Progress Report (Months 13–24)

Document Number	Deliverable D0.2 (months 13-24)
Project ref.	IST-2001-34460
Project Acronym	MEANING
Project full title	Developing Multilingual Web-scale Language Technologies
Project URL	http://www.lsi.upc.es/~nlp/meaning/meaning.html
Availability	Project Internal
Authors: German Rigau (UPV/EHU)	

meaning

INFORMATION SOCIETY TECHNOLOGIES

Project ref.	IST-2001-34460
Project Acronym	MEANING
Project full title	Developing Multilingual Web-scale Language Technologies
Security (Distribution level)	Project Internal
Contractual date of delivery	May 2004
Actual date of delivery	June 21, 2004
Document Number	Deliverable D0.2 (months 13-24)
Type	Report
Status & version	v Final
Number of pages	20
WP contributing to the deliberable	WP0
WPTask responsible	German Rigau
Authors	German Rigau (UPV/EHU)
Other contributors	
Reviewer	
EC Project Officer	Evangelia Markidou
Authors: German Rigau (UPV/EHU)	
Keywords:	
Abstract:	

Contents

1 Summary

Natural Language could be the most common and flexible interface between (especially non-expert) users and information systems. However, there is still a long way towards this goal. Current web access applications are based on words; MEANING [Rigau *et al.*, 2002] will investigate new ways for accessing to the Multilingual Web based on concepts, in order to provide to the web applications with capabilities that significantly exceed those currently available. MEANING will facilitate development of concept-based multilingual open domain Internet applications (such as Question/Answering, Cross Lingual Information Retrieval, Dialog systems, Summarisation, Text Categorisation, Event Tracking, Information Extraction, Machine Translation, etc.). Furthermore, MEANING will provide also a common conceptual infrastructure to describe Internet documents, thus facilitating knowledge management of web content. This common conceptual infrastructure is a decisive enabling technology for allowing the semantic web.

MEANING completed the analysis phase during the first year. The output of this crucial phase was the User Requirements report (D1.1), the Basic Design of the Architecture and Methodologies document (D2.1) and the Dissemination and Use Plan (D9.2).

During the second year we revised this analysis. The outcomes of this task has been the Basic Design of the Architecture and Methodologies document (D2.2). This report aimed to produce a detailed design of the development phase for the second MEANING round (months 13 to 24). It revised the overall methodology of MEANING including the standard protocols, formats, procedures, and evaluation criteria and the design Multilingual Central Repository database.

MEANING also continued the development phase of the project. This phase is central to the project, and aims to develop all the software tools and resources to produce the final MEANING outcomes. The project finished the second of three consecutive cycles, each one including the improvement of the already existing Linguistic Processors and the acquisition, word sense disambiguation, uploading and porting processes. The consortium has also performed the evaluation and assessment of the first MEANING cycle (D7.2) that has produced the second version of the Multilingual Central Repository (D4.2) and accompanying experiments and software tools for large–scale text processing: Linguistic Processors (D3.2), Acquisition (D5.2) and WSD (D6.2).

During the second year, the consortium also started the User Validation. Separately from the technical evaluation of MEANING performed in WP7, the consortium planned to carry out a user-based evaluation. The evaluation is separated in two different tasks:

- verification of the intermediate results after each production cycles
- demonstration of MEANING by integrating the results into existing web products

The purpose of the verification is to check whether the results satisfy the user–requirements (D1.1) and to provide the project with feedback on the applied methodology. The purpose of the demonstration is to show the feasibility of integrating the project into an existing industrial environment.

During the second year, the consortium has performed the first user–validation of the first MEANING cycle (D8.1).

For developing these reports, the consortium has investigated the short and long term exploitation and usage of the MEANING results in Human Language Technology (HLT) and software applications, the overall architectural design of the whole MEANING process and the concrete measures on how information about MEANING will be diffused within and outside the consortium.

A major achievement of the project has been the second version of the Multilingual Central Repository in which the conceptual information is stored (WP4). A prototype of the MCR can be consulted using the Web EuroWordNet Interface[1].

In addition, the project has finished the second versions of the Linguistic Processors for the five European languages involved in the project (WP3), and the second release of prototypes for the acquisition phase (WP5) and the disambiguation phase (WP6).

In the second year, we also investigated the language technology and the information systems that can directly benefit from the MEANING database and technology. The market prospects of the MEANING resources and tools are enormous.

The current technology of monolingual keyword search is clearly showing its limitation and severely hampered HLT application development. Nevertheless, the need for services and information handling has continued to increase with the growth of available information. The more information is available the more precise and exact people need to be able to find information. Rather than passive search, people want to be able to directly communicate with information systems, to instruct the mining and collection of information and the delivery of results according to their specific wishes. Such services can now only be offered in small domains and for limited data.

