REQUIREMENT ANALYSIS & SPECIFICATION OF THE METHODOLOGY

Deliverable D.2.1, WP2, BalkaNet, IST-2000-29388

Databases Laboratory (DBLAB)

Computer Engineering & Informatics Department

Patras University, Greece

GR 26500

Project coordinator: Prof. Dimitris N. Christodoulakis {dxri@cti.gr}
Project Web site: http://www.ceid.upatras.gr/Balkanet

BalkaNet

Identification number	IST-2000-29388
Type	Report – Document
Title	Requirement Analysis and Specification of the Methodology
Status	Final
Deliverable	D.2.1
WP contributing to the deliverable	WP2
Task	T.2.1 & T.2.2
Period Covered	September-December 2001
Date	January 2002
Version	6
Status	Confidential
Number of pages	79
WP / Task Responsible	CTI
Other Contributors	DBLAB, UAIC, RACAI, SABANCI, FI MU, MEMO, MATF, DCMB, PU, UOA

Authors	{Tzagarakis Manolis, Kapatsoulia Natassa} **CTI** {Assimakopoulos Dimitris, Stamou Sofia, Kyriakopoulou Maria, Avramidis Dimitris, Koutsoubos Ioannis, Mistou Eyh, Andrikopoulos Vasilis, Mathiou Stavria, Tsiopela Dimitra, Dimitropoulou Reggina, Spiliopoulou Stamatia-Irini} **DBLAB** {Cvetana Krstev} **MATF** {Oflazer Kemal, Cetinoglu Ozlem, Bilgin Orhan} **SABANCI** Cristea Dan **UAIC** Tufis Dan **UAIC** {Pala Karel, Smrz Pavel, Pavelek Tomas} **FI MU** Dutoit Dominique **MEMODATA** Koeva Svetla **DCMB**
EC Project Officer	Erwin Valentini
Project Coordinator	Professor Dimitris Christodoulakis Director of DBLAB Databases Laboratory, Computer Engineering & Informatics Department Patras University GR 26500, Greece Phone: +30 (61) 960 385 Fax: +30 (61) 960 438 E-mail: dxri@cti.gr
Keywords	Data Requirements, User Requirements, Applications of Semantic Networks, Lexical Resources, Information Retrieval, Language Engineering, WordNet Management System
Actual Distribution	Project consortium, Project Officer, EC

Abstract	In this report the general design of the BalkaNet multilingual semantic network is described based on the users and developers requirements and the state of the art in building semantic resources. During the specification of the methodology to be followed for the implementation of the project the EuroWordNet semantic database was regularly used as a reference point since BalkaNet's results are going to be merged with the EWN resource forming thus a European network of concepts. More specifically, the requirement analysis is based on the application of the project results, which is targeted towards Conceptual Indexing tasks in Information Retrieval applications. The usage of the lexical database in an Information Retrieval environment is taken as a starting point for the functional specification. In addition, to the direct requirements of the end users, the functional specification is based on the design of the Princeton WordNet 1,5 and the EuroWordNet, the structure and content of the resources that are going to be used, the quality of the extraction tools and the limitations set by the project's time frame. Deviations from the EuroWordNet might be due to: ❑ New applications of the project's results i.e. conceptual indexing tasks ❑ Different types and structure of the available lexical resources ❑ Inadequacy of the EuroWordNet in Information Retrieval tasks ❑ Achieving maximal compatibility across the monolingual Balkan WordNets while keeping compatibility with the EuroWordNet as well ❑ Quality of the tools for extracting information from the lexical resources Apart from the abovementioned deviations one great differentiation of the BalkaNet in comparison to the EuroWordNet network is the fact that the Inter-Lingual-Index of the former is going to be structure on the basis of a WordNet-like hierarchical structure. Despite the fact that the function of the ILI remains the same i.e. mapping concepts across language however its structure will be greatly differentiated from the one used in the EWN project. To achieve that that domain labels are going to be included in the monolingual WordNets and the Inter-Lingual-Index.
Status of abstract	Complete
Send on	January 2002

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Language resources are not only important for processing of information that is usually difficult to acquire but also because they impose a lot of challenges to the user who wishes to store, manipulate and retrieve it in an efficient way. Those challenges usually have to do with the amount of data that one has to handle but sometimes difficulties regarding the nature of the language itself and its properties come up.

One such widely used lexical resource is WordNet, a semantically cross-linked dictionary, which resembles the way that humans store and organize information in their memory. WordNet has known a great success as a lexical resource and a linguistic tool, for both language studies and as supporting resource in other large-scaled applications (e.g. Information Retrieval). The radical initial approach of WordNet was followed by EuroWordNet, which provided a similar infrastructure but incorporated more languages. Developers of EuroWordNet came across many interesting aspects of the task of handling and organizing multilingual resources in a semantic network and had to come up with new approaches and solutions.

Despite the abovementioned semantic networks when it comes to East and Central European languages such databases representing semantic relations are available to a limited extend and are by no means incorporated in any NL applications. The aim of the BalkaNet project is to develop a multilingual lexical semantic network representing the different structures of the Balkan languages in such a way that navigation through the conceptual space of monolingual and multilingual resources is facilitated with a minimum effort by the end user. The BalkaNet semantic network will be as much as possible build from available lexical resources since it is not only more cost-effective but it reassures at the same time the maintenance of language-dependent differences in the monolingual networks. The use of the BalkaNet's results will be demonstrated in an Information Retrieval environment. The expectation is that retrieval performance can greatly benefit from such a multilingual semantic network in terms of relevance[1] of the obtained results. More specifically, within the framework of the project an attempt will be made towards indexing documents in IR systems not in terms of their wordfroms (as performed by most search engines) but in terms of their conceptual meaning. Thus, language internal structures and semantic domains of the BalkaNet lexical resource will function as the basis against which documents will be checked prior to their indexing in IR system repositories.

To limit the scope of the work, the multilingual network will be primarily used for a single application that is conceptual indexing for information retrieval tasks. Given the innovation of the specific application it is easily understood that the impact of the results cannot be foreseen at this stage and end user requirements cannot be as complete as desired due to the novelty of the proposed approach and the limited time of the project to perform an extensive market research. However, the proposed application was selected for two main reasons. Firstly, due to the observation that WordNets already tested against IR systems did not significantly improve retrieval results. Their contribution concentrated mostly on the enhancement of recall results whereas at the same time they dropped precision too much. On the other hand, motivated by many studies on how end users interact with IR systems we attempt by

[1] Relevance is used here in terms of precision since the latter is measured using standard TF*IDF scores instead of semantic resources.

conceptual retrieval the acquisition of data relevant to the users' information needs. This will be achieved by classifying documents while being indexed according to the conceptual domain(s) they cover without taking wordforms and morphology into consideration. After all end users are interested in finding the desired information (or at least to approach it) and they pay little attention on lexical phenomena such as polysemy of lexical ambiguity problems that most search engines' developers try to deal with.

