. For instance, for Basque, being an agglutinative language with very rich morphological-syntactic information, MEANING will extract semantic relations that would be more difficult to capture in other languages. However, Basque is not largely present in the web as the others. MEANING technology will balance both gaps.

The partners have expertise in acquiring several different kinds of lexical information that will be needed for the next generation of WSD systems, but which is not available in any wordnet or comparable computational resource (e.g. subcategorisation frequencies, selectional preferences, domain information, domain terminology, topic signatures, conceptual clusters, specific lexico-semantic relations).

MEANING will develop procedures for porting and uploading the various types of information across languages via the EuroWordNet Inter-Lingual-Index to enrich each of the individual monolingual wordnets. UPC has the technology for the automatic alignment of different large-scale and complex semantic networks (Daudé et al, 1999, 2000, 2001). This technology will provide compatibility to the Multilingual Central Repository across the European wordnets, past and new. MEANING will also test the validity of samples of the mapped data in the target languages. The outcome will be a set of innovative wide-coverage repositories of lexico-semantic information.

### 5.4 Test-beds for evaluating the use of the new information in advanced multilingual applications for Web information access

Reliable WSD would benefit many types of established HLT application system (e.g. Machine Translation). It is likely that the same is true for the new generation of internet-based information access and management applications (for example automatic content-based web page indexing and automated email response). MEANING will establish the extent to which WSD would benefit representative, commercially important applications in the latter class. The project will produce the Multilingual Central Repository, large portions of the web annotated semantically, a concept-based Cross-lingual Information Retrieval system and a multilingual open-domain Question Answering system.

Currently, most language processing engines used in application systems are either developed manually from scratch or ported from other domains or applications in an expensive labour-intensive process. Since the rules are hand-coded and scarcely abstract away from the actual language used, it is difficult to adapt such systems to new domains.

The core results produced by MEANING will consist of comprehensive multilingual repositories encoding lexico-semantic information about word meanings, and accurate tools for tagging words with semantic classes. The existence of such technologies would constitute a significant advance in the state of the art, and greatly facilitate the construction of a new generation of HLT application systems.

## 6. COMMUNITY ADDED VALUE AND CONTRIBUTION TO EC POLICIES

MEANING will exploit the individual competence of the technological partners of the consortium integrating groups with complementary skills, including: empirical knowledge-based and Machine Learning know-how, computational thesaurus development, multilingual lexicon linking, automatic acquisition of linguistic knowledge from corpora and Word Sense Disambiguation (WSD). Furthermore, the technology centres in MEANING are specialized in several different European languages, which makes it possible to study the language-specificity of WSD and to exploit the possibilities of language-transfer of technology and solutions. A consortium with such a range of expertise can only be realised at a trans-national level. In bringing together these different technologies, MEANING gives considerably more "added value" to the research than would be possible at a purely national level, in line with EU RTD policy.

MEANING builds on the internationally-leading expertise in Europe on automatic acquisition of information about word meaning, developed in part through the EU Framework IV projects '**SPARKLE**: Shallow PARsing and Knowledge extraction for Language Engineering' and '**EuroWordNet**: Building multilingual wordnets with semantic relations between words', and their application in real scenarios, through the EU Framework V project '**NAMIC**: News Agencies Multilingual Information Categorisation'. These projects have put Europe at the forefront of research and development on word and concept-based computational resources. MEANING will provide a platform for maintaining this position, and in addition will drive a concerted effort into the use of this technology for improved word sense disambiguation and related concept-based NLP technologies. The achievement of European technological leadership is another facet of Community added value.

As well as combining technologies, MEANING will consolidate the associated language data to

form enriched multilingual thesauri. Large-scale resources of this type are scarce, and integrating the complementary types of data will establish a "critical mass" of resources at the European level, so that they achieve a much greater impact than they would individually. This is another motivation for carrying out this research and technology development collaboratively on a European basis.

MEANING tackles a problem that is important for a multilingual Europe. Progress in natural language processing research and development is heavily reliant on large-scale, accurate repositories of information about language. MEANING will build resources, which for example specify possible attributes of word-level concepts (e.g. subject domain) and relations between concepts (e.g. performer of an action). However, computational tools and data for deriving these attributes and relations are often available for a single language only. MEANING will bootstrap corresponding resources for other languages, mapping, uploading and porting concept-based data between five European languages. This requires a dynamic and responsive integrated effort that can only be achieved by carrying out the work at European level.

Enhanced broad and deep concept-based NLP technologies, of the type to be developed in MEANING, will be central to the next generation of advanced information services working across the Web. To be competitive in the global marketplace, all initiatives towards the exploitation of the "semantic content" of the textual data are vital. To support these it is strategically important that Europe will be in a position to develop, deliver and use such products and services. MEANING therefore contributes to EU policies relating to strengthening the international competitiveness of the European industry.

Public deployment of general-purpose, "semantic" information access services will provide individual European citizens and consumers with more reliable ways of navigating the largest repository of information currently available: the World Wide Web. In this way, the work addresses EU policies on the societal benefits that state of the art information society technologies can offer, for example by facilitating more highly focussed and customised access to Internet content. MEANING will develop concept-based resources in five European languages. Two of these languages are minority languages (Basque and Catalan), so it will contribute to EU policies on European economic and social cohesion by helping to facilitate uniform access to information for citizens whose current possibilities are restricted by virtue of language barriers.

# 7. CONTRIBUTION TO COMMUNITY SOCIAL OBJECTIVES

## 7.1 Introduction

MEANING is in accordance with some of the driving principles that are shaping EU’s social policies:

1. Universal services must be ensured together with network interconnection and the interoperability of services and applications throughout the Union. Similar measures are needed in other regions of the world, which also guarantee equal access.
2. Cultural and linguistic diversity should be protected and promoted.
3. Co-operation should be promoted with less economically advanced countries.
4. Economic operators must be made aware of the new opportunities, which the information society presents for them.

## 7.2 Improving quality of life, working conditions and contribution to improving employment.

The outcomes of MEANING will help reach some important social objectives in the following way:

- Working directly with the information market, it contributes to the developing "knowledge" policies by means of a new impetus for Community Research and Technological Development.
- Enhancing accessibility of the multilingual information in the Web to people, it contributes to improve living conditions to ensure that the benefits of growth promote a more cohesive and inclusive society. Technology as developed by MEANING will help provide natural interaction.
- The MEANING project provides a very effective contribution to the emerging Information Society. We are living a revolutionary period with profound impact of the new information and communication technologies. They are steadily changing the ways in which we live and work, transforming our societies.
- Through MEANING technology it will be possible to organize the documents in the web according to the concepts they refer to, independently of the languages they are written (for the moment Basque, Catalan, English, Italian and Spanish). European nations are a patchwork of distinct cultures, languages and traditions. Nowadays, the European Union is composed by fifteen members. In the next coming years ten more countries and their relative culture could make the number of languages even larger, complicating even more the management of all this knowledge.

## 7.3 Project's compliance with ethical requirements

In this multicultural panorama, one of the most pressing objectives that must be reached is the social cohesion and reciprocal comprehension among European cultures. Of course the distinct languages that are the expression of different cultures and their traditions, do not help the comprehension among the various EU members.

Multilingual technologies, including those provided by MEANING, are of fundamental importance to reach the objective of social cohesion in the European landscape. In fact social cohesion can be improved thorough the creation of new tools able to overtake language and cultural barriers. Multilingual technologies that promote accessibility of the information in other foreign languages are able to contribute actively to the enhancement of a closer European strengthening social integration and cohesion.

# 8. ECONOMIC DEVELOPMENT AND SCIENTIFIC AND TECHNOLOGICAL PROSPECTS.

Language pinpoints thoughts and concepts using words, but does that with words that can mean different things in different contexts, while at the same time different words or expressions can mean exactly the same thing or very similar things. Some scholars have therefore said that all

meaning is fluid and can only be defined in context. Other scholars nevertheless claim that meaning can at least partially be described and they even produced finite descriptions of concepts and relations between words. Wordnets are the well-known outcome of this approach. Practical experiments have shown that resources such as the English WordNet already improve current state-of-the-art NLP techniques but they have also shown that they are still not rich enough for commercial performance. Resources such as WordNet capture some essential aspects of meaning but would require massive further development to meet realistic requirements of genuine conceptual approaches to Knowledge Management and the Semantic Web. Even if such an investment is made, it requires continuous work to keep it up to date, because the vocabulary changes over time.

MEANING builds a bridge between words, concepts and contexts. In this respect, MEANING touches upon the most essential aspect of Language Technology. It gives lexical semantic resources sufficient body and context to take computational decisions on the ***meanings*** of words. It will also enable large-scale empirical studies of the usage of word meanings in contexts, which is impossible with the current corpora that are all word-based. MEANING thus takes language-technology from the word level to the concept level.

MEANING will directly improve the current state-of-the-art in Language Technology as it is incorporated in Information Retrieval, Summarisation, Information Extraction and Question Answering. Any technologies where word comparison can be replaced from more-precise concept comparison will solve many limitations of current software, without introducing new problems due to spurious expansions and ambiguities.

Furthermore, MEANING will make it possible to keep the developed resources up to date. By connecting concepts to context, it will be more efficient to discover and relate new meanings and words and customise the resources.

MEANING will also open more sophisticated developments. It will be possible to develop context-specific corpora that can form the basis for terminology development, translation resources, fact extraction and data mining. Furthermore, a more precise conceptual representation of the meaning of words makes it possible to develop dialogue systems and connect lexical semantic resources to more powerful ontological-based solutions. The latter systems usually lack the multilinguality and large-scale coverage that is provided by MEANING.

Finally, MEANING will make it possible to transfer acquired conceptual contexts and knowledge from one language to another. This is not only important for cross-lingual language-solutions in the European market but also to be able to quickly develop or expand new resources for new languages.

# 9. WORKPLAN

## 9.1 General Description

The duration of the MEANING project is 36 months. The work is subdivided into three main phases:

1. **Analysis** (User and Software requirements, overall architectural design definition; months 1-6).
2. **Development** (three releases of the software tools and resources; months 7-30).
3. **Assessment and Validation** (months 10-36).

The technical and administrative management are active throughout the whole project life-cycle, as are the dissemination and exploitation of the project results.

The workplan has been broken down into ten work packages (WP0-9):

- WP0 is devoted to Project Management;
- WP1 deals with the definition of the user requirements;
- WP2 is devoted to definition of the overall architectural methodology and design;
- WP3 to WP7 will develop and evaluate the MEANING tools and resources;
- WP8 deals with the user validation of tools and resources provided by MEANING and with the final demonstration;
- WP9 deals with Exploitation and Dissemination.

For each work package a WP leader is in charge to co-ordinate the work, to monitor the planned activities and to check the overall quality of the deliverables produced. The WP leader is technically responsible for the results achieved and to report them to the Project Management Board and to the Project Manager .

Five milestones at months 6, 12, 21, 30 and 36 have been fixed to assess the intermediate results and to identify the end of crucial phases of the project.

**Analysis (months 1-6)**

Purposes of this phase are:

- to refine the ideas about the tasks to be performed, computing equipment, definition of what is expected from the MEANING project in order to produce a complete definition of the *user requirements*;
- to provide the developers with an *overall architectural design* of the whole MEANING process.
- to analyse the statements of user requirements and produce a set of *software requirements* as complete, consistent and correct as possible;

The outcomes of this phase are:

- the User Requirements Report (deliverable D1.1);
- the Overall Architectural Design Document (deliverable D2.1);

WPs involved are from WP1 to WP8: in this early phase of the project users and developers work together to model the user services specifications and to define the technical specifications of all the MEANING components, which will be further developed in the next phase.