MEANING will allow conceptual processing of massive unstructured data. The MEANING database will make it truly possible to detect the appropriate meaning of words even in small contexts. More sophisticated services within the Customer Relationship Management area could be developed with reduced need for time-consuming knowledge acquisition (which is often ad-hoc and static). Information systems could then communicate with customers through real understanding of the meanings of words in phrases input by users and in the textual data that they are accessing. Systems would no longer be restricted to single domains and could also automatically be customized to further specific domains. Furthermore, since the knowledge stored in the MCR will be language independent, it will be possible to use MCR data to develop services and tools for new languages not currently covered in MEANING only providing an aligned wordnet.

[1] `http://nipadio.lsi.upc.es/wei.html` and
`http://nipadio.lsi.upc.es/cgi-bin/wei3/public/wei.consult.perl`

2 Status

2.1 Resources

Project effort for the 12-month reporting period (person-month)											
Participant's short name	**WP0**	**WP1**	**WP2**	**WP3**	**WP4**	**WP5**	**WP6**	**WP7**	**WP8**	**WP9**	**TOTAL**
UPC	4	1.4	3.3	7	7.4	7.4	7.2	2	1.3	1.8	42.8
ITC-irst	0	0	0	12	5	6	9	0	4.4	0	36.4
UPV/EHU	4.8	0	4	3.83	4.33	9.83	11.17	1	0	0.8	39.77
SuS	0	0	1	2	6	7	6	2	0	1	25
Irion	0	0	0	0	0	0	0	0	5.9	0	5.9
TOTAL	8.8	1.4	8.3	24.83	22.73	30.23	33.37	5	11.6	3.6	149.87

Status of deliverables				
Deliverable	**Current status (completed/ underway)**	**On schedule (yes/no)**	**Original completion date**	**Actual/planned completion date**
D0.2	Completed	Yes	M24	M24
D0.3	Completed	Yes	M24	M24
D2.2	Completed	Yes	M18	M18
D3.2	Completed	Yes	M21	M21
D4.2	Completed	Yes	M24	M24
D5.2	Completed	Yes	M21	M21
D6.2	Completed	Yes	M21	M21
D7.2	Completed	Yes	M24	M24
D8.1	Completed	Yes	M18	M18

Project effort in person-months			
Participant's short name	**Old Total, carried over from last Progress Report**	**This period's Total, carried over from 2.1. Resources**	**New TOTAL**
UPC	42.2	42.8	85
ITC-irst	31	36.4	67.4
UPV/EHU	21.1	39.77	60.87
SuS	26	25	51
Irion	17	5.9	22.9
TOTAL	137.3	149.87	287.17

3 Achievements

3.1 D0.1 Consortium agreement

The consortium decided to postpone this agreement until finishing the subcontrating of Irion and EFE companies. However, the consortium already have agreed a preliminary version.

3.2 D2.2 Basic Design of the architecture and Methodologies (second round)

This document is central to the project, and aims to produce a detailed design of the development phase (months 13 to 24). It includes the overall methodology of MEANING including the standard protocols, formats, procedures, and evaluation criteria and the design Multilingual Central Repository (MCR) database.

This document is the result of a planned revision of deliverable D2.1 (Basic Design of architecture and methodologies [2]). This report aims to produce a detailed design of the development phase for the second MEANING round (months 13 to 24). It revises the overall methodology of MEANING including the standard protocols, formats, procedures, and evaluation criteria and the design Multilingual Central Repository (MCR) database.

The consortium revises in this report the basic requirements for all modules involved in the project. MEANING is a long and very complex project with multiple interdependencies between workpackages. This deliverable tries to identify the current requirements needed for developping succsessfully the whole project. For each task all requirements must be revised, e.g. requirements for the Language Processors and infrastructure (WP3). This deliverable describes also for the second round the information flow and content data for uploading and porting information to the Multilingual Central Repository (MCR) (WP4), the experiments for the acquisition process (WP5) and word sense disambiguation (WP6), and a revision of the evaluation criteria needed for measuring the quality of the tools and resources produced by MEANING (WP7). In particular, this report provides the main guidelines:

- To identify the current requirements for the Language Processors and infrastructure (WP3) to be used in second round of MEANING.
- To define the timing, information flow and content data of the acquisition (WP5), word sense disambiguation (WP6), uploading and porting cycles (WP4) for the second round.
- To revise the main functionality of the MCR (WP4) including:
  - The content to be represented into the MCR (WP4).

[2] `http://www.lsi.upc.es/ nlp/meaning/documentation/D2.1.pdf.gz`

  - The process for uploading the data acquired from one wordnet to the MCR.
  - The process for porting the knowledge stored in the MCR to the respective wordnets.

- To define a set of large-scale knowledge acquisition experiments (WP5) for the second round.
- To define a set of large-scale word sense disambiguation experiments (WP6) for the second round.
- To revise the assessment and evaluation criteria to be used (WP7) for this round.

After a major revision of the whole architecture of the project no major changes were planned for the second round of MEANING. However, some minor changes in the Project Planning were devised.