In this documents we outline the user requirements in which the specification of the methodology for the implementation of the semantic network will be based. In addition, to the use requirements the type of information that is covered by the EuroWordNet project will be taken into consideration since we wish to keep maximal compatibility and incorporate as much as possible our results to the EWN database contributing thus to a common semantic network for most of the European languages.

However, deviations from the EuroWordNet design might be due to:

- ✓ Different structures of the Balkan languages
- ✓ The nature of the information stored in lexical resources from which the data to be incorporated into the BalkaNet will be derived.
- ✓ The poor quality or lack of electronic lexical resources for some of the participating languages
- ✓ The difficulty in accessing all necessary resources due to copy rights claims
- ✓ The inadequacy of the EuroWordNet for information retrieval use, which follows from its domain application.
- ✓ The possibility for the user to navigate through the desired conceptual domains without knowing the structure of the particular language and in most of the cases without even having to speak all the languages
- ✓ Achieving maximal compatibility across the different monolingual WordNets
- ✓ Trying to ensure that the structure of the semantic network will be independent of the underlying language and will resemble as much as possible the way in which human mental lexicons are organized.
- ✓ The limited exploitation of the participating languages due to the lack of NLP tools and infrastructures for lexical data acquisition.

Finally, our target application concentrates on the less-studied, thus equally important, Balkan languages and the deployment of new tools for their processing in an attempt to promote them in the EU community and thus make data easily accessible.

INTRODUCTION

WordNet is an electronic lexical thesaurus based on word meanings rather than word forms and divides the lexicon into the following categories: nouns, verbs, adjectives and adverbs. WordNet developed at Princeton University (Miller, 1990) holds semantic relations between English words organized around the notion of synsets. The success of WordNet has determined the emergence of several projects that aim at the development of WordNets for languages other than English. A European project, called EuroWordNet, (EWN) (Vossen, 1998) added eight other languages to this thesaurus resulting in a huge network of linguistic concepts that allow inter-lingual navigation and finding of translation equivalences between languages.

BalkaNet is an EC funded project (IST-2000-29388) that aims at developing a multilingual resource representing semantic relations among basic concepts of the following Balkan languages: Greek, Turkish, Romanian, Bulgarian, Czech and Serbian. The BalkaNet includes semantic relations existing in each of the above languages, as language internal relations, as well as among them, as equivalence relations to an Inter-Lingual-Index (ILI). The BalkaNet will as much as possible be built from available lexical resources so that it will be possible to combine information from independently created resources, making the final database more consistent and reliable while keeping at the same time the richness and diversity of the vocabularies of the languages involved. The main resources of information are going to be the individual monolingual WordNets that have already been developed or are currently under development for most of the participant languages. Where a monolingual Wordnet is not available dictionaries, thesauri or corpora of the respective languages will be used for the terminology extraction.

Members of the BalkaNet consortium are: University of Patras (Greece), Computer Technology Institute (Greece), University of Alexandru Ioan Cuza (Romania), Romanian Academy / Centre for Advanced Research in Machine Learning (Romania), Bulgarian Academy of Science (Bulgaria), Sabanci University (Turkey), Masaryk University / Faculty of Informatics (Czech Republic), Memodata (France), University of Plovdiv (Bulgaria), and the University of Athens (Greece). University of Belgrade / Faculty of Mathematics (Yugoslavia), Center Applied Research (the Netherlands) and Otenet / Internet Service Provider (Greece) are going to act as subcontractors.

For the development of the BalkaNet project a merge model approach will be adopted, meaning that each WordNet will be built separately in each language from independently developed resources and then linked to the most equivalent concepts in the ILI record. We aim at a total set of 15.000 comparable synsets in each language, corresponding with more or less 30,000 literals, covering generic vocabulary of the involved languages. The Part-Of-Speech (POS) distribution will be 65% nouns, 25% verbs, 5% adjectives and 5% adverbs. In addition, the monolingual Balkan WordNets developed from scratch within the framework of the project will comprise approximately 8,000 synsets whereas the number of synsets that will be added in already existing ones will be determined at a later stage. In addition, apart from the representation of generic vocabulary into the multilingual database a feasibility study will take place in order to test how domain-specific terminology can be incorporated into the semantic network under domain labels. Finally, the BalkaNet will be incorporated into the EWN semantic network resulting in a global semantic database covering conceptual areas of European languages. Thus, in order to keep

compatibility with the EuroWordNet the Language Independent Module, namely the Top-Concept Ontology, will be maintained along with the ILI records.

Despite the fact that the BalkaNet semantic network will be constructed in a similar way with the EWN, several new features will be implemented the most important of which concerns the development of a WordNet Management System (WMS) that will be publicly available at the end of the project. The main differentiation between the BalkaNet and the EWN relies on the reusability and openness of the tools and software that will be developed for the BalkaNet. More specifically, the EWN has been constructed by using the Polaris (Bloksma, 1996) tool, which has a few drawbacks. First and foremost it is a commercial stand-alone tool designed solely for WordNet maintenance that cannot be easily adapted to a new application and also runs only in Microsoft Windows.

For the implementation of the BalkaNet project we will develop a (inter) networked tool that helps partners in coordinating their work online. Although it would be technically possible, we do not want to create a fully Internet-based Web Polaris tool, since the Web cannot yet deliver a full-blown graphical user interface, and this would unnecessarily restrict local editing of WordNets. However, keeping all the benefits of the Web, such as distributed work environment, concurrent access to the data and multiple views of the data will be achieved through the WMS. Thus, what we intend to develop is a WMS that with a proper protocol will allow the local tools to retrieve the required information. However, since the Internet is not always reliable the offline operation of local tools will be the primary mode of the WMS whereas the online one will be considered as an extra facility. This way, a WMS that supports both online and offline integration with local tools, plus a good dedicated online interface is going to be a powerful federated platform for quick and coordinated development of the monolingual WordNets while at the same time the construction of a multilingual Balkan WordNet will be feasible.