The milestone to be achieved is then related to the achievements of the objectives above described:

- **M1 – Month 6**
  - User requirements;
  - Overall Architectural Design and Software Requirements

**Development (months 7-30)**

This phase is central to the project, and aims to develop all the software tools and resources to produce the final MEANING outcomes.

The development is organised in three consecutive cycles involving WP3-7. Four work packages, from WP3 to WP6, have been identified to carry out the development of the software tools. They will carry out the three consecutive acquisition, word sense disambiguation, uploading and porting processes, while WP7 is devoted to assess and evaluate the tools developed, the process carried out and the resources produced.

WP3 is devoted to develop the Linguistic Processors for each language involved in the project. Picture 1 summarises the MEANING data flow. Each development cycle consists of:

- WP6 (WSD): Word Sense Disambiguation using the local wordnets and the enriched knowledge ported from the Multilingual Central Repository (WSD0, WSD1, WSD2).

- WP5 (Acquisition): Local acquisition of knowledge using specially designed tools and resources, corpus and wordnets (ACQ0, ACQ1, ACQ2).

- WP4 (Integration): Uploading the acquired knowledge from each language into the Multilingual Central Repository and porting to the local wordnets (PORT0, PORT1, PORT2).

After each cycle WP7 is devoted to the evaluation and assessment of the software tools and resources produced in MEANING.

WSD0 and ACQ0 will start simultaneously using the already existing knowledge placed into the local wordnets. The knowledge acquired during this phase will be uploaded into the Multilingual Central Repository and will be ported (PORT0) to the local wordnets. Next cycles of WSDi and ACQi will start simultaneously using the knowledge acquired from the previous phase (PORTi-1, sense-tagged and syncatically annotated corpora, etc.). That is, WSDi+1 will benefit from ACQi and ACQi+1 from WSDi.

The first cycle consist on the independent acquisition and WSD using the knowledge currently available in the local wordnets (ACQ0, WSD0) and the first porting of those results (PORT0).

At the end of the second cycle MEANING will benefit ACQ1 with the knowledge placed into the Multilingual Central Repository and the sense-tagged corpora provided by WSD0; and also, WSD1 with the knowledge from the Multilingual Central Repository and the syntactically annotated corpora provided by ACQ0.

At the end of the third and final cycle MEANING will have the results from a complete sequence of Acquisition and WSD: ACQ2 over results from WSD1 (resulting from ACQ0); and WSD2 over results from ACQ1 (resulting from WSD0). And also the corresponding PORTings.

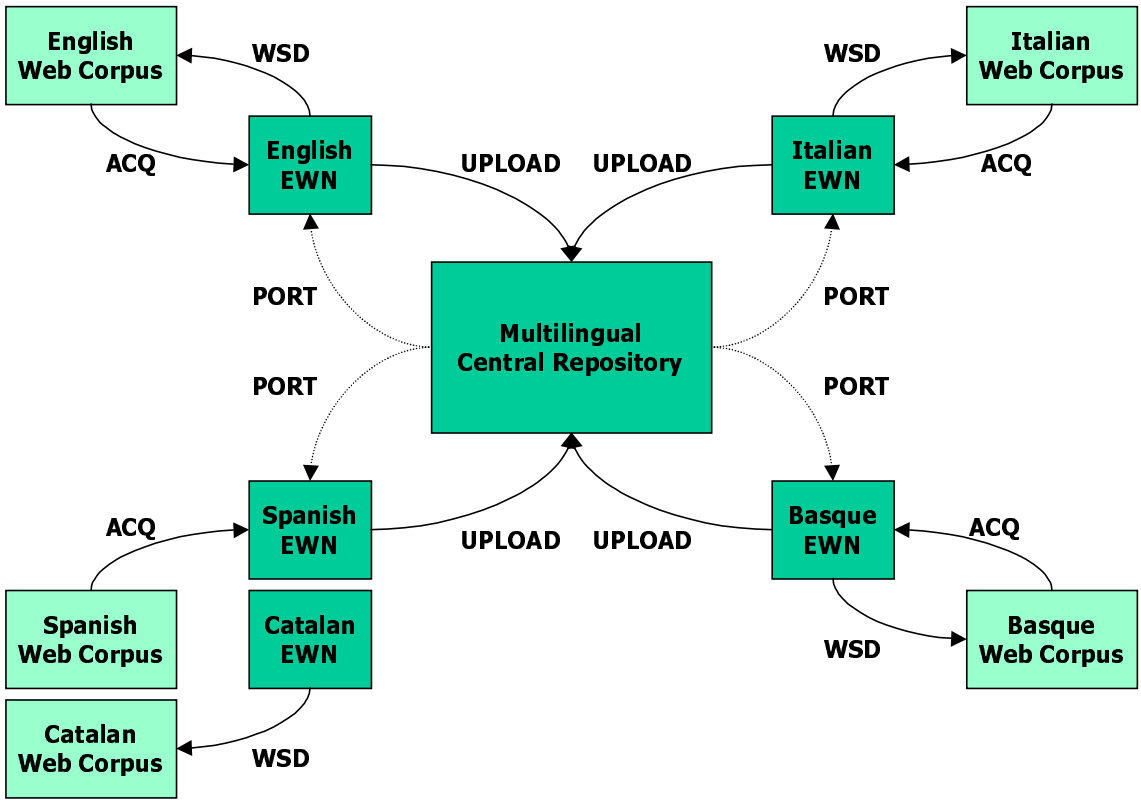



**Figure 1:** MEANING data flow.

Major milestones and output results during this development regard the delivery of the software tools and resources, whose three version submissions will be at 12, 21 and 30 months.

The end of this phase will coincide with the starting of the validation activities on the demonstration.

Milestones:

- **M2 – Month 12**
  - First release of Linguistic Processors (English, Italian, Spanish, Catalan and Basque);
  - First release of Multilingual Central Repository Software
  - ACQ0: First acquisition process
  - WSD0: First Word Sense Disambiguation process
  - PORT0: First upload and porting processes

- **M3 – Month 21**
  - Second release of Linguistic Processors (English, Italian, Spanish, Catalan and Basque);
  - Second release of Multilingual Central Repository Software
  - ACQ1: Second acquisition process.
  - WSD1: Second Word Sense Disambiguation process
  - PORT1: Second upload and porting processes

- **M4 – Month 30**

- Final release of Linguistic Processors (English, Italian, Spanish, Catalan and Basque);
- Final release of Multilingual Central Repository Software
- ACQ2: Final acquisition process.
- WSD2: Final Word Sense Disambiguation process.
- PORT2: Final upload and porting processes

Next sections will provide for each component, details on the different (sub)problems and proposed solutions for each component.

**Language Processors and Infrastructure (WP3)**

We will re-engineer and scale up existing robust language processing software tools (for English, Italian, Spanish, Catalan and Basque) in accordance with the WP1 assessment of the software requirements for MEANING. The software tools will form part of the systems developed for acquisition (WP5) and word sense disambiguation (WP6).

The functionality of these tools include:

- tokenisation and sentence boundary detection
- lemmatisation
- part of speech tagging
- noun group chunking
- robust shallow parsing
- named-entity recognition and categorisation (e.g. into location, company or product names)
- keyword, topic and terminology detection
- text classification (e.g. ECONOMIC, SPORT domains)

We will direct further development and refinement effort via assessment of accuracy and speed of the tools within the context of WP5 and WP6. This workpackage will produce five software tools: ELP (English language processor), ILP (Italian language processor), SLP (Spanish language processor), CLP (Catalan language processor) and BLP (Basque Language processor).

Immediately on project start-up we will carry out such engineering actions as are necessary to equip each partner with fast Internet access, and sufficient processing power and storage space. Due to the amount of data MEANING will be processing, WP4-6 will involve heavy computing and the need for large amounts of storage. The consortium is therefore budgeting for workgroup server-level computational resources and high capacity hard disk arrays.

**Integration (WP4)**

The Multilingual Central Repository acts as a multilingual interface for integrating and distributing all the data produced by MEANING. The different knowledge acquired from each wordnet will be uploaded and ported across languages and resources, maintaining the compatibility among them. The knowledge acquired from each language during the three cycles will be consistently upload into the Multilingual Central Repository, granting the integrity of all the data produced by MEANING. After each MEANING cycle, all knowledge acquired and integrated into the Multilingual Central Repository will be then distributed across local wordnets.

Thus, the Multilingual Central Repository will include modules for:

- Uploading the data acquired from one language to the Multilingual Central Repository.
- Porting the knowledge stored into the Multilingual Central Repository to the local wordnets.
- Checking the integrity of the data stored in the Multilingual Central Repository.

This workpackage will perform the three consecutive processes for uploading and porting the knowledge acquired from each language to the respective local wordnets: PORT0, PORT1, PORT2.

**Acquisition (WP5)**

We will design and apply advanced automatic learning techniques for acquiring linguistic knowledge from large quantities of raw (unannotated) text obtained from large texts collections (the web and large streams of news from News Agencies). Participants 1-4 all have prototype acquisition systems for the various types of information that will be learned, preliminary results from which have led to numerous recent publications at international level. In this workpackage (WP5) the acquisition systems will be refined and applied to large amounts of language data.

The acquisition systems are comprised of lower-level linguistic processing tools, which are the subject of WP3. For instance, UPV/EHU's topic signature acquisition system contains an Internet document retrieval engine, a tokeniser, a lemmatiser and a statistical classifier. As another example, UoS's subcategorisation acquisition system contains a tokeniser and sentence boundary detector, a lemmatiser and part of speech tagger, and uses a robust shallow parser to identify putative predicate subcategorisation frames.

Over the course of the project the acquisition phase takes place three times for each language (ACQ0, ACQ1 and ACQ2). The input of the last two will be a large word sense disambiguated corpus. This corpus is produced by downloading text from automatically-selected web documents and applying the WSD system (WP6) resulting from the previous acquisition-porting cycle. ACQ0 is an exception, as the corpus will not be sense disambiguated. The acquired knowledge is stored in the local wordnet, uploaded to the Multilingual Central Repository (WP4), and is then ported to the other languages.

Each acquisition phase uses the results of the previous cycle together with further refined linguistic processing tools (WP3). The details of the methodology and schedule for acquiring the various kinds of information will be elaborated in WP2 (Methodology and Design). However, we envisage that the first acquisition phase, using already existing knowledge in the local wordnets, will learn subcategorisation frequencies, topic signatures, terminology and domain information; the second phase will key them to word senses induced by WSD0, and will additionally learn new senses, coarser-grained sense clusters and selectional preferences; and the third will learn specific lexico-semantic relations and thematic role assignments for nominalizations and diathesis alternations. In addition, each phase will provide an increasingly larger set of automatically retrieved examples per word sense.

As WP6 integrates these types of lexico-semantic information into WSD, each acquisition phase will be used to refine information acquired in a previous phase, so the results of the phases should be of successively higher quality. This will be verified in WP7 through measurements of the quality of the information and the accuracy of WSD.

**Word Sense Disambiguation (WP6)**

We will design and apply a combination of unsupervised Knowledge-based and supervised Machine Learning techniques that will provide a high-precision system that is able to tag running text with word senses in real time. The system will use the word senses in the Multilingual Central Repository as the sense inventory for nouns, verbs, adjectives and adverbs. The qualitative leap will be possible because of the combination of several factors:

- A system that acquires a huge number of examples per word from the web (WP5). Some of the examples will automatically come with a sense tag (WP5), as well as a domain categorization tag (WP3).