3.3 D3.2 Second Release of the LPs

This deliverable reports an analysis of the situation of each partner of the MEANING project with respect the availability of Linguistic Processors, Linguistic Resources and Corpora. Moreover, it aims at making a plan for the improvement of currently available resources and the development of futher resources useful in the acquisition (WP5) and disambiguation (WP6) phases of the project.

The MEANING approach heavily relies on linguistic information automatically acquired form large-size corpora for five different languages (Basque, Catalan, English, Italian and Spanish). Such corpora need to be processed at several levels of linguistic analysis, including morphology, syntax, and semantics. As a consecuence, it is of the utmost importance that partners of the project are adecuately equipped with state of the art processors for their respective languages and for the kind of linguistic data they want to acquire and use in word sense disambiguation.

This deliverable provides also a summarized description of the construction of the MultiSemCor corpus (Working Paper WP3.4) and the MEANING Italian corpus (Working Paper WP3.5).

Up to now, all MEANING linguistic processors present high level capabilities and coverage. However, one of the major goals of WP3 is to harmonize current and future developments. Harmonization is necessary at different levels. At one level, harmonization will ensure the availability and the progressive improvement of all the tools and resources required in the different phases of the project. Another level of harmonization will ensure common behaviours of the systems developed at the different partners sites. To this extent, MEANING has choosen standard formats and datasets whenever possible.

3.4 D4.2 PORT1

This document describes the second version of the Multilingual Central Repository (MCR1) and the first Porting process (PORT1). We provide a brief description of the knowledge currently uploaded and integrated into MCR1, including a brief description of a general Upload/Porting architecture. Finally, we provide a full description of the second Porting process.

The MCR follows the model proposed by the EuroWordNet project. EuroWordNet [Vossen, 1998] is a multilingual lexical database with wordnets for several European languages, which are structured as the Princeton WordNet [Fellbaum, 1998].

MEANING works with five wordnets corresponding to five European languages (Basque, Catalan, English, Italian and Spanish). The MCR acts as the sense inventory for nouns, verbs, adjectives and adverbs for all the languages involved in the project. 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 connected.

Working Paper WP4.4 provides an extended summary of the second upload process in the Multilingual Central Repository. The second version of MCR, MCR1 integrates five local wordnets (including five versions of the English Princeton WordNet from WN1.5 to WN2.0), the eXtended WordNet, an upgraded version of the EuroWordNet Top Ontology, WordNet Domains, the Suggested Upper Merged Ontology, and three large set collections of Selectional Preferences and Instances.

Working Paper 4.4 also provides a revised design of the Multilingual Central Repository (MCR). It contains a detailed description of the main software components of the MCR (database design, the Web Interface to the MCR, APIs, XML format for importing/exporting data, etc.) and a detailed summary of the content uploaded for MCR0 (local wordnets, EuroWordNet Top Ontology, MultiWordNet Domains, and large set collections of Selectional Preferences, etc).

All this knowledge acquired locally will be uploaded and ported across the rest of languages via the EuroWordNet ILI, maintaining the compatibility among them. MEANING will perform three consecutive processes for uploading and porting the knowledge acquired from each language to the respective local wordnets: PORT0, PORT1, PORT2. The knowledge acquired from each language during the three cycles will be consistently upload into the MCR, granting the integrity of all the data produced by the project. After each MEANING cycle, all knowledge acquired and integrated into the MCR will be then distributed across the local wordnets.

In that way, the ILI structure (including the Top Ontology and the Domain Hierarchy) will act as a natural backbone to transfer the different knowledge acquired from each local wordnet to the rest of wordnets.

Thus, the MCR includes system modules for:

- Uploading the data acquired from one language to the MCR.
- Porting the knowledge stored into the MCR to the local wordnets.

- Checking the integrity of the data stored in the MCR.

Working paper WP4.5 describes the initial research steps towards the Top Ontology for the Multilingual Central Repository (MCR) built in the MEANING project. The current version of the MCR integrates five local wordnets plus five versions of Princeton's English WordNet, three ontologies and hundreds of thousands of new semantic relations and properties automatically acquired from corpora. In order to maintain compatibility among all these heterogeneous knowledge resources, it is fundamental to have a robust and advanced ontological support. This report studies the mapping of main Sources of Ontological Meaning onto the wordnets and, in particular, the current work on the MEANING Top Concept Ontology.

MCR0 integrates now into the same EuroWordNet framework (using a new version of Base Concepts, the Top Ontology and the MultiWordNet Domains) five local wordnets (with five English WordNet versions) with hundreds of thousand of new semantic relations, instances and properties fully expanded. All wordnets gained some kind of new knowledge comming from other wordnets by means of the first porting process. In fact, the resulting MCR0 is the largest and richest multilingual lexical–knowledge ever build.

3.5 D5.2 ACQ1

This deliverable summarises the work carried out in the second MEANING cycle concerning the acquisition of lexical information, to feed into improved word sense disambiguation.