So far, EWN shares the same concept of a multilingual synset in the ILI. New records can only be added at the tail of the file and are maintained by a central authority that issues periodical releases of a new ILI record replacing the previous one. The WMS will provide a more flexible reference scheme that enables local WordNets to keep references to the ILI even while the latter is significantly restructured. The benefit behind using the WMS is that project developers will be tightly linked with other WordNets and valuable suggestions for new terminology fields will be facilitated. The WMS will be as open to the user as possible since it will be fairly easy for the users to develop and add their own components to the system. This can be accomplished by either encapsulating in the system capabilities for “plugging in” other applications, or by deploying the system under a free source license, with or without the data. In this way, it will be easy for users to use the same platform for their work and keep at the same time the data compatible.

Fundamental in the design of the WMS is that each partner will retain full responsibility and independence of his local WordNet and at the same time they will be able to view other WordNets and check how compatible they are. A new browser (editor) developed for the BalkaNet project will be able to work with WordNet files written in XML and it will also employ client-server architecture. The above tools will be developed in Linux platform and the results will be widely available at the end of the project. The central infrastructure of the WMS is going to be a federated database along with necessary communication protocols and Linux-based tools, which

will run locally and provide central services. The main output of the WMS will be a publicly available WordNet editor and a database viewer, which will export both monolingual and multilingual data stored in the central BalkaNet database. Summarizing, the (inter) networked WMS is going to be a platform independent tool that will enable the individual development of the monolingual WordNets while at the same time linking of WordNets into a central database will be feasible.

DRAWBACKS OF WORDNET 1.5 AND HOW EUROWORDNET DIFFERENTIATED?

Drawbacks of the Princeton WordNet 1,5

- ✓ Lack of "formal" and "telic" hyponymic
- ✓ Relations. Noun files in WordNet were developed before realized the importance of this distinction
- ✓ No distinction between proper and common nouns, or between mass and count nouns. Thus some nouns have to be re-classified
- ✓ Lack of information about the semantic distance between two related words. WordNet makes not effort to weight the meaning differences.
- ✓ Wordnet gives information only about the lexical level and its usefulness for knowledge engineering is limited.
- ✓ Lack of compound concepts
- ✓ Small number of entailment relations (entailment is a unilateral relation)
- ✓ Absence of case relations

In order to overcome the above problems the following can be done:

- ✓ Add better distinguished relations, and thematic roles
- ✓ Enrich the morphology
- ✓ Add a domain level, which refers to the topic and thus relates objects to events.

Differences between the EuroWordNet and the Princeton WordNet 1,5

Inter-Lingua-Index: is an unstructured fund of English concepts manly taken from WN 1,5 with the only purpose to provide an efficient mapping across the individual languages. Each synset in the monolingual WordNets has at least one equivalence relation with the synset or record in the ILI. The only organization of the ILI is the linking of ILI records with the top-concept ontology and with the domain labels. All language-independent pieces of information are stored in the ILI concepts. Changes in the domain, the top-concepts and at instances have to be specified only for the ILI records and not for each monolingual WordNet. One partner, who notifies the others so that they update their monolingual WordNets, is performing the update of the ILI. Added ILI records have to be glossed in English. The ILI is a simple list without

internal structure but the translation relations might be complex in the sense that languages may have different types of equivalence relations with the records but no relations exist between the index records. A drawback of the ILI is that the transitivity of the relations cannot be guaranteed.

Top-Concepts: It is a hierarchy of language-independent concepts, reflecting explicit opposition relations. There is consensus in the internal relations of the top-concepts and if formed by the top level of synsets from each language. The top-concepts are linked to the ILI records

Domains: The purpose of the domains is twofold. First it can be used directly in IR tasks to group concepts in a different way and secondly by using domains the generic vocabulary can be separated from the domain-specific vocabularies, overcoming thus ambiguity problems. Domains are language-neutral and they are stored at the ILI concepts

Note: Both top-concepts and domain labels can be transferred via the equivalence relations of the ILI to the language specific meanings and next via the language internal relations to any other meaning in the WordNets.

Instances: These are mostly proper nouns and unlike domains are linked in the individual WordNets to their classes. This is due to the fact that they are not easily derived from the lexical resources. Instances are minimally added in the EWN by hand just to illustrate the possibility of future extension and customization.

Differences in relations between WN 1,5 and EWN

- ✓ The use of **labels in relations** (e.g. conjunction/disjunction, non-factive[2], reversed, negation)
- ✓ **XPOS relations** (synonymy and hyponymy)
- ✓ A more global **near-synonym relation** (the distinction is relevant for IR because it makes it possible to precisely predict which words can be expected to replace other words in text)
- ✓ **Sub-event relations** instead of entailment (it is useful for closely related verbs and the difference in the direction of the entailment can be expressed by the labels factive and reversed).
- ✓ Use of **role relations** between entities and events (used when hyponymy does not express the relation in a salient way).

Why XPOS relations were allowed in EWN?

The main reason concentrates on the different syntactic role of nouns and verbs in English since approximately 30% of all noun senses refer to an event, state of relation. Many of these could have synonymy or hyponymy relations with verbs or adjectives or should at least be related to the top-concepts of the verbs.

[2] Non-factive indicates that a causal relation does not necessarily hold

- ✓ In other languages than English it is difficult to distinguish nouns and verbs or the distinction does not even exist [Lyons 1977]

- ✓ From an IR point of view the same information can be coded in an NP or in a sentence, thus by unifying nouns and verbs in the same ontology is will be possible to match expressions with different syntactic structures but comparable content.

- ✓ By merging verbs and abstract nouns it is possible to link mismatches across languages that involve a POS-shift.

Thus, instead of distinguishing among nouns and verbs in the EWN there is a distinction between first-order (nouns) and high-order (both nouns and verbs) entities. As a consequence new relations were introduced in EWN that did not exist in WN 1.5 such as "noun-to-verb-hypernym", "verb-to-noun-hyponym", "noun-to-verb-synonym" and "verb-to-noun-synonym".

PART A

USER REQUIREMENT - FUNCTIONAL REQUIREMENT ANALYSIS

The BalkaNet project aims at building a multilingual lexical database consisting of WordNets in several Central and Eastern European languages. Each language specific WordNet will be structured along the same lines as the Princeton WordNet and the EuroWordNet, i.e. synonyms are grouped in synsets (synonym sets), which in their turn are related by means of basic semantic relations such as hyponymy, meronymy, antonymy etc. All language specific WordNets will be stored in a central lexical database system. Equivalence relations between synsets in different languages will be made explicit in the so-called Inter-Lingual-Index, which will be adopted from the EuroWordNet project. The Index is an unstructured collection of concepts with the only purpose to provide an efficient mapping across the different languages. Where necessary the Inter-Lingual-Index will be modified or further extended with new concepts in order to reflect the lexicalization patterns of the Balkan languages.