- The use of sophisticated linguistic information, such as, syntactic relations, semantic classes, selectional restrictions, subcategorization information, domain, etc. The Linguistic Processors (WP3) and the Multilingual Central Repository (WP4) with the acquired knowledge (WP5) will provide this linguistic information.

- Efficient margin-based Machine Learning algorithms (e.g. Boosting and Support Vector Machines).

- Novel algorithms that combine tagged examples with huge amounts of untagged examples in order to increase the precision of the system. (e.g. co-training and transductive learning modes of margin-based classifiers).

The consortium have large experience on applying different techniques and knowledge types to WSD. In this workpackage they will combine their expertise and develop qualitatively improved systems.

The final system is not to be produced in one single step. We plan to develop the final system in a serious of three phases. Each phase will have an increasing amount of examples and richer acquired knowledge (WP5).

- WSD0 will produce a baseline system for all polysemous words (ca. 23.000) in the local wordnets, trained on a small number of automatically retrieved training examples. The baseline system should show improvement with respect to a system trained on currently existing corpora (e.g. SemCor).

- WSD1 will produce a first system for all polysemous words in the Multilingual Central Repository. This system will use the knowledge acquired in ACQ0 and an bigger number of automatically retrieved training examples.

- WSD2 will produce the final system, which will be trained with a significant subset of the web, and will take advantage of the knowledge induced in ACQ1.

**Assessment and Validation (months 10-36)**

**Evaluation and Assessment (WP7)**

This workpackage will deal with the technical evaluation and assessment of all software and data produced by MEANING. Specifically:

- The linguistic processors and infrastructure (WP3)
- The knowledge acquired and placed in the Multilingual Central Repository (WP5)
- The knowledge ported to the different languages (WP4)
- The performance of the WSD system (WP6)

The evaluation will be performed internally using standard evaluation measures like precision and recall. Evaluation and assessment will be performed three times at the end of each acquisition, disambiguation, uploading and porting cycle, and twice for each release of the linguistic processors. After the evaluation the improvement with respect to the previous cycle will be assessed. This will allow to measure the progress of the project, and to take the necessary actions in case of deviations.

**User validation (WP8)**

Separately from the technical evaluation of MEANING that is described in WP7, we will carry out a user-based evaluation. The evaluation is separated in two separate task:

- verification of the intermediate results after each production cycle in MEANING
- demonstration of MEANING by integrating the results in existing web products.

The purpose of the verification is to check whether the results satisfy the user-requirements and to provide the project with feedback on the applied methodology. The verification will directly assess the evaluation criteria formulated in WP1 (user-requirements) to the intermediate project results. This will result in a report (D8.1, D8.2 and D8.3) after each cycle, stating the quality of the results according to these criteria.

The demonstration will show the feasibility of integrating the project results into an existing industrial environment. Demonstration will be carried out at the final project workshop or review. In addition to the demonstration, D8.4 will describe the way and ease of the integration.

**Exploitation and dissemination**

The exploitation activities, inserted in the specifically devoted (WP9) work package, deal with the prospects and opportunities for the commercialisation of the potential products and services coming from the pre-marketable MEANING outcomes. A continuos tuning to the market evolution and demand for the product is required by the all the partners, particularly the users and the industrial ones, and a specific exploitation strategy will be studied and carried out, able to open the horizons coming from the use of the new technology. The Exploitation strategy will be reported in the Dissemination and Exploitation Plans (deliverables D9.1 and D9.2), which will be reviewed and amended as the project progresses. If the first, preliminary version is expected to concern opportunities of short-term marketing of the project results, the final version of this document will include a real market-oriented final report, able to guide the industrial partners involved in the Consortium towards a concrete exploitation of the MEANING technology. This will follow the distinction between short-term and long-term exploitation, as is made in WP1 for the user-requirements.

As for the dissemination activities, that means is to present and diffuse the results of the projects supported by the European Community outside the Consortium and outside the IST Programme, the MEANING Consortium will define an active dissemination strategy, closely involving the user groups. The Dissemination and Implementation plans (deliverables D9.1 and D9.2) will describe concrete measures on how information about MEANING will be diffused within and outside the

Consortium and the respective roles of the involved partners. Among the other things:

- the Universities and Research Associations involved in the project will work in disseminating results (both research and application related) in scientific sites for the whole duration of the project in the form of articles in journals or conference proceedings as well as presentations at scientific events, fairs, workshops and conferences. The consortium will specifically target evaluation schemes such as SENSEVAL and TREC.

- In addition, the project will organise two workshops inviting relevant researchers involved in the MEANING technologies. The first workshop will be carried out during the first year of the project to obtain detailed feedback for the MEANING design. The second workshop will be held at the third year of the project. In this case the main goal will be to present the main results of MEANING and to promote an in depth discussion about the main achievements and drawbacks of the MEANING technology.

- The dissemination materials listed in Appendix X will supply general information about the project and also will serve as handout on events like exhibition, concentration meetings, etc. The dissemination materials will be updated when results become necessary to report. In this way, they will be able to reflect at any time the global status of the project.

- The consortium will directly disseminate the results to a user-group of interested companies and institutes. We will investigate the possibility of channelling the results via existing networks such as the Global Wordnet Association, EAGLES-ISLE, ELDA, LDC and ELSNET. Channelling may involve direct mailing, collaborative meetings and workshops, or publication of results in newsletters. This will enlarge the scope of the project to many other languages and research groups, as well industrial parties in Europe.

- The results of MEANING will be public and free. This includes both the acquired knowledge in the Multilingual Central Repository and the technologies. For this purpose, we will develop open-source license agreements. The results can directly be down-loaded from the web and we will investigate distribution via agencies such as ELDA and LDC for minimal costs. Sharing of other resources and tools within the project will be agreed upon in the Consortium agreement.

## References

Agirre E. and Martínez D. *Exploring automatic word sense disambiguation with decision lists and the Web*. Proceedings of the Workshop “Semantic Annotation And Intelligent Annotation” organized by COLING 2000. Luxembourg. 2000.

Agirre E. and Martinez D. *Learning class-to-class selectional preferences*. Proceedings of the Workshop "Computational Natural Language Learning" (CoNLL-2001). In conjunction with ACL'2001/EACL'2001. Toulouse, France. 2001.

Agirre E. and Martinez D. *Decision Lists for English and Basque*. Proceedings of the SENSEVAL-2 Workshop at ACL/EACL'01, Toulouse, France. 2001.

Agirre E. and Martinez D. *Knowledge sources for Word Sense Disambiguation*. Proceedings of the Fourth International Conference TSD 2001, Plzen (Pilsen), Czech Republic. 2001. Published in the Springer Verlag Lecture Notes in Computer Science series. Václav Matousek, Pavel Mautner, Roman Moucek, Karel Tauser (eds.) Springer-Verlag.

Agirre E., Ansa O., Martínez D. and Hovy E. *Enriching WordNet concepts with topic signatures*. Proceedings of the NAACL workshop on WordNet and Other lexical Resources: Applications, Extensions and Customizations. Pittsburg. 2001.

Agirre E., Garcia E., Lersundi M., Martinez D. and Pociello E. *The Basque task: did systems perform in the upperbound?* Proceedings of the SENSEVAL-2 Workshop. In conjunction with ACL'2001/EACL'2001.

Toulouse, France. 2001.

Agirre E. and Martinez D. *Integrating selectional preferences in WordNet*. Proceedings of the first International WordNet Conference. Mysore, India, 2002.

Agirre E., Ansa O., Arregi X., Arriola J.M., Diaz de Ilarraza A., Pociello E. and Uria L. *Methodological issues in the building of the Basque WordNet: quantitative and qualitative analysis*. Proceedings of the first International WordNet Conference. Mysore, India. 2002.

Atserias J., Climent S., Farreres J., Rigau G. and Rodríguez H., *Combining Multiple Methods for the Automatic Construction of Multilingual WordNets*. Proceedings of the International Conference "Recent Advances on Natural Language Processing" RANLP'97. Tzigov Chark, Bulgaria, 1997.

Benítez L., Cervell S., Escudero G., López M., Rigau G. and Taulé M., *Methods and Tools for Building the Catalan WordNet*. Proceedings of the ELRA Workshop on Language Resources for European Minority Languages, First International Conference on Language Resources & Evaluation, Granada, Spain. 1998.

Blum A. and Mitchel T. *Combining labelled and unlabeled data with co-training*. In Proceedings of the 11th Annual Conference on Computational Learning Theory. 1998.

Carroll, J. and McCarthy, D. *Word sense disambiguation using automatically acquired verbal preferences*. In Computers and the Humanities. Senseval Special Issue, Vol. 34, No 1-2. 2000.

Daudé J., Padró L. and Rigau G., *Mapping Multilingual Hierarchies using Relaxation Labelling*, Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC'99). Maryland, United States, 1999.

Daudé J., Padró L. and Rigau G., *Mapping WordNets Using Structural Information* , 38th Anual Meeting of the Association for Computational Linguistics (ACL'2000). Hong Kong, October 2000.

Daudé J., Padró L. and Rigau G., *A Complete WN1.5 to WN1.6 Mapping*, Proceedings of NAACL Workshop "WordNet and Other Lexical Resources: Applications, Extensions and Customizations". Pittsburg, PA, United States, 2001.

Escudero G., Màrquez L. and Rigau G., *Boosting Applied to Word Sense Disambiguation*. Proceedings of the 11th European Conference on MachineLearning, ECML 2000. Barcelona, Spain. 2000. Lecture Notes in Artificial Intelligence 1810. R. L. de Mántaras and E. Plaza (Eds.). Springer Verlag 2000. Also as a research report LSI-00-03-R. Dept. de Llenguatges i Sistemes Informàtics. UPC. Barcelona. 2000.

Escudero G., Màrquez L. and Rigau G., *Naive Bayes and Exemplar-Based approaches to Word Sense Disambiguation Revisited*. Proceedings of the 14th European Conference on Artificial Intelligence, ECAI-2000. Berlin, Germany. 2000. Also as a research report LSI-00-05-R. Dept. de Llenguatges i Sistemes Informàtics. UPC. Barcelona. 2000.

Escudero G., Màrquez L. and Rigau G., A Comparison between Supervised Learning Algorithms for Word Sense Disambiguation. Proceedings of Fourth Computational Natural Language Learning Workshop (CoNLL-2000). Lisbon. Portugal. 2000.

Escudero G., Màrquez L. and Rigau G., *An Empirical Study of the Domain Dependence of Supervised Word Sense Disambiguation Systems*. Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC'00). Hong Kong, China. 2000.

Escudero G., Màrquez L. and Rigau G., *Using LazyBoosting for Word Sense Disambiguation*. Proceedings of 2nd International Workshop “Evaluating Word Sense Disambiguation Systems”, SENSEVAL-2. Toulouse, France. 2001.

Fellbaum C. editor. *WordNet An Electronic Lexical Database*. The MIT Press. 1998.

Ide, N. and Vèronis, J. *Introduction to the special issue on word sense disambiguation: The state of the art*. Computational Linguistics, 24 (1), 1998.

Kilgarriff A. and Rosenzweig J. *Framework and Results for English SENSEVAL*. Computers and the Humanities. 34. 2000.

Korhonen A., Gorrell, G. and McCarthy D. *Statistical Filtering and Subcategorization Frame Acquisition*. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. Hong Kong. 2000.

Leacock, C. Chodorow, M. and Miller, G.A. *Using Corpus Statistics and WordNet Relations for Sense Identication*, Computational Linguistics, 24(1), 1998.

Magnini B. & Cavaglià G., *Integrating subject field codes into WordNet*. In Proceedings of the Second Internatgional Conference on Language Resources and Evaluation LREC’2000, Athens. Greece. 2000.