The initial design of WP5 in the first MEANING cycle comprised five experiments covering the proposed ACQ0 topics and some of ACQ1. Most of the experiments gave promising results. For the second round, the consortium decided to continue all the experiments, and add new ones to cover the topics originally envisaged for ACQ1 and ACQ2.

The ACQ1 phase of MEANING continued to focus on acquiring lexical information from large text colections (corpora currently available) rather than collecting data from the web. However, some experiments (for instance, Experiment 5.D below) are processing large amounts of data acquired directly from the web.

Below we describe eleven experiments (named A–K) that cover the topics proposed in the MEANING technical annex for ACQ0 and ACQ1 and also start to address topics in ACQ2.

- ACQ0:
  - Subcategorization frequencies (experiment 5.A)
  - Topic signatures (experiment 5.D)
  - Terminology/collocations (experiment 5.B)
  - Domain Information (experiment 5.C)
- ACQ1 (using WSD0)

  - New senses (experiment 5.J)
  - Coarser-grained sense clusters (experiment 5.H)
  - Selectional preferences (experiment 5.G)

- ACQ2 (using WSD1)
  - Specific lexico-semantic relations (experiment 5.I)
  - Thematic-role assignments for nominalizations
  - Diathesis alternations

In summary, the consortium planned to perform the following experiments for ACQ1:

Experiment 5.A Investigating multilingual acquisition for verbal predicates.

Experiment 5.B Detecting new noun-noun and adjective-noun collocations.

Experiment 5.C Acquiring domain information for named entities.

Experiment 5.D Acquiring topic signatures from large corpora and the web.

Experiment 5.E Acquiring examples of words used in particular senses from large text collections.

Experiment 5.F Acquisition of lexical knowledge from MRDs.

Experiment 5.G Acquiring improved selectional preferences.

Experiment 5.H Acquisition of sense clusters.

Experiment 5.I Acquiring multiwords: phrasal verbs.

Experiment 5.J Acquiring new senses.

Experiment 5.K Acquiring sense frequencies.

Experiment 5.A performs two preliminary studies into how the knowledge obtained in a language can help in the acquisition process performed in other languages. These experiments aim also to shed some light on the use of the knowledge acquired using general and domain corpora.

Experiment 5.B explores new ways to detect new noun–noun and adjective–noun collocations from large amounts of English text. The experiments investigated a new collocation technique based on analysing the possible substitutions for synonyms within candidate phrases. The experiments also show that WordNet has low coverage of multi-word expressions.

The final goal of experiment 5.C is to automatically acquire domain information for named entities. Two methodologies are being tested: the first one is an application of *Term*

categorization, a supervised approach which aims to assign a domain label to terms in a large corpora. The second approach is directly derived from the usupervised *domain–based* techniques developed at ITC-irst for WSD. The result of the experiments will be a large repository of domain classified named entities.

Experiment 5.D is devoted to acquisition of Topic Signatures from corpora and the web Topic signatures associate a topical vector to each word sense. The dimensions of this topical vector are the words in the vocabulary and the weights capture which are the words closer to the target word sense. This experiment has three main goals: a) to check whether the current technique for constructing topic signatures is satisfactory, b) to assess the usefulness of topic signatures on MEANING related tasks, particularly, domain information acquisition for word senses, and clustering of word senses, and c) to report the current status on the large–scale acquisition of topic signatures. An automatic program has already collected 1000 snippets from Google for each monosemous noun. The total amount of monosemous nouns in WN1.6 is 90,645. The examples take nearly 10 Gigabytes of data.

Experiment 5.E focuses on acquiring form large text collections examples of words used in particular senses. The goal of this experiment is to test the feasibility of automatically generating arbitrarily large corpora for supervised WSD training. For this purpose a new software tool has been developed. EXRETRIEVER characterises automatically each synset of a word as a query (using mainly, synonyms, hyponyms and the words of the definitions); and then, using these queries to obtain sense examples (sentences) automatically from large text collections.

In experiment 5.F, the consortium plans to acquire semantic relations from a Basque, Spanish and English monolingual dictionaries.

Experiment 5.G is exploring methods to acquire selectional preferences which instead of covering all the noun senses in WordNet, just give a probability distribution over a portion of "prototypical classes", where that portion can be disambiguated and where the disambiguation is performed using a ratio of types in a class, rather than tokens.

There is considerable literature on what makes word senses distinct, but there is no general consensus on which criteria should be followed. From a practical point of view, the need to make two senses distinct will depend on the target application. In experiment 5.H a set of automatic methods to hierarchically cluster the word senses in WordNet are been explored.

Experiment 5.I is investigating new methods to determine if some words correspond to phrasal verbs or just verbs+prepositions.

In experiment 5.J three possibilities have been identified to acquire new senses:

1. Classifying new terms using Topic Signatures [Alfonseca and Manandhar, 2002]
2. Classifying new terms using web directories [Santamaria *et al.*, 2003]
3. Training Machine Learning models for semantic classes rather than word classes (see experiment WP6.J Semantic Class Classifiers).