For the successful implementation of the project requirements set by two distinct groups, namely the end users and the developers have to be specified so that there is a coherence in the decisions made within the framework of the project and the latter forms a useful application. Thus, each group has to set its requirements separately based on their expectations and objectives, which to some extend might share some common points. More specifically, the requirements set by the developers concentrate mostly on the development of a multilingual semantic database comprising of WordNets in different Balkan languages. In this respect particular attention has to be paid to the data of the semantic resources and as such to the lexical resources (both their quality and quantity) that will form the basis for the actual development of the semantic database. Moreover, attention should be paid and special care should be given to the architecture and implementation of the tools for the processing of the resources and for the integration of the data in each monolingual network. Finally, the requirements of the developers are also targeted against the incorporation of each monolingual network in the final database and the interlinking of the individual WordNets via the Inter-Lingual-Index.

However, in order to meet the abovementioned requirements, developers have to keep in mind during the implementation of the project the final application of BalkaNet and what send users are expecting out of it. Consequently, user requirements have to be clearly determined since they are going to form the basis on which the data and structure of the network will be implemented. In light of the above and concentrating on the fact that the project's results are targeted towards Information Retrieval tasks the consortium has performed a limited (due to time considerations) market search in order to trace areas that could benefit from such a multilingual database. At this point it should be noted that despite the fact that BalkaNet is mainly targeted towards Balkan languages we wished that our application and thus the infrastructure of our network is language-independent and if possible of application-independent. At this early stage of work we mostly focused on a language-independent application.

Information Retrieval

The amount and the detail of the information stored in Information Retrieval (IR) systems grow exponentially and end users of such systems wish to easily and quickly

retrieve the desired information (regardless of format or medium) with the minimum effort. The performance of information retrieval systems is measured with traditional Recall and Precision Scores, which most of the times are provided by the systems and their integrated mechanisms (e.g. index, ranking algorithms etc.). Despite the fact that many such mechanisms are being continuously developed or enhanced, the amount of data and users is growing too fast and as a result there are many cases reported where end users are either unable to efficiently use an IR system or unable to find the information they are looking for. Of course the abovementioned problems are partly due to the lack of experience by the end users. In many cases though it is the system to blame for such inconveniencies since they do not always provide help or advanced facilities to inexperienced end users.

In order to enhance system performance and make it interact in a friendly way with the end users various techniques have been incorporated in them in an attempt either to help users interpret retrieval results or to make the system interpret users' needs. Towards the last direction many Natural Language Processing (NLP) techniques have been incorporated in IR systems since natural languages queries form the intermediary between users and the system provided that an information need is most of the times expressed in natural language. Such techniques vary from query expansion (using semantic information) to query enhancement (using morphological information) and some of them have proved to enhance retrieval results for some languages whereas others either deteriorate or leave intact IR performance.

With respect to the incorporation of semantic information in IR systems several attempts have been made and most of them concentrate on the integration of thesauri in search engines. With the emergence of WordNet semantic networks it was attempted that such resources could be beneficial for IR tasks. In particular, both Princeton WordNet 1,5 and EuroWordNet were tested against IR environments and several report on their performance are available. Princeton WordNet aimed at query expansion tasks and it precision enhancement whereas EuroWordNet was mostly targeted towards recall enhancement. Despite the fact that at some cases some slight improvements of IR were reported by using semantic networks however many problems were also traced. The most important of which are summarized below:

- Query expansion using synsets deteriorated precision scores
- Multilingual Information Retrieval (MIR) did not greatly improve search results due to inconsistencies of the Inter-Lingual-Index
- Query disambiguation could not be resolved by the WordNet itself due to the lack of disambiguators
- Some of the language internal relations included in WordNet were not taken into consideration in IR applications

However, there might be other IR applications in which such a lexical resources could help and these are currently being examined. Once such application that has been examined by the BalkaNet consortium concerns the application of WordNets in performing conceptual indexing or conceptual classification tasks. After a limited market research was performed it was observed that web directories (e.g. Yahoo) and other classification techniques currently used by search engines are not extensively based on linguistics information stored in lexical databases, thesauri, corpora etc.

Our estimation is that the hierarchical structure and content of BalkaNet could greatly contribute towards this direction in an attempt to classify retrieved results in such a

way as to be meaningful to end users while at the same time satisfying their search results. Thus, the envisaged application of the BalkaNet database is targeted towards conceptual classification tasks, which will be performed via the domain labels that will be incorporated in it. The possibility of searching information in pre-specified directories, which will emerge from the network's conceptual domains, will increase the productivity and accuracy of any environment where a large amount of documents are used every day. This implies that the type of end users that benefit from conceptual classification tasks will be highly differentiated from specialized and professional experts to practically any type of users. The conceptual domains of the BalkaNet are going to function partly as directories in which information (documents) conceptually belonging to the underlying domain is going to be stored. This information will be useful fro retrieving information that belong to a desired domain without having to browse all documents retrieved, since all retrieved documents will be default fall in this category.

The implications of this resources or the user community will be extremely significant, because of the fact that the underlying semantic database will be able to produce a qualitative module for information retrieval products. The possibility of performing conceptual searching on the web will provide new services and will open up new market possibilities.

Functionality of project's results in Information Retrieval

Users' requirements from an Information Retrieval system concentrate on flexibility, ability to locate relevant documents as a response to their search requests, on-line help facilities, visual representation of the search results, advanced searching techniques (e.g. key-word browsing) etc. The aim of the BalkaNet project is to provide a basic semantic resource along with its infrastructure that could be incorporated in an information retrieval system with the aim of enhancing accuracy of the retrieval results and improving functionality of the system from the end users' point of view.

Given the current state of the art in information retrieval and taking into consideration the IR applications in which WordNet have already been tested against, the contribution of BalkaNet is targeted towards helping end users in retrieving information relevant to a particular conceptual domain (e.g. environment, medicine etc.). Thus, performing conceptual retrieval of texts instead of exact keyword matching as performed by most search engines significantly contributes to the following:

- Tracing of the desired information within an index that covers many topics or thematic areas, which most of the times are not unknown even by IR developers (except from the case of close lexical databases)
- Retrieving information from the web even if the keywords issued by end users are not included in the indexing keyword
- Keeping separate indices of the system where documents falling to pre-specified conceptual domains will be stored
- Facilitating updating of the indices
- Creating web directories that are language-independent
- Providing the possibility of incorporating new domains, sub-domains etc, from various lexical resources and languages

Once the BalkaNet database is available, it will be possible for end users to automatically chose their conceptual area of interested and perform their search request against the index they wish obtaining thus **better relevance scores** of their retrieved results. More specifically, the main contributions of the resource are going to be:

- Classification of documents in the index of IR systems
- Semantic indexing of documents
- Improvement of relevance of the obtained results
- Searching for information in the desired conceptual domain (directory)
- Helping towards interpretation of the search results

More specifically, in each IR system where BalkaNet's results will be incorporated there are going to be two separate indices. One, which is by default used by the system and where all documents reached by spiders are stored and an advanced one where documents will be stored in the indices in which they conceptually belong. That means that documents prior to be indexed are going to be processed according to their keywords (title terms, fonts of terms, position of terms). Then the semantic network will be consulted in order to trace in which of the domain(s) those keywords fall in. once the domains are traced the documents in question will be indexed in the respective indices.