Magnini B., Strapparava C., Pezzulo G. and Gliozzo A. *Using domain information for word sense disambiguation*. Proceedings of 2nd International Workshop “Evaluating Word Sense Disambiguation Systems”, SENSEVAL-2. Toulouse, France. 2001.

Martínez D. and Agirre E. *One Sense per Collocation and Genre/Topic Variations*. Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. Hong Kong, 2000.

McCarthy, D. and Korhonen, A. *Detecting verbal participation in diathesis alternations*. Proceedings of the 17th

International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics COLING-ACL'98. Montreal, Canada. 1998.

McCarthy D., *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Aternations, Subcategorization Frames and Selectional Preferences*. Ph.D. thesis, University of Sussex. 2001.

McCarthy, D., Carroll J. and J. Preiss. J. *Disambiguating noun and verb senses using automatically acquired selectional preferences.* Proceedings of the SENSEVAL-2 Workshop at ACL/EACL'01, Toulouse, France. 2001.

Mihalcea R. and Moldovan D. *An automatic method for generating sense tagged corpora*. In Proceedings of American Association for Artificial Intelligence, AAAI'99. 1999.

Miller G. *Five papers on WordNet,* Special Issue of International Journal of Lexicogrphy 3(4). 1990.

Miller G. Leacock C., Randee T. and Bunker R. *A Semantic Concordance*, in proceedings of the 3rd DARPA Workshop on Human Language Technology, 303-308, Plainsboro, New Jersey, March, 1993.

Ng, H. T. and Lee H. *Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach*. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics. 1996.

Ng. H. T. *Getting Serious about Word Sense Disambiguation*. In Proceedings of the ACL SIGLEX Workshop: Tagging Text with Lexical Semantics: Why, what and how?, Washington, USA, 1997.

Ribas F., On Acquiring Appropriate Selectional Restrictions from Corpora Using a Semantic Taxonomy. Ph.D. thesis, Universitat Politècnica de Catalunya. 1995.

Rigau G., Taulé M. Fernandez A. and Gonzalo J., *Framework and Results for the Spanish SENSEVAL*. Proceedings of 2nd International Workshop "Evaluating Word Sense Disambiguation Systems", SENSEVAL-2. Toulouse, France. 2001.

Roventini A., Alonge A., Calzolari N., Magnini B. and Bertagna F. *ItalWordNet: a Large Semantic Database for Italian*. In Proceedings of the Second Internatgional Conference on Language Resources and Evaluation LREC'2000, Athens. Greece. 2000.

Vossen P. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dordrecht. 1998.

Yarowsky D., *Unsupervised word sense disambiguation rivaling supervised methods*. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, ACL'95. 1995.

## 9.2 Workpackage List

| Work-package No | Workpackage title | Lead contractor No | Person-months | Start month | End month | Phase | Deliv-erable No |
|---|---|---|---|---|---|---|---|
| 0 | Project Management | UPC | 24 | 0 | 36 | R | D0.1<br>D0.2<br>D0.3 |
| 1 | User Requirements | UPC | 4 | 0 | 6 | R | D1.1 |
| 2 | Methodology and Design | UPV/EHU | 21 | 0 | 24 | R | D2.1 |
| 3 | Linguistic Processors and Infrastructure | ITC-IRST | 48 | 0 | 27 | R | D3.1<br>D3.2<br>D3.3 |
| 4 | Integration | UPC | 48 | 0 | 30 | R | D4.1<br>D4.2<br>D4.3 |
| 5 | Acquisition | UoS | 74 | 0 | 27 | R | D5.1<br>D5.2<br>D5.3 |
| 6 | Word Sense Disambiguation | UPV/EHU | 74 | 0 | 27 | R | D6.1<br>D6.2<br>D6.3 |
| 7 | Evaluation and Assessment | UoS | 26 | 10 | 30 | R | D7.1<br>D7.2<br>D7.3 |
| 8 | User validation | UPV/EHU | 6 | 13 | 36 | D | D8.1<br>D8.2<br>D8.3<br>D8.4 |
| 9 | Exploitation and dissemination | UPV/EHU | 28 | 0 | 36 | D | D9.1<br>D9.2<br>D9.3<br>D9.4 |
| | TOTAL | | 353 | | | | |

## 9.3 Workpackage descriptions

# Project Management

| **Workpackage number :** | 0 | | **Start date or starting event:** | | | M0 | | |
|---|---|---|---|---|---|---|---|---|
| **Participant:** | | **UPC** | UPV/EHU | | | | | |
| **Person-months per participant:** | | 12 | 12 | | | | | |

**Objectives**

To co-ordinate the project in accordance with the requirements of the Commission and the MEANING Consortium.

To plan, organise, monitor, control and lead the variety of activities needed by the MEANING project, that is study and research on advanced language processing methods, assessment of the research results, design and development of high quality software and resources enforcing best practices, user-driven monitoring and consistency checking of the achievements.

To deliver the MEANING software and resources within budget, according to the schedule and with the required quality level.

To manage all financial matters (preparing accounts, providing payments, etc).

To monitor the production of the reports and deliverables, assessing their quality and submitting them to the Commission.

**Description of work**

Project Management organisation and methodology will be responsible for the assessment and continuous monitoring of work progress and evaluation of achievements. Activities related to this workpackage will therefore address the following items:

- Overall administrative project management, which includes all aspects of financial and contractual administration. The Administrative Management Board (AMB) co-ordinated by UPC will be in charge for all the administrative, organisational, information and strategic issues.
- Overall technical project management. The Technical Management Board (TMB) co-ordinated by UPV/EHU will be responsible for all the technical issues (technical choices, monitoring supervision). All the technical activities at workpackage level (emission of deliverables, milestones achievements) will be co-ordinated by the respective WP leaders, which will report to the AMB, the TMB and to the Project Manager.

A specific group of deliverables are then foreseen to provide the various Boards, WP Leaders and the Project Manager with instruments to evaluate the current state of the project.

Finally, a Periodic Progress Report will be provided to report the project progress intermediate steps, milestones and results achieved.

**Deliverables**

- D0.1 Consortium Agreement.
- D0.2 Periodic Progress Reports. See Appendix X for further details.
- D0.3 Periodic Management Reports. See Appendix X for further details.

**Milestones and expected result**

M36: Project delivered according to the time schedule within budget.

# User-requirements

| **Workpackage number :** | 1 | | **Start date or starting event:** | | | M0 | | |
|---|---|---|---|---|---|---|---|---|
| **Participant:** | | **UPC** | UPV/EHU | | | | | |
| **Person-months per participant:** | | 3 | 1 | | | | | |

**Objectives**

To provide the criteria and perspective to design the architecture and methodology of the project and the scope of the work. The industrial partners in the project will address the short-term exploitation and usage of the MEANING results in language technology and software applications. From this, they will derive the general criteria and directions for MEANING in the form of explicit user-requirements. These requirements will be the basis for developing the verification criteria in WP8 but will also be input to the architecture and design of the project and the definition of the methodologies.

In addition, the users will also investigate the longer-term scope of the technology and resources developed in MEANING. This is particularly important because of the innovative character of the technology. The long-terms scope will not be used for the evaluation criteria of the results but will be incorporated in the dissemination plans and activities.

**Description of work**

- Inventorizing the applications and the language technology that can benefit from the MEANING results.
- Defining the short-term and long-term scope of MEANING technology.
- Formulating the general requirements for MEANING outcomes to be of use for industrial exploitation.

**Deliverables**

D1.1. User-requirements for MEANING [months 6, 15, 24]

**Milestones and expected result**

M6: Specification of the user-requirements and basic design of the architecture and methodologies of MEANING.

# Methodology and Design

| **Workpackage number :** | 2 | | **Start date or starting event:** | | M0 | | |
|---|---|---|---|---|---|---|---|
| **Participant:** | **UPC** | ITC-IRST | UPV/EHU | UoS | | | |
| **Person-months per participant:** | 6 | 6 | 6 | 3 | | | |

**Objectives**

To define the overall methodology of MEANING including the standard protocols, formats, procedures, and evaluation criteria and the Multilingual Central Repository database.

The partners will define all the requirements for all modules involved in MEANING. For each MEANING task and cycles all requirements must be identified i.e. requirements for the Language Processors and infrastructure (WP3). This workpackage will produce also for each cycle the information flow and formats for uploading and porting data to the Multilingual Central Repository (WP4), for the acquisition process (WP5), word sense disambiguation (WP6) and the evaluation criteria needed for measuring the quality of the tools and resources produced by MEANING (WP7). The partners will also define the main functionalities and the linguistic content to be represented into the Multilingual Central Repository.

**Description of work**

- Inventorizing the applications, resources and language technology currently available by the partners.
- To identify the requirements for the Language Processors and infrastructure (WP3) to be used in MEANING.
- To define the linguistic content to be represented into the Multilingual Central Repository.
- To define the timing, information flow and formats of the acquisition, word sense disambiguation, uploading and porting cycles.
- To design the main functionality of the Multilingual Central Repository (WP4) including:
  - The process for uploading the data acquired from one language to the Multilingual Central Repository.
  - The process for porting the knowledge stored in the Multilingual Central Repository to the respective wordnets.
- To define the assessment and evaluation criteria to be used in WP7.

**Deliverables**

D2.1. Basic design of the architecture and methodologies of MEANING [months 6, 15, 24].

**Milestones and expected result**

M6: Specification of the user-requirements and basic design of the architecture and methodologies of MEANING.

# Linguistic Processors and Infrastructure

| **Workpackage number :** | 3 | | **Start date or starting event:** | | M0 | | |
|---|---|---|---|---|---|---|---|
| **Participant number:** | UPC | **ITC-IRST** | UPV/EHU | UoS | | | |
| **Person-months per participant:** | 16 | 16 | 8 | 8 | | | |

**Objectives**

To equip each partner with sufficient computational resources and internet bandwidth.
To provide robust language processing software tools (for English, Italian, Spanish, Catalan and Basque) which will form part of the systems developed for acquisition (WP5) and word sense disambiguation (WP6) including:

- tokenisation and sentence boundary detection
- lemmatisation
- part of speech tagging
- noun group chunking
- robust shallow parsing
- named-entity recognition and categorisation (e.g. into location, company or product names)
- keyword, topic and terminology detection
- text classification (e.g. ECONOMIC, SPORT domains)

**Description of work**

- Equipping each partner with fast internet access, and sufficient processing power and storage space.
- Re-engineering and scaling up existing robust language processing software tools in accordance with the WP1 assessment of the software requirements for MEANING.
- Further developing and refining tools directed by assessment of their accuracy and speed within the context of WP5 and WP6.

**Deliverables**

D3.1 ELP (English language processor), ILP (Italian language processor), SLP (Spanish language processor), CLP (Catalan language processor) and BLP (Basque Language processor) Software (first release).

D3.2 ELP, ILP, SLP, CLP and BLP Software (second release).

D3.3 ELP, ILP, SLP, CLP and BLP Software (final release).

**Milestones and expected result**

M9: First release of the Linguistic Processors (English, Italian, Spanish, Catalan and Basque)
M18: Second release of the Linguistic Processors (English, Italian, Spanish, Catalan and Basque)
M27: Final release of the Linguistic Processors (English, Italian, Spanish, Catalan and Basque)

# Integration

| **Workpackage number :** | 4 | | **Start date or starting event:** | | | M0 | |
|---|---|---|---|---|---|---|---|
| **Participant:** | **UPC** | ITC-IRST | UPV/EHU | UoS | | | |
| **Person-months per participant:** | 18 | 10 | 10 | 10 | | | |

**Objectives**

To design and develop the Multilingual Central Repository for uploading and porting the knowledge acquired across languages and resources and maintaining the compatibility among them.