A complete new design of this experiment is expected for the final round.

The first sense heuristic which is often used as a baseline for supervised WSD systems frequently outperforms WSD systems which take context into account. This is largely because of the skewed frequency distribution of the senses for most words. Experiment 5.K is developing a method for automatically ranking WordNet senses. We experiment with nouns only at this stage. We rank noun senses using the "nearest neighbours" in a thesaurus acquired from automatically parsed text.

3.6 D6.2 WSD1

The initial design of WP6 included for the first MEANING cycle six experiments covering most of the WSD0 topics and some for WSD1. All of them are undergoing. For this second round, the consortium decided to continue all the experiments, adding new experiments for covering topics of WSD1 and WSD2.

This working package has two main objectives. On the one hand we will develop state-of-the-art all-words WSD systems for all languages that will act as baselines. The systems will be based on currently existing tagged corpora for English and on unsupervised knowledge-based methods for languages other that English.

On the other hand, several experiments have been designed to explore how to improve current WSD technology. Some experiments try to explore algorithms and more informed features in order to improve the accuracy of supervised WSD systems. Other experiments seek to break the acquisition bottleneck, using a combination of automatically acquired examples, or suplementing labelled data with large amounts of unlabelled data via bootstrapping techniques or transductive algorithms.

Below we summarize the current status of a set of twelve experiments (A – L), and their relation to the working papers and papers related to this task.

Experiment 6.A All-words WSD systems for English

Experiment 6.B High Precision English WSD for bootstrapping

Experiment 6.C WSD based on automatic acquisition of high quality sense examples from large text collections.

Experiment 6.D Transductive approach using Support Vector Machines on labelled and unlabelled data.

Experiment 6.E All-words WSD systems for the rest of languages

Experiment 6.F Contribution of linguistically more informed features in supervised WSD

Experiment 6.G Unsupervised WSD

Experiment 6.H Boostrapping

Experiment 6.I Effect of Sense Clusters

Experiment 6.J Semantic Class Classifiers

Experiment 6.K Effect of Ranking Senses Automatically

Experiment 6.L Disambiguating WN Glosses

Experiment 6.A plans to evaluate different systems to produce a baseline state-of-the-art all-words system for English. This system will be evaluated in the SENSEVAL-3 English all–words task. Experiment 6.E evaluates the portability of those systems to the other languages in MEANING.

As experiments 6.B, 6.C and 6.D, explored different ways to bootstrap supervised WSD into unsupervised or minimally supervised systems, they have been integrated into a common Experiment 6.H (Boostrapping). Experiment 6.B aims to produce high precision systems that could feed supervised systems. Experiment 6.C tries to acquire automatically examples for word senses and train supervised systems with them. Experiment 6.D adds unlabelled data to existing training data to test whether the performance improves.

Experiment 6.F examined features as used by present supervised WSD systems. It will specially analyze the contribution of syntactic and semantic features.

After examining the results of experiment 6.A, the baseline system for WSD1 in English have been deployed in this round. Experiment 6.E provides results for languages other than English. Experiments 6.B, 6.C and 6.D provides preliminary results on breaking the acquisition bottleneck, and experiment 6.F provides clues about the contribution of additional features. Experiment 6.G is devoted to study different unsupervised WSD. Experiments 6.I and 6.K test in the WSD scenario the results of experiment 5.H (Sense Clusters) and 5.K (Ranking Senses Automatically). Instead of learning word experts from SemCor (classifiers that learn to distinguish word senses), experiment 6.J plan to learn multiple semantic class experts (classifiers that learn to distinguish semantic classes). Finally, experiment 6.L uses a set knowledge–driven heuristics to sense disambiguate the WordNet glosses.

The last cycle of WSD will produce all-words systems which improve the baseline system deployed in WSD1. The experiments 6.B to 6.K will provide clues on the best way to acomplish this.

Several research lines have been tried with mixed results for the all-words systems. A prototype all-words WSD system for English has been constructed on SVM trained on Semcor. The system would fare among the 2 best system in the Senseval-2 all-words competition. The all-words WSD system for other languages are under development, and we expect to have the system running for the next Senseval competition the spring of 2003.

3.7 D7.2 Evaluation and assessment of Meaning1

The objectives of work package 7 are:

- to evaluate the quality and accuracy of the developed software and the acquired data; and

- to assess the progress of the project, and if necessary, provide the necessary information to devise corrective actions.

Evaluations of quality and accuracy are given in the relevant work packages (e.g. D6.2 describes word sense disambiguation systems and also presents relevant quantitative evaluation results).

In this document we therefore give a qualitative assessment of the progress made in the second MEANING cycle, and make proposals for work that could be carried out in the final cycle. These proposals will be refined and made more concrete in Deliverable D2.3 (design of the third MEANING cycle).