Apart from indexing documents based on semantic information BalkaNet is also going to contribute to the actual performance of search requests. More specifically, once end users issue their keywords the system will automatically inform users on the conceptual domains in which their keyword belongs to. Then users will be asked to specify their domain of interest by clicking of the respective link and documents matching the initial keywords and found in the index of their interested will be retrieved.

In this way, the benefit for the end users is that they are going to view documents that are close to their information need excluding thus from their search documents that are irrelevant to their search request. For example if an end user enters as keyword the term "mouse" the system will not be able to understand by itself whether the users is interested in animals or computers. By using the BalkaNet database while indexing, documents that refer to animals and contain the term "mouse" are going to be classified under the domain "Animals", whereas documents about computers that also contain the term "mouse" will be classified under the domain "Computers". Once the users issues his keyword he will be notified by the system that his keyword might be about two different topics (namely animals and computers) and thus he will be asked on which of the two he is interested in. In case he wishes to retrieve information relevant to one of the above topics he will click on the respective link otherwise either he will clink on both links (if he does not know what he is actually looking for) or he will not clink at either link (if he wishes to retrieve any kind of document that might possibly contain his keyword).

Finally, BalkaNet semantic network can also be used in other NLP and IR applications in order to obtain a better performance in natural language understanding by means of considering the lexical semantic relationships. Such applications vary from lexical publishing to machine translation tasks and query expansion to word sense disambiguation applications.

BALKANET'S APPLICATION AND FUNCTIONALITY IN IR

Conceptual Indexing

The impact of WordNet has not only changed the way lexicography and linguistic research in general is being conducted but also opened up possibilities for new fields of NL applications. The question that arises is whether there can be meaningful data in WordNet without a meaningful structure and how the latter through the usage of semantic anchors and links can contribute to improvements of the performance of widely used IR applications. Towards the above directions the application of the BalkaNet project aims at shedding light on how lexical structures incorporated in IR and NL systems affect user's interaction with them and we especially focus on languages that are not widely studied. The reason for emphasizing on less studied languages is twofold. Firstly, the great differentiation that exists among the structures of these languages imposes the need for language and data independent structures for the homogeneous representation and manipulation of semantic information. Secondly, due to the fact that semantic networks have not yet been widely incorporated in IR systems for these languages and there is little evaluation results available to the research community. Finally, through semantic networks the structure of Romance languages can be further exploited and compared against Latin-origin languages in an attempt to enhance data retrieval systems and thus meet end users' needs in a better and more meaningful way.

The amount of data presented to end users of IR systems is growing explosively. While access and speed of such systems is being improved the user is more and more faced with the problem on how to deal with the massive amount of information, where to find what one needs in the huge network of information. Regardless of format or medium all this information has to be labeled, classified and stored in a meaningful way, in order to be retrieved by the user with the minimum effort required. The easiest, more direct and flexible way to access any kind of information is by means of natural language interfaces enabling end users to formulate search requests in general language and navigate through the retrieved information without having to know how a particular system works.

The BalkaNet semantic network can be used directly in many NLP applications ranging from word sense disambiguation tasks, language-learning tools and dictionary publishing. In addition the multilingual database may serve as a staring point for large lexical knowledge bases or as a source of semantic information to improve grammar and spelling checkers.

One envisaged application of the BalkaNet concerns its incorporation in Information Retrieval (IR) systems in order to support conceptual text retrieval as opposed to exact keyword matching. The available linguistic tool, which will support textual research will not be restricted to the English language, therefore access to the stored data will be facilitated for people who are non-English native speakers making thus multilingual and IR feasible. The most immediate application of the BalkaNet in IR tasks concerns the conceptual indexing of the documents, thus improving relevance of the obtained results. Much work has been conducted on how lexical information improved retrieval results but little attention has been paid to the contribution of conceptual information that is accessible via language internal relations.

Applying hierarchical links of concepts in indexing mechanisms allows us to search and query the text on the basis of topics, as identified by the lexical trees (ontologies) that are formed rather than on the basis of keywords, which is the most common form of search and index mechanism of text. Our approach concentrates on creating a lexical forest from the user's query and match this to conceptual tress (domains) that will be stored to indices.

The failing of keyword searches have long been known in the information retrieval community. With the proposed approach we give the user the ability not only to search for concepts that are related (synonymous) to the concepts specified in the query but also the ability to search for information within the conceptual domain they consider of interest. The closer the concepts in the query are the domains in the top of the network's hierarchy the stronger the evidence of the relatedness between the retrieved results and the specific conceptual domain.

Text collections can be indexed in IR systems in terms of the domain labels included in the ILI records. Thus the differentiation of the BalkaNet in comparison to the EuroWordNet project concerns the fact that the Inter-Lingual-Index will be a structured fund of the superset of all concepts encountered in all participating languages. The motivation behind structuring the ILI originates from various problems related to the mapping of senses in EWN. More specifically, because of high level of sense differentiation in ILI records there is a danger that conceptual equivalencies across WordNets are not linked to exactly the same sense of the English translational equivalent but instead connected to distinct ILI concepts reflecting different senses of the same word. In order to account for these diverging mappings from local WordNets onto ILI concepts, domain labels are going to be included in the ILI records and the latter are going to be structured on the basis of the domain labels and the top ontology so that terms linked to the ILI correspond to the same conceptual domain even if they are not exact translational equivalents, due to language differentiations. In addition, a structure ILI would mean a grouping of the ILI concepts that belong to the same conceptual domain enabling thus a preliminary clustering of terms and as a consequence a preliminary clustering of documents indexed on the basis of the ILI.

Such a representation can be used to perform language independent text retrieval since what is taken into consideration is the conceptual area that is the users interest and not the exact keywords a user types when searching for information. This approach differentiates substantially from keyword-based retrieval as performed by most search engines. Since in our case keywords entered by end users do not have to be indexing keywords.