The Multilingual Central Repository will include modules for:
- Uploading the data acquired from one language to the Multilingual Central Repository.
- Porting the knowledge stored in the Multilingual Central Repository to the local wordnets.
- Checking the integrity of the data stored in the Multilingual Central Repository.

To perform the process for uploading and porting the knowledge acquired from each language to the respective local wordnets.

**Description of work**

- To design and develop the Multilingual Central Repository
- To perform PORT0, PORT1 and PORT2 uploading and porting process.

**Deliverables**

D4.1 PORT0. First release of the Multilingual Central Repository, software and data uploaded and ported in PORT0.
D4.2 PORT1. Second release of the Multilingual Central Repository, software and data uploaded and ported in PORT1.
D4.3 PORT2. Final release of the Multilingual Central Repository, software and data uploaded and ported in PORT2.

**Milestones and expected result**

M12: End of PORT0. First release of the Multilingual Central Repository and data from PORT0.
M21: End of PORT1. Second release of the Multilingual Central Repository and data from PORT1.
M30: End of PORT2. Final release of the Multilingual Central Repository and data from PORT2.

# Acquisition

| **Workpackage number :** | 5 | | **Start date or starting event:** | | M0 | | |
|---|---|---|---|---|---|---|---|
| **Participant:** | UPC | ITC-IRST | UPV/EHU | **UoS** | | | |
| **Person-months per participant:** | 18 | 18 | 18 | 20 | | | |

**Objectives**

To design and apply advanced unsupervised learning techniques for acquiring lexical syntactical and semantic information from large quantities of unannotated text and latter on with annotated text.

This workpackage will develop automatic techniques to acquire domain terminology, sense-labelled subcategorization frequencies, topic signature, domain information, new senses, coarser-grained sense clusters, specific lexico-semantic relations, selectional preferences, thematic role assigments for nominalizations, and diathesis alternations.

The results will be evaluated in WP7 (Evaluation and Assessment).

**Description of work**

- Downloading text for each language from automatically-selected web documents and applying the WSD system (WP6) resulting from the previous acquisition-porting cycle.
- First acquisition phase, using already existing knowledge in the local wordnets, will learn subcategorisation frequencies, topic signatures, terminology and domain information.
- Second phase will key them to word senses induced by WSD0, and will additionally learn new senses, coarser-grained sense clusters and selectional preferences.
- Third phase will learn specific lexico-semantic relations and thematic role assignments for nominalizations and diathesis alternations.
- Storing the information in the local wordnets for the respective language, and uploading it to the Multilingual Central Repository (WP4).

**Deliverables**

D5.1 ACQ0. First release of the acquisition software and data acquired from ACQ0.
D5.2 ACQ1. Second release of the acquisition software and data acquired from ACQ1.
D5.3 ACQ2. Final release of the acquisition software and data acquired from ACQ2.

**Milestones and expected result**

M9: End of ACQ0. First release of the acquisition software and data from ACQ0.
M18: End of ACQ1. Second release of the acquisition software and data from ACQ1.
M27: End of ACQ2. Final release of the acquisition software and data from ACQ2.

# Word Sense Disambiguation

| **Workpackage number :** | 6 | | **Start date or starting event:** | | M0 | | |
|---|---|---|---|---|---|---|---|
| **Participant:** | UPC | ITC-IRST | **UPV/EHU** | UoS | | | |
| **Person-months per participant:** | 18 | 18 | 20 | 18 | | | |

**Objectives**

Produce a high-precision system that is able to tag all open class words in running text with word senses in real time. The system will be evaluated in WP7 (Evaluation and Assessment), and will use the following:

- Linguistic features extracted from the corpus using the linguistic processors (WP3). The linguistic features will include the linguistic knowledge acquired in WP5 and stored in the Multilingual Central Repository for each of the target word senses (WP4), including automatically retrieved examples for each word sense (WP5)
- State of the art Machine Learning techniques (e.g. decision lists, boosting) to learn from the training features.
- Extensions of the Machine Learning techniques to richer features (e.g. including syntactic information).
- Extensions of the Machine Learning techniques to use unannotated examples (e.g. co-training and novel transductive models of margin-based classifiers).

**Description of work**

The work involved includes the following:

- The extension of state-of-the-art ML techniques to profit form rich linguistic features
- The extension of state-of-the-art ML techniques to profit from untagged examples, e.g. co-training.

The system will be developed in three steps:

1. A baseline system (WSD0) using the extended unsupervised Machine Learning techniques and standard linguistic features, trained on a number of examples automatically acquired from the web.
2. A first system (WSD1) using the extended unsupervised Machine Learning techniques, richer linguistic features as provided by the linguistic processors and the Multilingual Central Repository (ACQ0). It will be trained with a bigger number of automatically retrieved examples for all polysemous words in the Multilingual Central Repository.
3. A final system (WSD2) using the technology developed in WSD1, but using the improved central repository (ACQ1) and trained on a significant part of the web for all polysemous words.

**Deliverables**

D6.1 WSD0. First release of the WSD system and the corpus disambiguated with WSD0.
D6.2 WSD1. Second release of the WSD system and the the corpus disambiguated with WSD1.
D6.3 WSD2. Final release of the WSD system and corpus disambiguated with WSD2.

**Milestones and expected result**

M9: End of WSD0. First release of the WSD system and the corpus annotated using WSD0
M18: End of WSD1. Second release of the WSD system and the corpus annotated using WSD1.
M27: End of WSD2. Final release of the WSD system and the corpus annotated using WSD2.

# Evaluation and Assessment

| **Workpackage number :** | 7 | | **Start date or starting event:** | | M10 | | |
|---|---|---|---|---|---|---|---|
| **Participant:** | UPC | ITC-IRST | UPV/EHU | **UoS** | | | |
| **Person-months per participant:** | 6 | 6 | 6 | 8 | | | |

**Objectives**

To evaluate the quality and accuracy of the developped software and the acquired data. Objective measures on significant samples of the data and software results will be provided.

To assess the progress of the project, and if necessary, provide the necessary informatio to devise corrective actions.

**Description of work**

To evaluate and asses each of the two releases of the linguistic processors.

To evaluate and asses each of the three releases of the knowledge acquired and uploaded to the Multilingual Central Repository.

To evaluate and asses each of the three releases of the ported wordnets.

To evaluate and asses each of the three releases of the WSD system.

**Deliverables**

D7.1 Results of the first evaluation and assesment of linguistic processors, knowledge acquired, knowledge ported and WSD system

D7.2 Results of the second evaluation and assessment of linguistic processors, knowledge acquired, knowledge ported and WSD system

D7.3 Results of the final evaluation and assessment of knowledge acquired, knowledge ported and WSD system

**Milestones and expected result**

M12: End of first evaluation and assesment of the first cycle.

M21: End of second evaluation and assesment of second cycle.

M30: End of final evaluation and assesment of final cycle.

# User-validation

| **Workpackage number :** | 8 | | **Start date or starting event:** | | | M13 | | |
|---|---|---|---|---|---|---|---|---|
| **Participant:** | | UPC | **UPV/EHU** | | | | | |
| **Person-months per participant:** | | 3 | 3 | | | | | |

**Objectives**

To verify the progress of the project with respect to the user-requirements, to provide feedback to the project partners and to demonstrate the feasibility of integrating the project results in an existing language technology product.

The evaluation is separated in two separate task:

- verification of the intermediate results after each production cycle in MEANING
- demonstration of MEANING by integrating the results in the existing products of the company, including mono and multilingual Information Retrieval, multilingual Question Answering, classification, routing and filtering systems.

The purpose of the verification is to check whether the results satisfy the user-requirements and to provide the project with feedback on the applied methodology. The verification will directly assess the evaluation criteria formulated in WP1 (user-requirements) to the intermediate project results. This will result in a report after each cycle, stating the quality of the results according to these criteria.

The demonstration will show the feasibility of integrating the project results into an existing industrial environment. Demonstration will be carried out at the final project workshop or review.

**Description of work**

1. Specification of formal criteria for evaluationg the intermediate results of MEANING.
2. Assessing the formal criteria to each cycle in MEANING
3. Providing feedback to the consortium from a user-perspective
4. Adapting the commercial technology to incorporate WSD technology and MEANING resources
5. Converting the MEANING results to the application environment
6. Preparing the demonstration of the integrated MEANING results
7. Reporting on the demonstration

**Deliverables**

D8.1 Verification of MEANING-1, Report

D8.2 Verification of MEANING-2, Report

D8.3 Verification of MEANING-3, Report

D8.4 Demonstration of MEANING in Language Technology, Report

**Milestones and expected result**

M36: Integration of the MEANING technology and resources in existing commercial language technology and the demonstration of the integration in an end-user application.

# Exploitation and Dissemination

| **Workpackage number :** | 9 | | **Start date or starting event:** | | M0 | | |
|---|---|---|---|---|---|---|---|
| **Participant number:** | UPC | ITC-IRST | **UPV/EHU** | UoS | | | |
| **Person-months per participant:** | 5 | 9 | 9 | 5 | | | |

**Objectives**

Main objective of the activities in this WP is to define the route to the exploitation of the MEANING results. This includes clarity over IPR, responsibilities, territories and post-RTD collaboration. This will be formalized in the Consortium Agreement and the public License Agreements of the MEANING results. The License Agreements will be based on the open-source model.

Moreover, the work will be focused on ensuring awareness of the MEANING project existence, aims and expected results amongst the language-technology and web-related companies. The users and industrial partners, will be in charge to stay tuned to market evolution and demand for the product or service that will be the final result of the MEANING project. As for Universities and Research associations, whose main functions are teaching and research, they will mainly exploit by diffusion of expertise and research results, to business by consultancy and availability for contractual research and development, and otherwise by the normal dissemination channels of publication, conference papers, organised workshops and courses, and the world-wide-web.

In addition, the project will organise two workshops inviting relevant researchers involved in the MEANING technologies. The first workshop will be carried out during the first year of the project to obtain detailed feedback for MEANING. The second workshop will be held at the third year of the project. In this case the main goal will be to present the main results of MEANING and to promote an in depth discussion about the main achievements and drawbacks of the MEANING technology

**Description of work**

1. Consortium agreement stating the IPR and rights for using resources and technology within the consortium.
2. Dissemination plan targeting at actual events and networks to promote MEANING.
3. Participation in international workshops, conferences and evaluation schemes, especially Senseval and Trec.
4. Organisation of two workshops: one during the first year of the project organised by UPV/EHU and second during the thrid year organised by ITC-IRST.
5. Development of license and distribution agreements for the project results.
6. Publication of the results on the web.
7. Studying the route to the market and the possibilities for further development of MEANING resources and technology.

**Deliverables**

D9.1 Project Presentation.

D9.2 Dissemination and Use Plan.

D9.3 Public Reports (including annual and final public reports).

D9.4 Technology Implementation Plan and Distribution agreements.

**Milestones and expected result**

M30: Distribution agreements D9.4) should be completed and signed by all involved parties.