One of the main points made in this report and also on D3.2, is that the project needs to harmonise the capabilities of its tools: for instance, implementing named entity recognition and text classification in each of the various languages. The output should conform to a common set of categories, e.g. MUC for named entites, and IPTC for text categorisation.

We performed a very complex process to upload correctly all this knowledge into a single multilingual repository. Once finished the first part of the upload process the data released by the different partners (just checking errors and inconsistencies), a more complex second part has been performed. The correct integration of every piece of information into the MCR. That is, linking correctly all this knowledge to the ILI. This second part involved a complex cross checking validation process and usualy a complex expansion/inference of large amounts of semantic properties and relations through the semantic structure.

Having all this types of different knowledge and properties completely expanded and covering the whole MCR1, a new set of inference mechanism can be devised in order to further infer new relations and knowledge. For instance, new relations can be generated when detecting particular *semantic patterns* occurring for some synsets having certain ontological properties, for a particular Domains, etc. That is, new relations can be generated when combining different methods and knowledge. For instance, when several relations derived in the integration process have particular confidence scores greater than certain thresholds. We also suggest further analysis of this possibility.

However, without these new inference tools (i.e. without having infered extra knowledge) in this porting process all the knowledge integrated into the MCR can be ported (distributed) to the local wordnets. That is, this process finish producing exporting XML files for all local wordnets. The results of the full porting shows the feasibility and fertility of the MEANING approach (after uploading and after porting).

For SensEval-3, MEANING proposed a common framework for a lexical sample task in order to evaluate supervised learning systems for WSD in for Basque, Catalan, English, Italian, Spanish and Romanian (this language not included in MEANING). Trail materials were distributed on January 2004. Training materials include a small set of labelled examples and a large set of unlabelled examples.

The MEANING consortium plans to participate with a common system in a unique task: English all-words. Several MEANING partners also plan to participate individually in several other tasks, including several lexical-sample tasks, Word-Sense Disambiguation of WordNet Glosses or automatic subcategorization acquisition, etc.

In the next MEANING cycle, we will research how the current content of MCR1 can help LP2, ACQ2 and WSD2. Specifically, how MCR1 can be used in the other MEANING work packages.

3.8 D8.1 User Validation

The major objective of MEANING is to provide innovative technology for building and deriving enriched lexical resources in which the meaning of words is related to contexts. In these contexts words are used in a particular meaning. A large collection of actual uses of word meanings will enable the development of high-precision technology that directly and precisely relates language to the actual message or meaning, rather than just the pattern of words. Eventually, conceptual language technology systems can be developed that actually interpret the information expressed in natural language, make the adequate inferences and generate the needed responses back in natural language. However, there is also a direct short-term benefit from such a database for current language technologies such as information retrieval, classification and naive dialogue systems. If words can precisely be mapped to meaning, the current systems will be more precise and effective, information can be compressed and systems will be faster.

Irion Technologies, as a language-technology company, is interested in any technique that deals with the many ways in which languages encode information in expressions. Since there is no one-to-one mapping between expressions and information and since the number of expressions is in principle infinite (and in practice very large), it is absolutely necessary that methodologies will be available to determine the correct and/or relevant mappings in context. Irion has a multilingual semantic network that is comparable to MCR. The core focus of the company is on building generic cross-lingual technology and end-user applications that exploit this database. Once the Irion database is connected to the MEANING database, the results of MEANING can directly be tested.

During the second year, the consortium has performed the first user–validation of the first MEANING cycle (D8.1). The purpose of the verification is to check whether the results satisfy the user–requirements (D1.1) and to provide the project with feedback on the applied methodology. The purpose of the demonstration is to show the feasibility of integrating the project into an existing industrial environment.

In general, there are two major user-perspectives towards resources such as the MEANING database. First of all, the database will make certain existing NLP technologies more precise and indirectly improve end-user applications that can make use of these technologies. Secondly, it will be possible to develop new technologies and applications, which are currently too complex due to the ambiguity and vagueness of linguistic expressions.

Within the limitations of the project, we narrowed the scope of the demonstration and testing environment to 3 major areas:

- Cross–lingual information retrieval
- Cross–lingual classification

- Dialogue systems on unstructured text

These applications are been extended with the MEANING knowledge database to illustrate the capacity of conceptual information processing. They are currently most directly related to the naive keyword indexing and search and therefore can show the capacity to improve this technology. The dialogue systems clearly show the capacity of MEANING to contribute to high-precision information accessing.