Software developers and telematic-users have not had such much opportunity to experiment or gain experience with applications based in the semantic processing of information. Semantic resources are hardly available (especially for the Balkan languages) and most certainly there are not multilingual resources with sufficient semantic information available for these languages. In this respect, we expect that the clarity on the user-requirements and needs will arise from the availability of such a multilingual lexical resource and the possibility for people to work with it. The development of the BalkaNet will clarify the emerging use of future applications and will contribute to the development of new approaches towards IR tasks.

The motivation behind Conceptual Indexing Applications

The main application of the EuroWordNet and Princeton Wordnet 1.5 was targeted towards performing Information Retrieval (IR) tasks however their contribution to IR was did not always meet end users' requirements and did not greatly improved retrieval results mostly due to the lack of disambiguation techniques and morphological information in WordNets. Nevertheless, EuroWordNet apart from multilinguality also introduced a new element, and probably a more important, namely the semantic and domain classification using the Top and Domain ontology of the ILI records.

However, due to the fact that the Inter-Lingual-Index was an unstructured list of concepts many problems have been raised, i.e. if a synsets belongs to all Domain ontologies available how can we distinguish between them? Is the ILI structure enough for the efficient connection of two monolingual WordNets? Would it be more efficient to leave the flat file and move to another structure?

Drawbacks of non-structured ILI

By using a non-structured list of records, the only way to provide continuance and consistency among WordNets is by reassuring that every synset will be linked to some ILI record, either directly (explicitly) or via other connected synsets (implicitly).

Even with the usage of complex equivalence relationships among local WordNets and the ILI, in many cases the equivalence control for two synsets of different languages is totally impossible. For example: the synset A of one language is connected to an ILI record and synset B of another language is connected to the same ILI record, in case a hyponymic relation of the underlying synsets does not have a link to the ILI (while the condition of connection through its hypernyms is retained) we cannot possible know if A and B are equivalents and even if we check it the only way to allow verification or not of the above relation is by importing a new ILI record connected to both A and B. Thus, the major drawback of the unstructured ILI is that transitivity of the relations cannot be guaranteed, i.e. a Greek and a Turkish word that have equivalence link to the same Interlingua concept are not automatically useful Greek-Turkish translation pairs.

Another drawback is the connection of many ontologies with the same ILI record. In this case, all the attributes of the ontologies are inherited by those synsets linked to the particular ILI-record and describe the synsets of their hyponymic tree. In case a distinction no longer exists after a certain point of the tree, there is no way of rejection.

Benefits of a structured ILI

It's worth exploring the possibility of structuring the ILI by structuring and treating ILI as a WordNet but that will support a restricted number of relationships. More specifically, if we adopt the convention that the development of the monolingual WordNets will start from a part of the Basic concepts' tree, then for that particular tree the maximum compatibility and coverage among different WordNets is higher than if we simply started with the basic concepts.

If two trees of different languages are fully connected with an ILI tree, then the synsets of these trees can be used the same way in a wide range of applications. For different ontologies (classifications) we can have different trees on ILI, so the mapping procedure of different languages is much simpler based on a given

classification. ILI can still be used as a non-structured list of records, supporting total compatibility with the EuroWordNet model. In every ILI record any extra information, that might be useful, can be added independently languages and in the ideal case, if a solution to a problem of one language is found then the same solution can be applied to the other connected languages with the least effort. In addition, the same data can have many different representations. Finally, from the point of research, such a change might contribute in a number of applications and is extremely interesting considering it, if not applying it.

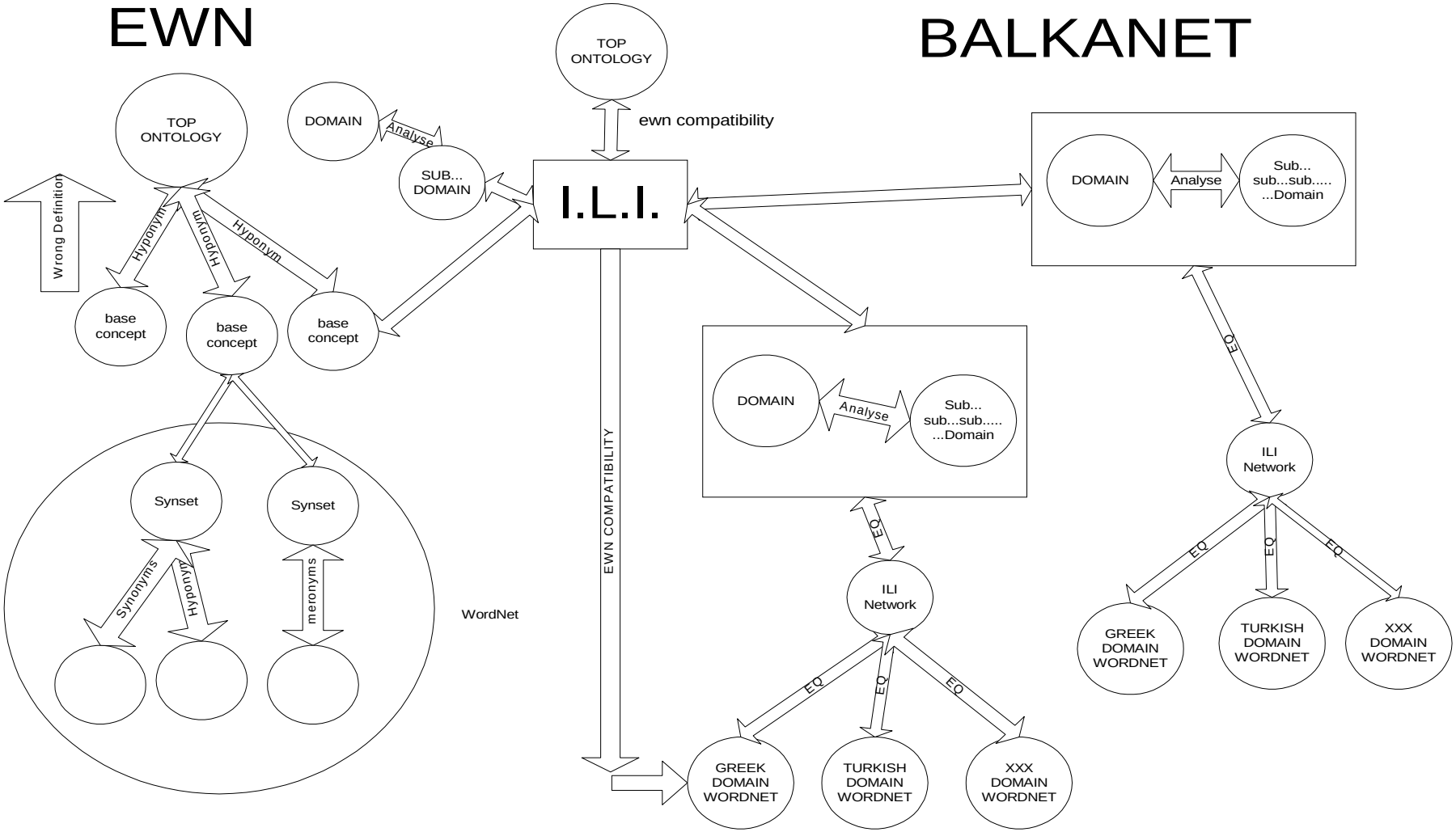
Summarizing, a structured Inter-Lingual-Index would gives us more flexibility to maintain and control the database and to be able to modify and change the ILI without having an effect on the monolingual WordNets.