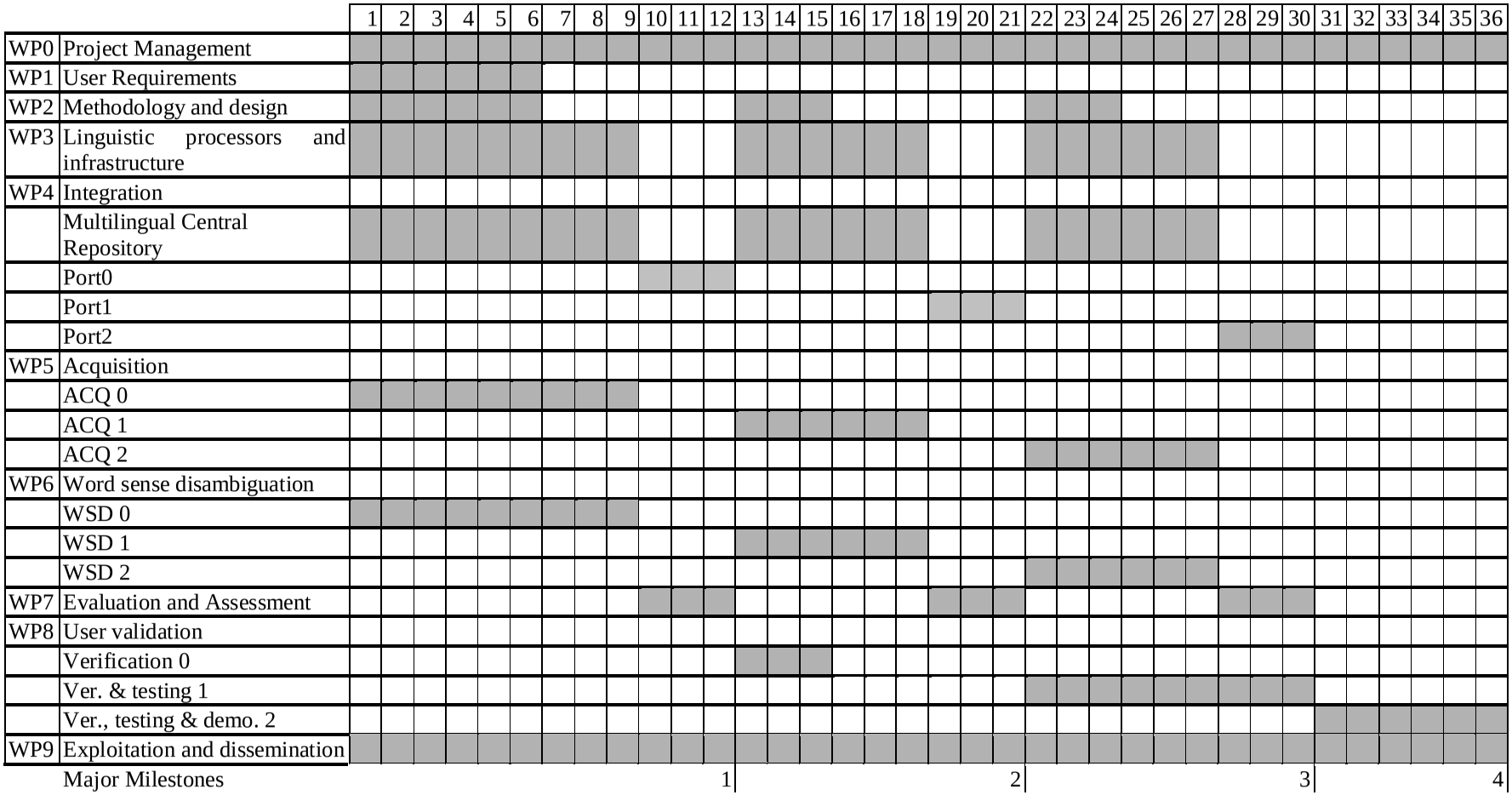
M36: Results of MEANING are published on the web.

## 9.4 Deriverables List

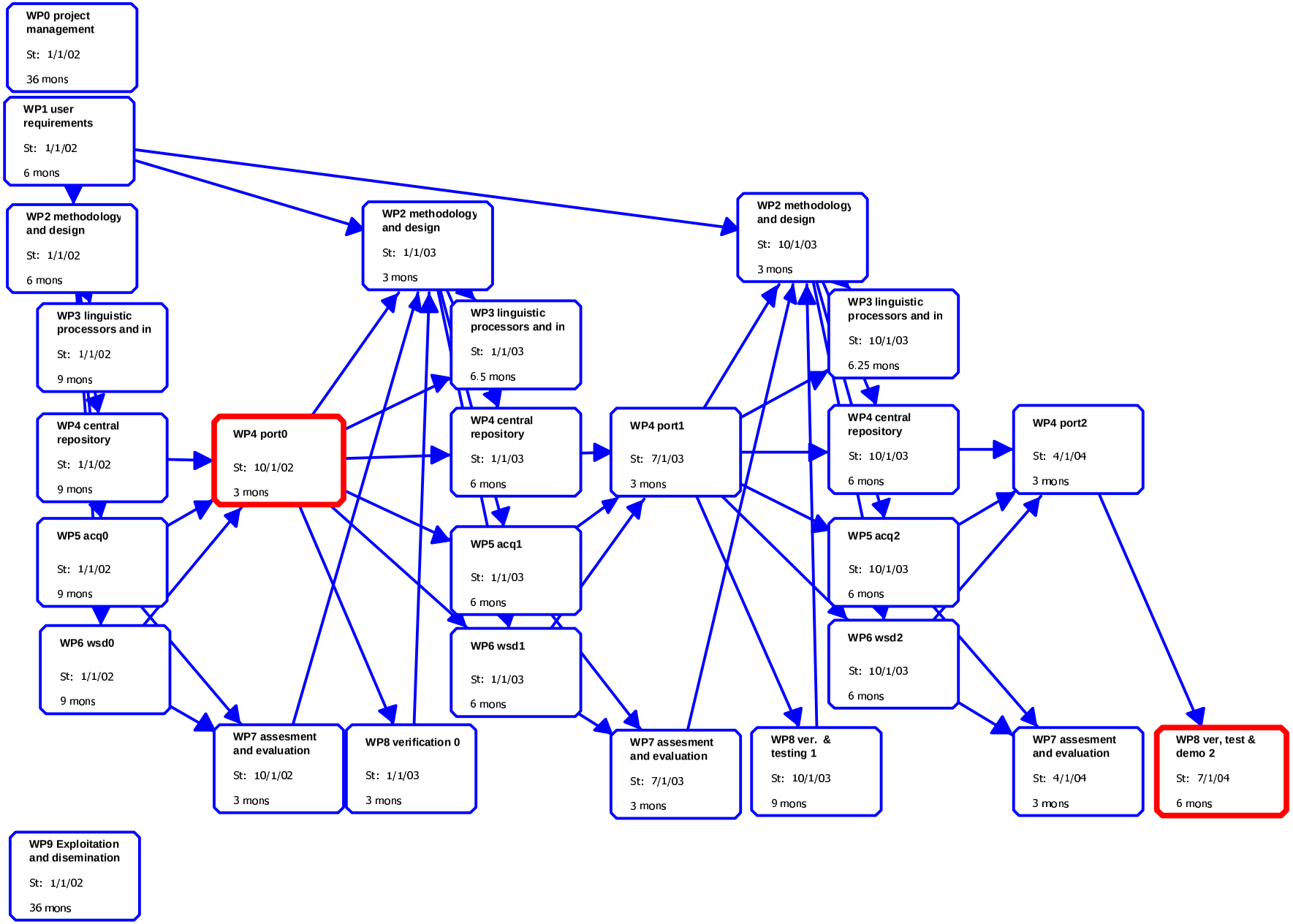
| Del. No | Deliverable name | WP no. | Lead participant | Estimated person-months | Del. type | Security | Delivery date |
|---|---|---|---|---|---|---|---|
| D0.1 | Consortium Agreement | 0 | UPC | 1 | Agreement | Pub. | 2 |
| D0.2 | Periodic Progress Report | 0 | UPV/EHU | 3 | Report | Rest. | See Appendix X |
| D0.3 | Periodic Management Report | 0 | UPC | 3 | Report | Rest. | See Appendix X |
| D1.1 | User-requirements | 1 | UPC | 1 | Report | Pub. | 6, 15, 24 |
| D9.1 | Project presentation | 9 | UPV/EHU | 1 | Presentation | Pub. | See Appendix X |
| D9.2 | Dissemination and Use Plan | 9 | UPV/EHU | 2 | Report | Pub. | 6 |
| D2.1 | Basic design of the architecture and methodologies | 2 | UPV/EHU | 3 | Report | Pub. | 6, 15, 24 |
| D3.1 | First release of the Linguistic Processors | 3 | ITC-IRST | 2 | Prototype | Int. | 9 |
| D5.1 | ACQ0 | 5 | UoS | 2 | Report Prototype Data | Pub. | 9 |
| D6.1 | WSD0 | 6 | UPV/EHU | 2 | Report Prototype Data | Pub. | 9 |
| D4.1 | PORT0 | 4 | UPC | 2 | Report Prototype Data | Pub. | 12 |
| D7.1 | Evaluation and assessment of MEANING 1 | 7 | UoS | 2 | Report | Pub. | 12 |
| D8.1 | Validation of MEANING 1 | 8 | UPV/EHU | 2 | Report | Pub. | 15 |
| D3.2 | Second release of the Linguistic Processors | 3 | ITC-IRST | 1 | Prototype | Int. | 18 |
| D5.2 | ACQ1 | 5 | UoS | 2 | Report Prototype Data | Pub. | 18 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| D6.2 | WSD1 | 6 | UPV/EHU | 2 | Report Prototype Data | Pub. | 18 |
| D4.2 | PORT1 | 4 | UPC | 2 | Report Prototype Data | Pub. | 21 |
| D7.2 | Evaluation and assessment of MEANING 2 and technology watch | 7 | UoS | 3 | Report | Pub. | 21 |
| D8.2 | Validation of MEANING 2 | 8 | UPV/EHU | 2 | Report | Pub. | 24 |
| D3.3 | Final release of the Linguistic Processors | 3 | ITC-IRST | 1 | Prototype | Int. | 27 |
| D5.3 | ACQ2 | 5 | UoS | 2 | Report Prototype Data | Pub. | 27 |
| D6.3 | WSD2 | 6 | UPV/EHU | 2 | Report Prototype Data | Pub. | 27 |
| D4.3 | PORT2 | 4 | UPC | 2 | Report Prototype Data | Pub. | 30 |
| D7.3 | Evaluation and assessment of MEANING 3 | 7 | UoS | 2 | Report | Pub. | 30 |
| D8.3 | Validation of MEANING 3 | 8 | UPV/EHU | 2 | Report | Pub. | 30 |
| D8.4 | Demonstration of MEANING | 8 | UPV/EHU | 2 | Demonstration | Pub. | 36 |
| D9.3 | Public reports | 9 | UPV/EHU | 4 | Report | Pub. | See Appendix X |
| D9.4 | Technology Implementation Plan | 9 | UPV/EHU | 2 | Report | Pub. | See Appendix X |

## 9.5 Project Planning and Timetable



## 9.6 Graphical Presentation of project components



## 9.7 Project management

### 9.7.1 Management Structure and Techniques

The management of the project will be organised with the following structure:

- An **Administrative Management Board (AMB)** co-ordinated by UPC and composed of one Management Representative from each partner with the aim of monitor and managing items that affect the contractual terms of the project.
- A **Technical Management Board (TMB)** co-ordinated by UPV/EHU and composed by technical experts coming from each partner, which is in charge to take decisions on all technical issues.
- A **Project Director** from UPV/EHU to assume the technical direction of the project, who will, be the chair of TMB, be member of AMB and be the main reference for the EC.
- An **Administrative Director**, from UPC, and reporting to the Project Director.