To perform the validation tests, Irion has designed a complete framework and dataset to test the Irion systems. This framework uses the Reuters collection containing 23307 English news for the period 20-August-1996 until 19-August-1997. The Irion framework will allow the MEANING consortium:

- To check in a common test bed the current performance of the Irion Systems with and without the MEANING technology.
- To demonstrate the feasibility of integrating the MEANING results in three different Language Technology products from Irion:
  - TwentyOne: Cross-Lingual Information Retrieval system
  - Adjust: Cross-Lingual Classification System
  - Pidgin: Cross-Lingual Question Answering Dialog System

The content of deliverable D8.1 is basically the following:

- Description of the Irion systems, including all their components, etc.
- Description of an evaluation scenario and formal criteria.
- Performance of the Irion systems as they are now that is, a formal evaluation of the current Irion systems withoutÂ?MEANING.
- Assessment of the MEANING1 outcomes, that is the Irion vision of the MCR and software
- Plan of integration, that is, how Irion plans to integrate the initial MEANING outcomes into the Irion systems. For instance, Irion may consider that not all the knowledge and software tools available are worthy to include into a prototype system for evaluation.
- Conclusions: the feedback Irion to the MEANING consortium.

Next deliverables (D8.2 and D8.3) should integrate also the complete design of an evaluation framework for an end–user scenario. The evaluation will be performed in a real scenario provided by EFE. During the meeting in Madrid we discussed several possible scenarios. In particular, we decided to investigate a multilingual database of pictures:

FOTOTECA. This database receives about 800 pictures everyday. Each picture have a caption mainly in Spanish and English. Now, these captions are translated for multilingual access.

EFE has provided a small sample of two month of text captions.

Some key points about the initial EFE scenario:

- They receive around 800 pictures everyday.
- There are Spanish (from EFE) and English texts (from EPA and AP).
- EFE is translating most of the English texts.
- 500 words per text on average.
- Users usually ask for Named Entities: Persons locations and Events.
- The text is in XML format.

4 Management

No major deviations have been detected from the current workplan. However, after the experience of the first MEANING cycle, we planned to extend the second cycle three months. In that way, each MEANING cycle will cover a complete whole year of analisys, development, evaluation and assessent.

Although it was initially decided to be subcontracted by UPV/EHU, finally, the University of Sussex has successfully subcontracted Irion technologies for a total amount of 70,000 Euro to contribute to Work Packages 1 (user requirement), 8 (user validation) and 9 (exploitation and dissemination).

Finally, Reuters will not be subcontracted by the consortium. After a teleconference with Reuters at Brighton, they provided to the consortium a *letter of intend* including their expected contribution in MEANING as consultants. However, in the meanwhile (summer 2002) Reuters suffered an in depth internal restructuring (i.e. most of the R+D section of Reuters has moved to California). After no signal from Reuters the consortium agreed to explore the active participation of other similar End-User companies (i.e. EFE, Sharp, VanDale, etc.). These companies will be also invited to become members of the MEANING user-group.

During the last year, the consortium finished the negotiations with the Spanish EFE news agency, in order to substitute the End-user role of Reuters in MEANING. EFE is currently collaborating with UPC and UPV/EHU teams in a national project called HERMES [3]. EFE is being subcontracted by the UPC for a total amount of 30,000 Euro to contribute to Work Package 1 (user requirement), 8 (user validation) and 9 (exploitation and dissemination).

[3] `http://terral.lsi.uned.es/hermes/`

5 Awareness

MEANING is also exploring the possibilities to set up a user-group of trans-national European companies. The consortium is also in contact with the Global WordNet Association and EuroTerm and BalkaNet IST projects that extend and develop wordnets in specific domains and other languages, which could be added to the MEANING MCR or tap information from the MEANING knowledge base to their own language. MEANING is also in contact with DEEP THOUGHT[4] project for exchange of results and concertation to support robust deep NLP. Furthermore, MEANING is closely cooperating with and contributing to SENSEVAL competition.

To evaluate the MEANING results, the consortium collaborated with two other projects: SWAP[5] and EDAMOK[6]. Very brifly, a SWAP partner located in Mayorca is interested in evaluating the EDAMOK technology on distributed knowledge management; such technology, requires, among others, linguistic processors and the MCR developped in MEANING.

The ESPERONTO[7] project also contacted with us to start a collaboration between the two projects. In this case, they also need linguistic processors and wordnets for the languages considered (English, Spanish and Catalan). MEANING could provide the necessary tools and resources for this project.

We are also in close contect with the BALKANET project in order to coordinate current and future developements of both projects.

Regarding dissemination, during the first year, the consortium published 41 papers in 17 International Conference Proceedings, 4 International WorkShops, 3 Journals and one book, covering different MEANING areas: from WP2 to WP8.

During the second year, the consortium published 51 papers in 32 papers in International Conference Proceedings, 13 papers in International WorkShops and 6 papers in international journals. These papers also covers different MEANING working parts: from WP2 to WP9.