Description of the architecture

From now on, the concept of the Domain is introduced to the ILI. The Domain is the equivalent of the Top Ontology and the Domain Ontologies and aims to the transfer, firstly to the ILI and then to the local WordNets, of the semantic or other attributes that it represents. Each element of the Domain is developed with the structure of a tree in the new ILI and each one of those trees represents concepts that belong to that particular Sub Domain and thus Domain.

ILI has a WordNet structure (there must be examined which relationships will be supported – certainly hypernymy and hyponymy) so that we have a net of concepts and not just simple records. Every concept of the ILI should be mapped to a domain through complex equivalence relationships (that should be examined, too). In proportion with the number of the Domains the corresponding domain-centric conceptual nets are created. Those nets are then connected to the local WordNets, and not one record connected to a tree like in the EWN, but one ILI tree connected to a local WordNet tree. Also, in the ILI will be kept information for the connection of every tree with the local WordNets and this way the connection and comparison of the local WordNets (which are domain WordNets from now on) will be more simple and efficient.

The architecture of the structure ILI is illustrated in the following scheme:


EWN
BALKANET
TOP ONTOLOGY
ewn compatibility
TOP ONTOLOGY
DOMAIN
Analyse
SUB... DOMAIN
I.L.I.
Wrong Definition
Hyponym
Hyponym
Hyponym
base concept
base concept
base concept
Synset
Synset
Synonyms
Hyponym
meronyms
WordNet
EWN COMPATIBILITY
DOMAIN
Analyse
Sub... sub...sub..... ...Domain
EQ
ILI Network
GREEK DOMAIN WORDNET
TURKISH DOMAIN WORDNET
XXX DOMAIN WORDNET
DOMAIN
Analyse
Sub... sub...sub..... ...Domain
EQ
ILI Network
EQ
EQ
EQ
GREEK DOMAIN WORDNET
TURKISH DOMAIN WORDNET
XXX DOMAIN WORDNET

Towards a structured Interlingua

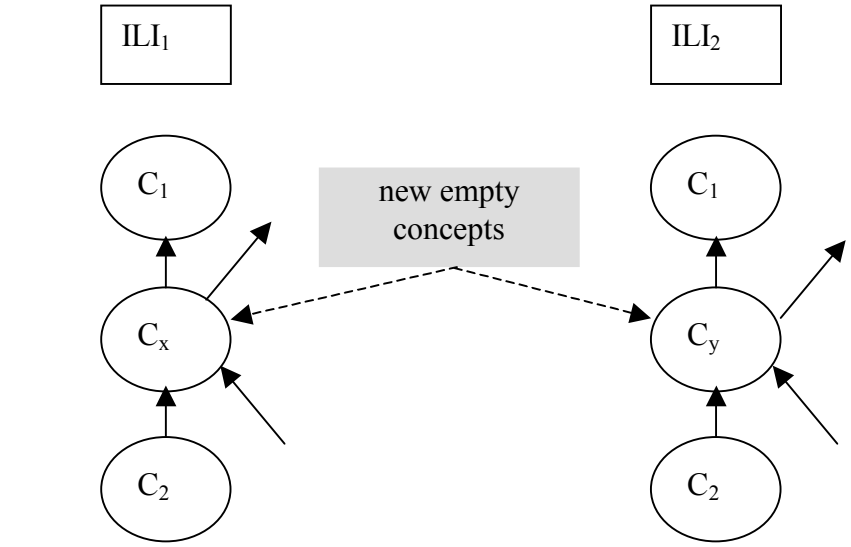
The structured ILI has three new features (compared to EuroWordNet ILI):

- It is rich in hierarchical relations (i.e. each of its concepts is part of a hierarchy). These concepts are mainly English WN 1.5 concepts with their corresponding glosses.
- Those concepts which do not belong to WN 1.5 are empty synsets that have a gloss associated with them and which are linked to at least one lexicalized synset belonging to some of the Balkan languages. The empty synsets of ILI represent linguistic concepts, which do not have a correlation in English but which have lexicalizations in at least one of the Balkan languages. The gloss of the corresponding synset in a language L is translated into English and associated to an empty English synset, and this synset is itself, like all the others, included in the hierarchy. The empty synsets in ILI allow for the inter-BalkaNet alignment of concepts specific to our languages but which do not have equivalents in English (e.g. "sarma"). As such, our network of WordNets will do more than EuroWordNet succeeds to do with respect to inter-language alignment. Note that once the language-L-to-ILI alignment is accomplished, there is no need to explicitly record the hierarchy in language L, as it can be at any time recuperated from ILI. The storage in the database of the L language WN could be useful only for a direct access to the hierarchy, not intermediated by the ILI, in case when the access is restricted to language L WN alone.