### 9.7.2 Project Co-ordination and Leadership

**Project Director**

The technical leadership of the project will be assured by the Project Director, who will be in charge of its day-to-day running, with responsibility for implementing decisions taken by the AMB and TMB, and taking decisions between meetings. This person will also be the principal interface of the project towards the EC, be responsible for the submission of periodic management reports to the EC, and ensure that the Consortium fulfils all its contractual responsibilities towards the Commission, including those in respect of submission of cost statements. This person will be member of both AMB and TMB.

**Administrative Management Board**

An Administrative Management Board will be created to specifically address the main administrative, organisational, information and strategic decisions concerning the project. It is composed of one Management Representative from each partner and is chaired by the Project Director. The AMB is formally empowered by the Consortium Agreement to take decisions affecting the budget and the objectives of the project, changes and exploitation agreements, and is the highest authority for conflict resolution, with the Project Director having a casting vote if necessary. The AMB will meet when needed, regularly.

**Technical Management Board**

The Technical Management Board, made of technical experts from all the partners will be chaired by the Project Director, will handle the technical management and execution of the project. It will take day-to-day technical decisions, with the participation of experts when necessary and will report to the AMB. It will implement the strategy, the choice of techniques, supervising the monitoring of the results.

The TMB will work extensively by using electronic mail, Web based bulletin boards and will meet

regularly. Meetings will usually be held at the same time as the AMB and more often only if required.

**Workpackage direction**

Each workpackage and each task is under the responsibility a workpackage leader. She/he organises the suitable contacts between the concerned partners and is in charge of producing the project deliverables.

### 9.7.3 Project Administrative Management

**Administrative Director**

In line with the guidelines suggested by the EC, a separate project administration function will be maintained. This will be under the responsibility of the Administrative Director, a senior administrator nominated by the administrative co-ordinating partner. This person will report to the Project Director, and be in charge of Financial and Project Administration. Financial management will include support for cost claims and interfacing the EC on this subject, management of payments to partners, budget control and monitoring, support to partners for cost-related interaction with the EC. Project administration tasks include project contracts, assemblage of material for periodic reporting, internal information exchange, meeting preparation and follow-up.

**Reference documents**

Project co-ordination will be guided by major reference documents that define the objectives, the work programme and the operational procedures of the MEANING project:

- The MEANING Project programme.
- The Consortium Agreement to be signed between the partners to specify issues not included in the European Commission contract (decision procedure, conflict resolution, exploitation, etc.)
- The Implementation Plan, containing structure and contents of all the deliverables, time schedule (including internal milestones and achievements) and (internal) interactions of all the project activities.

**Communications**

Daily communication between all participants will be assured using electronic mail and Web based bulletin boards.

# 10. CLUSTERING

MEANING will participate to IST support activities in the framework of the CLASS[3] (Collaboration in Language and Speech Science and Technology) initiative in the area of Crosslingual Information and Knowledge Management. Among the goals of this cluster there is the specification of a standard reference platform/architecture which could serve as a base

[3] http://www.class-tech.org/

for LT-based improvements in language identification, summarisation, categorisation, retrieval, clustering, relevance ranking, information extraction, information presentation/visualisation, and knowledge discovery. The next meeting of this cluster will be held at LREC, May 2002, in Las Palmas.

Several HLT projects are associated to this CLASS cluster, including BINDEX, C-ORAL-ROM, CLARITY, CROSSMARC, KERMIT, LIMBER, LIQUID, MEMPHIS, MKBEEM, MUCH MORE, MUMIS, NAMIC, PEKING, SAFE, TQPRO and CLEF, which is of particular interest for MEANING. CLEF, Text Retrieval System Evaluation activity, co-ordinated in Europe by the DELOS Network of Excellence for Digital Libraries and organised in collaboration with the US National Institute of Standards and Technology (NIST) and the TREC Conferences. The CLEF series of system evaluation campaigns aims at promoting research and development in Cross-Language Information Retrieval by (i) providing an infrastructure for the testing and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts, and (ii) creating test-suites of reusable data which can be employed by system developers for benchmarking purposes.

# 11. OTHER CONTRACTUAL CONDITIONS

## 11.1 Subcontracting

UPV/EHU will subcontract Irion Technologies BV (Netherlands) for a total amount of 70000 Euro, and UPC will subcontract Reuters (UK) for 30000 Euro. The subcontractors will contribute to Workpackages 1 (user requirements), 8 (user validation) and 9 (exploitation and dissemination).

Irion Technologies have large experience on using and porting large-scale NLP systems from one domain to another for a particular end-user application. MEANING will also contribute to EU policies on fostering the uptake of technology developments by SMEs, subcontracting Irion Technologies BV.

## 11.2 Travel outside the EU Member States and Associated States

The exploitation and dissemination activities (Workpackage 9; section 9.1) call for project participants to present results of the project at scientific events, fairs, workshops and conferences, particularly targetting evaluation schemes such as SENSEVAL and TREC. These activities may involve travel outside EU member states, for example to the US and Asia. The project participants therefore take it that Commission authorisation is given to use project funds to attend the following relevant and prestigious international conferences and evaluation exercises, either to present papers or to organise workshops affiliated to them:

- ACL / EACL / NAACL (Annual Meeting of the Association for Computational Linguistics, also European and North American ACL Chapter Meeting)
- COLING (International Conference on Computational Linguistics)
- EMNLP / WVLC (ACL SIGDAT Empirical Methods in Natural Language Processing /

Workshop on Very Large Corpora)

- HLT (Human Language Technology Conference)
- IJCAI (International Joint Conference on Artificial Intelligence)
- International Conference on Wordnet
- IWPT (ACL SIGPARSE International Workshop on Parsing Technologies)
- LREC (International Conference on Language Resources and Evaluation)
- SENSEVAL (Evaluating Word Sense Disambiguation Systems)
- TREC (Text REtrieval Conference)

For other travel outside the EU and Associated States, the Project Officer will be asked for specific approval in advance.

## 11.2 Other Specific Project Costs

Each partner has budgeted for producing one audit certificate: the budgeted cost for full cost partners being 8000 Euro each, and for additional cost partners, 4000 Euro.

In support of the exploitation and dissemination actions (Workpackage 9), UPV/EHU will organise a project workshop during the first year of the project, and ITC-IRST will organise one during the third year, each at a cost of 24000 Euro. MEANING will invite to these workshops researchers involved in technologies relevant to the project.

# APPENDIX A – CONSORTIUM DESCRIPTION

The MEANING consortium is composed by four academic partners (UPC, ITC-IRST, UPV/EHU and UoS). Partners in the consortium, with the exception of UPV/EHU, have a long tradition in participating in EU actions, having co-ordinated several successful EU projects.

The academic partners are all well-known players in Language Technology, and specially Natural Language Processing. They have a long profile developing language resources and have been involved in the construction of wordnets for languages other than English: UPC for Spanish and Catalan, ITC-IRST for Italian and UPV/EHU for Basque. The representative of Irion was the project manager of the LE-EuroWordNet project, which guarantees both a high degree of competence and a strong commitment of the industrial partner in all the project phases. All the involved groups have a strong motivation in developing technologies to enhance the usability of wordnets and using them in real applications. Also all partners are involved with WSD, as demonstrated by their participation at the SENSEVAL competition. Although there is a broad overlapping in their interests, partners have addressed the WSD problem from different, even if complementary, points of view, which makes the MEANING collaboration a unique opportunity to produce significant scientific impulse.

The role of each partner in the workplan reflects this situation. UPC, UoS, UPV/EHU and ITC-IRST will all work on their respective languages, each providing resources and tools to be used in the acquisition/porting cycle of the project. Responsibilities are balanced among the partners, with UPC serving as coordinator and Irion being responsible for user requirements and user validation.

The integration of the knowledge acquired and uploaded into the Multilingual Central Repository will be co-ordinated by the UPC. They will develop the technology for the automatic alignment of large-scale and complex knowledge bases. This technology will provide compatibility to the Multilingual Central Repository across the European wordnets, past and new. ITC-IRST invested a large effort studying the linguistic relations between the levels of information present in documents and their representation into WordNet. They will co-ordinate the development of the Linguistic Processors. UPV/EHU, having much experience studying the performance and developing efficient Knowledge-based and Machine Learning algorithms for WSD, will co-ordinate this part of the project. UoS has carried out pioneering work on large-scale automatic acquisition of linguistic knowledge and their application to WSD. So, they will co-ordinate this process and also the evaluation and assessment workpackage.

The Consortium will subcontract two companies: Reuters and Irion. Over the past 150 years, the news agency has led the way with new innovations in the dissemination and use of news and information. We plan to use their expertise toapply the MEANING technologies in real scenarios. Irion Technologies is a software company that provides linguistic software products. They add value to search engines by combining natural language technology with information retrieval. Their products include translation between a large number of languages, automatic classification, automatic hyperlinking, multilingual semantic networks and multimedia processing tools.

Least, but not last, people involved in the project know each other since a long time, which is an additional and important guarantee of cohesion of the Consortium.

**1) TALP Research Center, Universitat Politècnica de Catalunya (UPC)**

Participant's address: Jordi Girona, 1-3
E-08034 Barcelona
Spain
URL: http://www.talp.upc.es
Director: Climent Nadeu

TALP (Research Center for Language and Speech Technology and Applications) is a Specific Research Center in Universitat Politècnica de Catalunya (UPC), devoted to technology and applications of the natural language processing techniques, either for spoken or written language. It's formed by two research groups in UPC: The Natural Language Processing Research Group from the Software Department (LSI), and the Speech Processing Group from the Signal Theory and Communications Department (TSC).

TALP Research Center belongs to ELSNET (European Network of Excellence in Human Language Technologies) and is members of the Reference Center in Language Engineering (CREL) of the Catalan government. There are 37 researchers working at TALP, 26 of them are lecturers in the Telecommunications or Computer Science curricula at UPC. Since the academic year 1999-2000 TALP has been offering the new European Master in Language and Speech.

*Relevant European project references*

TALP has been active in many successful Third and Fourth Framework projects, in some case with the role of coordinating partner. ACQUILEX (Esprit), ACQUILEX II (Esprit), EuroWordNet (LE), NAMIC (IST), SpeechDat (TELEMATICS), VIDAS (ACTS), HANDY (CRAFT), SALA, SpeechDat-Car (TELEMATICS), COST 250, INTERFACE (IST), FAME (IST).

**CV: German Rigau**

Ph.D. and B.A. in Computer Science by the Universitat Politecnica de Catalunya (UPC). Currently teaching at the Computer Science Faculty of the UPC as an Associate Professor. He is also doing research at the TALP Research Center of the UPC. He has published over thirty-refereed articles and conference papers in the area of Natural Language Processing and in particular Acquisition of Lexical Knowledge and Word Sense Disambiguation. He has been involved in several European research projects (ESPRIT BRA ACQUILEX, ACQUILEX II, LE EUROWORDNET, LE NAMIC) and Spanish National research projects (ITEM, HERMES). He has also participated in both last editions of the international competition of SENSEVAL. Currently, he is member of the Association for Computational Linguistics (ACL) and the Spanish Society for Natural Language Processing (SEPLN).

**CV: Horacio Rodríguez**

Industrial Engineer by the Universitat Politècnica de Catalunya (UPC). Degree on Physics by the Universitat de Barcelona (UB) and Ph.D. in Computer Science by the Universitat Politècnica de Catalunya (UPC). Currently teaching at the Computer Science Faculty of the UPC. He is also a researcher at the TALP Research Centre of the UPC. He has published over thirty refereed articles and conference papers in the area of Natural Language Processing and in particular Acquisition of Lexical Knowledge and using empirical methods in NLP tasks. He has been advisor of six Ph.D. Thesis. He has been involved in several European research projects (ESPRIT BRA ACQUILEX, ESPRIT BRA ACQUILEX II, LE EUROWORDNET, LE

NAMIC) and Spanish National research projects (ITEM, HERMES). Currently, he is member of the European Association for Artificial Intelligence (ECCAI) and the Spanish Society for Natural Language Processing (SEPLN).