We also performed successfully the first MEANING workshop [8]. The workshop was held in San Sebastian the 10th, 11th and 12th of April, 2003. The total number of participants was 40. We invited eight relevant researchers on Acquisition, Word Sense Disambiguation and representation and management of knowledge:

- Walter Daelemans[9]
- Rada Mihalcea[10]
- David Yarowsky[11]

[4] http://www.eurice.de/deepthought/index.htm
[5] http://swap.semanticweb.org
[6] http://edamok.itc.it
[7] http://www.esperonto.net
[8] http://ixa.si.ehu.es/Ixa/local/meaning-workshop
[9] http://cnts.uia.ac.be/~walter/home.html
[10] http://www.cs.unt.edu/~rada/
[11] http://www.cs.jhu.edu/~yarowsky/

- Fernando Gomez[12]
- Dekang Lin[13]
- Alexander Maedche[14]
- Anna Korhonen[15]
- Julio Gonzalo[16]

The workshop was divided into 8 hour-slots: 4 hour for external presentations, 2 hours for internal presentations and 2 hours for panel discussion. We reserved 1 hour-slot for speaker (45 min. exposition, 15 min questions/discussion). Mainly, one day devoted to WSD and the other to acquisition. We also reserved 4 hour-slots for MEANING presentations: 1 hour at the beginning for presenting the whole project, and three more for WP5 (ACQ), WP6 (WSD) and WP4 (PORT).

The consortium was invited to submit an article describing the major achievements of MEANING for the ELSNews.

German Rigau was invited to the Global WordNet Conference GWC'2004[17] in Brno, Czech Republic, the 20th to 23th of January to describe the current status and outcomes of the MEANING project.

With respect the user group, each partner has provided a initial list of possible companies and institutions to become members of the MEANING user-group (i.e. Reuters, Sharp, EFE, Larousse, VanDale, etc). The MEANING web site has a form for subscription as a member in the MEANING user group. A newsletter will be also maintained to keep the MEANING user group informed.

6 Conclusions

The project aims to collect and analyse language data from the web and building multilingual lexical knowledge bases to support open domain word sense disambiguation. Although, this is a very chalanging research area, the vast majority of technical goals mentioned in the work-plan has been correctly achieved. The results of this project are extremely satisfactory, and in several aspects better than expected.

A major achievement of the project has been the second version of the Multilingual Central Repository in which the conceptual information is stored (WP4). A prototype of the MCR can be consulted using the Web EuroWordNet Interface[18].

[12] http://www.cs.ucf.edu/~gomez/
[13] http://www.cs.ualberta.ca/~lindek/
[14] http://www.aifb.uni-karlsruhe.de/Personen/viewPersonenglish?id_db=50
[15] http://www.cl.cam.ac.uk/users/alk23
[16] http://sensei.lsi.uned.es/~julio/
[17] http://www.fi.muni.cz/gwc2004/
[18] http://nipadio.lsi.upc.es/wei.html and
http://nipadio.lsi.upc.es/cgi-bin/wei3/public/wei.consult.perl

The consortium expects to announce the distribution of the second release the Multilingual Central Repository (MCR1) during final year of the project. The distribution will consider several licenses. A free license will contain all Princeton wordnets included into the MCR, all the mapping between all these wordnets, the MEANING Top Concept Ontology, the SUMO ontology and the eXtended WordNet. A free license for research purposes will also include the Basque, Catalan, Italian and Spanish wordets, the WordNet Domains and the selectional preferences acquired from SemCor and BNC.

In addition, the project has finished the second versions of the Linguistic Processors for the five European languages involved in the project (WP3), the second prototypes for the acquisition phase (WP5) and the second releases of the disambiguation systems (WP6).

During the second year, the consortium published 51 papers in 32 papers in International Conference Proceedings, 13 papers in International WorkShops and 6 papers in international journals. These papers also covers different MEANING working parts: from WP2 to WP9.

We also performed successfully the first MEANING workshop [19]. The workshop was held in San Sebastian the 10th, 11th and 12th of April, 2003. The total number of participants was 40 including eight invited relevant researchers on Acquisition, Word Sense Disambiguation, and representation and management of knowledge.

Now, after the second cycle, the consortium is ready to start validating the project outcomes in several Irion concept-based internet applications, including CLIR, Text Classification and Question Answering dialogue systems. The consortium also expects to involve the EFE company in the validation process in a very near future.

Moreover, the consortium has decided to make the results publicly and freely available. This decision will garantee the exploitation of the project results by the HLT community. The consortium expects an important impact on the HLT sector at large, both in its knowledge technology side and on multilinguality.

References

[Alfonseca and Manandhar, 2002] E. Alfonseca and S. Manandhar. An unsupervised method for general named entity recognition and automated concept discovery. In *Proceedings of the 1st International Conference on General WordNet*, Mysore, India, 2002.

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

[Rigau *et al.*, 2002] G. Rigau, B. Magnini, E. Agirre, P. Vossen, and J. Carroll. Meaning: A roadmap to knowledge technologies. In *Proceedings of COLING'2002 Workshop on A Roadmap for Computational Linguistics*, Taipei, Taiwan, 2002.

[19] `http://ixa.si.ehu.es/Ixa/local/meaning-workshop`

[Santamaria *et al.*, 2003] C. Santamaria, J. Gonzalo, and F. Verdejo. Automatic association of web directories with word senses. *Compututational Linguistics, Speccial Issue on web as a corpus*, 29(3):485–502, 2003.

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