**2) Istituto Trentino di Cultura - Instituto per la Ricerca Scientifica e Tecnologica (ITC-IRST)**

Participant's address: via Santa Chiara
38100 Trento, Italy
Director: Oliviero Stock

The Istituto Trentino di Cultura (created 1962 by the Autonomous Province of Trento) has as its objective both scientific excellence and innovation and technology transfer to companies and public services. In its areas of competence, ITC collaborates with the main actors in world-wide research and it works in synt ony with the European Union Programs. The total budget is currently about 17 M Euro. Research activities are carried in scientific and technological areas, advanced computer science, microelectronics, physics, mathematical sciences and in human sciences. IRST, ITC Centre for Scientific and Technological Research, is a point of reference in the international scientific community and, at the same time, a hub for the development of technologies and application s with social and economical impact. Personnel at ITC-IRST is about one hundred people on a permanent basis, and about 50 people on "soft" money. Altogether ITC-IRST budget amounts to about 10 MEuro. Half of ITC-IRST direct costs are covered by industrial contracts and European and National contracts. So far over 40 European contracts of diverse kind have been carried on by ITC-IRST. A substantial portion of ITC-IRST activities are in information technology (mostly in user - friendly and intelligent systems), with projects organised in three Divisions. Other areas of activity are microsystems (facilities include a clean room, the speciality is innovative microsensors), and in some applied physics areas. Altogether the activity is organised in five Divisions: Interactive Sensory Systems (ISS), Cognitive and Communications Technologies (CCT), Automatic Reasoning Systems (ARS), Microsystems (MS), Physics-Chemistry of Surfaces and Interfaces (PCS) and a Tele-medicine Laboratory (TeleMed). CCT is directly involved in the present proposal. Altogether about 15 scientists and some 15 junior researchers are involved in the division.

Research activities of CCT include Natural language-based dialogue, automatic generation of texts and spoken utterances, information extraction from texts, development and maintenance of linguistic resources, question/answering, multimedia and multimodality. ITC-IRST is a member of the European Network of Excellence in Natural Language and Speech (ELSNET). ITC-IRST has been active in many successful EU funded projects, in some case with the role of coordinating partner. Among the relevant ones: FACILE (LE), GIST (LRE); HIPS (Esprit), SPEECHDATCAR (LE); SPEEDATA (LE); TAMIC (MLAP), TAMIC-P (LE), TRANSTERM (LRE), VODIS2 (LE), CHARADE (Esprit); CARICA (Esprit), NESPOLE!, M-PIRO, RENAISSANCE, CLASS.

**CV Bernardo Magnini:**

Bernardo Magnini is Senior Researcher at ITC-IRST (Istituto per la Ricerca Scientifica e Tecnologica). He graduated at the University of Bologna with a thesis on Philosophy of Language. At ITC-IRST he is involved in the TCC (Cognitive and Communication Technology) division, where he coordinated several projects. His research interests are dialogue systems, human-computer interaction, natural language processing technologies with particular emphasis for lexical semantics and linguistic resources. He participated in several national and international projects,

among which TRANSTERM (Creation, Reuse, Normalisation and Interpretation of Terminologies in Natural LanguageProcessing Systems - LRE Project 062-055); GIST, (Generating InStructional Text, LRE Project 062-09); TAMIC-P (Transparent Access to Multiple Information for the Citizen-Pensions); ILEX (realization of a Lexical Database for Italian); TAL (Trattamento automatico della lingua).

**3) Computer languages and systems, University of the Basque Country (UPV/EHU)**

Participant's address: Manuel Lardizabal pasealekua, 1
E-20018 – Donostia
Spain
URL: http://www.ji.si.ehu.es; http://ixa.si.ehu.es
Director: Armando Bilbao

The Computer Languages and Systems department of the University of the Basque Country has a rich pool of research groups, ranging from database research, intelligent tutoring systems to natural language processing. The NLP research group is currently formed by 13 lectures and a large team of collaborators, PhD students and postgraduate research staff, totaling 25 members. The group is a Reference Group in the Language Industry Cluster in the Basque Country.

The group has a tight interaction with the industry both local and international, including local publishers, internet service providers and newspapers. It has produced commercial products like the Xuxen orthographic corrector for Office2000™ and QuarkXpress™, the Elhuyar bilingual dictionary integrated in Office2000™, the Jalgi (www.jalgi.com) web browser and the Egunkaria (www.egunkaria.com) news browser. Our linguistic technology has been used to produce the Basque reference corpus, released by the Royal Academy of Basque. The group has been involved in several Basque local research projects, in two national projects (ITEM, HERMES), in the developer group of EuroWordNet (producing the Basque WordNet) and in one European FEDER project (Hiztegia 2002). During the 1994-2000 the group had a total budget of 0.8 MEuros, excluding lecturers salaries and office expenses.

**CV: Eneko Agirre**

PhD and B.A: in Computer Science by the University of the Basque Country, M.Sc. in Cognitive Science by the University of Edinburgh. Lecturer in the Computer Science Faculty of the University of the Basque Country. He has published over forty refereed articles and conference papers in Natural Language Processing, mainly in the areas of Lexical Knowledge Acquisition and Word Sense Disambiguation. He is a member of the programme committee for the TSD Conference from 2001, and in the International WordNet Conference 2001. He is a reviewer for several major conferences including IJCAI and ACL. He is the site coordinator for the HERMES Spanish national project, and has been involved in the Item Spanish national project and other local projects. He has participated in the last two editions of the SENSEVAL competition, also being the Basque task organizer

**4) Cognitive and Computing Sciences, University of Sussex (UoS)**

Participant's address: Falmer
Brighton BN1 9QH
UK

URL: http://www.cogs.susx.ac.uk
Director: Richard Coates

The group in Cognitive and Computing Sciences, University of Sussex specialising in natural language processing consists of around 15 faculty, doctoral and postdoctoral researchers, five of whom are permanent members of staff. The group is one of the largest in the UK of researchers focusing on statistical and corpus-based approaches to automatic analysis of text.

Recent and current research projects, funded by the EU and by UK national research councils, include the development of shallow parsing technology for English together with corpus-based lexical acquisition techniques, basic research into statistical parsing, automatic simplification of text, efficient wide-coverage parsing using lexicalised grammars, design and implementation of multilingual inheritance-based lexicons, robust parsing by stochastic optimisation, and the construction of large treebanks of written and transcribed spoken English. The value of projects running within the past three years totals some 1.3 MEuro.

The group is active within the international research arena; several members are past and present editorial board members and programme chairs of major book series, journals and international conferences and summer schools; the group also contains the current Secretary of the European Chapter of the Association for Computational Linguistics (EACL), and a member of the executive board of the European Network of Excellence in Human Language Technologies (ELSNET).

**CV: John Carroll**

Reader in Computer Science and Artificial Intelligence at the University of Sussex (UoS), with B.A. and Ph.D. from University of Cambridge. He has published over fifty refereed articles and conference papers in the area of Natural Language Processing, mainly on parsing and lexical acquisition. He was programme chair of the 6th International Workshop on Parsing Technologies (IWPT'00), and has organised international workshops on Robust Parsing (at ESSLLI'96), Evaluation of Parsing Systems (at LREC'98) and Efficiency in Large-scale Parsing Systems (at COLING'00). He is currently on the editorial boards of both the major journals in the field of NLP: Computational Linguistics and Natural Language Engineering. He has been involved, as researcher or principal investigator in European research projects (ACQUILEX, ACQUILEX II, SPARKLE) and several UK national projects. He has participated in both international SENSEVAL competitions. He has recently been invited researcher/associate professor at Stanford University and Tokyo Institute of Technology.

**5) Reuters Limited (Reuters)**

Participant's address: 85 Fleet Street
London EC4P 4AJ
UK
URL: http://www.reuters.com

Today, Reuters technology enables is 663,200-strong client base to access information and real-time quotes on over 960,000 financial instruments including equities, bonds and derivatives from 257 exchanges and over the counter markets. Market prices, news and data are updated as much as 8,000 times a second and distributed instantaneously. The company provides historical information on over 40,000 companies. In addition, more than 30,000 headlines, including third party

contributions, and over eight million words are published daily in over 26 languages. Reuters provides technology which helps its financial customers serve their clients better and to achieve greater efficiencies and cost-savings in the processing of information within their organisations.

Reuters is committed to making the financial markets work on the Internet. It aims to deliver existing financial services, such as analytics, real-time quotes and over the counter pricing, and enhanced services, such as Instant Messaging, over Internet and mobile devices, for instance desk- and laptops, mobile phones and handheld devices. With some 150 years experience in applying electronic information to business needs, it is one of the companies best placed to achieve this.

**Key Person: Richard Willis** (Head of Research and Standards Group).

### 6) Irion Technologies B.V. (Irion)

| | |
|---|---|
| Participant's address: | Bagijnhof 80<br>P.O.Box 2849<br>2601 CV Delft,<br>The Netherlands |
| URL: | http://www.irion.nl |
| CEO: | Joop van Gent |

Irion Technologies is a small software company that develops language-based information retrieval systems for Internet companies and for software houses. Irion combines advanced language technology with the best statistic and heuristic approaches to information retrieval. Current Irion Products are: language-identification, multilingual semantic network, summarization, classification, morphological analysers, shallow parsers, named-entity recognition, cross-lingual retrieval and search engines. Currently, all language technology is available for and across 7 European languages. In the next years, Irion will expand this coverage to many more languages.

Irion was started a year ago and spans a team of 8 developers and computational linguists. Irion is supported by a group of experts in IR and language technology, from TNO, Van Dale and a number of universities. Funded by TNO, Van Dale and three established venture capital companies, Irion thus provides the sound basis of established international companies.

Irion is interested in language technology that brings information retrieval closer to understanding and meaning. Multilingual semantic networks play an important role in this. However, word-sense-disambiguation (WSD) is essential for making language-technology work. Without WSD, a multilingual and mono-lingual semantic network is only of limited use. At the moment, first generation WSD technology is incorporated into the Irion products. Irion plans to enhance this technology.

**CV: Piek Vossen**

Chief Technolopgy Officier at Irion Technologies. He has a B.A. and a Ph.D. at the University of Amsterdam. He has published many articles in the field of computational lexicology and Natural Language Processing. He worked as a senior researcher at the University of Amsterdam in several EC projects (Acquilex I and II, Sift, EAGLES, and EuroWordNet I and II) and national projects (Links, Like). He was the site manager for Amsterdam in the Sift project and the coordinator of

EuroWordNet, in which the University of Amsterdam was the main contractor. He organised several workshops on wordnets (EACL/ACL in Madrid 1997, EACL/ACL-Senseval in Toulouse 2001) and 2 conferences (Euralex 1998 conference in Amsterdam and the 1st WordNet conference 2002 in Mysore, India).

He has been an active member of the Ansi-committee on Ontology Standards and the Eagles Lexicon Group, both involved in the standardization of ontologies, wordnets and lexical semantic resources. In 1999, he joined Sail-Labs as a senior researcher and manager, where he worked for almost 2 years on the development of customization techniques and tools for exploiting multilingual wordnets in information retrieval and classification. He is a board member of the European Association of Computational Linguistics (EACL) and a founder and president of the Global Wordnet Association (GWA).