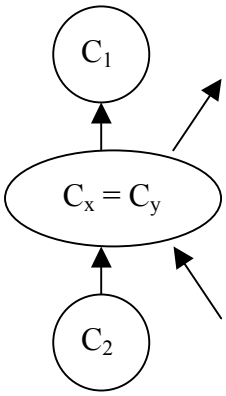
With respect to this extremely dense scheme of ILI, I see three problems:

- Concepts in language L WN, which have no clear alignment with ILI concepts (for instance, where the near-synonymy alignment relations were used in EuroWordNet). One possibility would be to reproduce the language L concept in ILI, translating the gloss in English, as suggested above, thus permitting to other languages to use this concept for an alignment with their specific concepts. Another possibility would be to maintain the near-synonymy relations, mainly for hard-to-take or temporary decisions, and to allow for a later decision, moderated by the consortium, who will discuss and agree upon inclusion or not in ILI of empty synsets (for instance, a more restricted view would include empty synsets only in the case of at least two languages sharing the same concept).
- The dynamics of ILI updating, namely what should be the methodology to update ILI in case of simultaneous development of individual WordNets? There, again, two strategies could be imagined:
  - A synchronous update of ILI – in which ILI is kept on the main Internet site of the consortium and the updating services of ILI are controlled by a server that sees the ILI as a shared resource. This strategy supposes software to support and schedule the accesses to ILI but no software and no human intervention for post-processing. The drawback could be a slow access during the updating process, since the ILI resource is remote. In such a vision, whoever needs a concept not originally in ILI either introduces it there, or finds an empty concept there, already introduced by a partner who created it before.

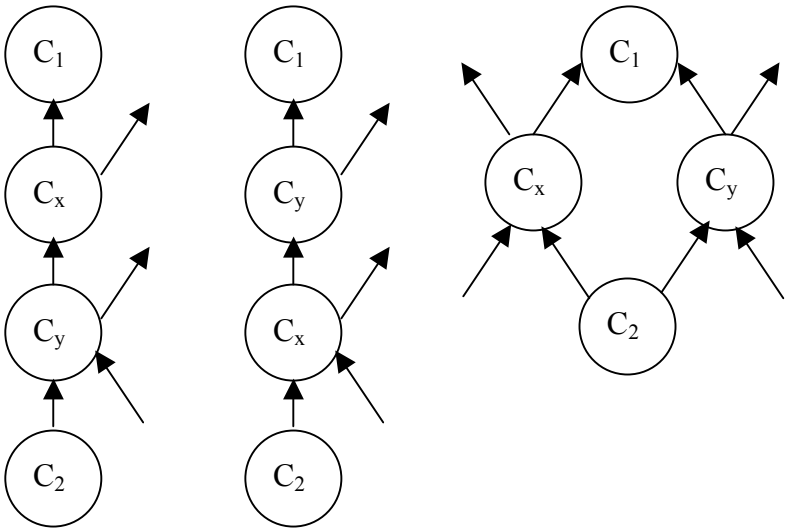
- A synchronous update of ILI – in which an original version of ILI (perhaps the WN 1.5 itself) is replicated to all partners, they do their updates independently, and from time to time a software is run to search for new entries in all versions of ILI and group those which are on the same levels (see figure). A human judge will decide upon the unification or multiplication of the empty synsets.



unification



different types of multiplication



- Proper names in one language could have no correspondent in ILI. In general, a view in which to allow for leaf concepts in the hierarchies not to be necessarily represented in ILI could be adopted.

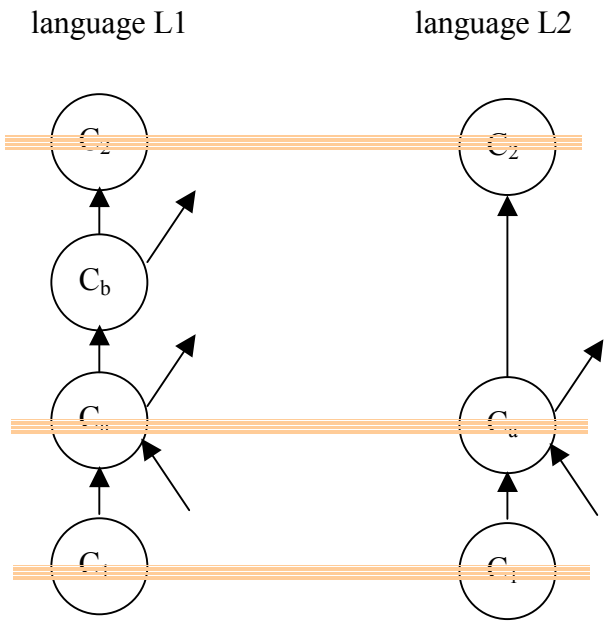
❖ At a certain level of ILI, basically, at the level of basic concepts, there will exist a tagging of those concepts with domain labels. So a label placed on an

ILI concept C is inherited, in one direction, by all the concepts below it and, in another direction, through the language-L-to-ILI alignment, by that concept of language L aligned with C, as well as by all those concepts placed below it in the language L hierarchy.

Such a philosophy is defended, in my opinion by three observations:

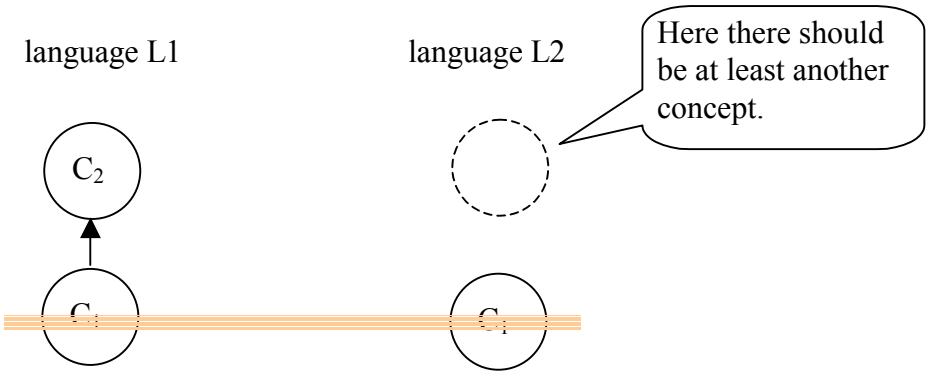
- The hypernymy relation (and its opponent – hyponymy) has a language independent definition (see the classics, Miller, Felbaum, etc.) and reflects a subsumption relation from a more particular linguistic concept to a more general one. As such, if H is the hypernymy relation, $^+$ is the Kleene operator (H^+ meaning at least one H relation), and by $C_1^{L1} = C_1^{L2}$ is meant that concept C_1 in language L1 is aligned with concept C_1 in language L2, then:

  if $C_1^{L1}\ H^+\ C_2^{L1}$ and $C_1^{L1} = C_1^{L2}$ and $C_2^{L1} = C_2^{L2}$, then it is a must that $C_1^{L2}\ H^+\ C_2^{L2}$ even if the chain of hypernymy relations in one language and the other could be of different lengths. The difference in lengths could be induced by the existence of concepts in the chain of language L_1 to which no concepts in language L_2 correspond (see figure):



- The top concepts are the same in all languages and are compulsory;
- No two top concepts are placed in a hierarchy (therefore one above the other).

One corollary of this system of axioms is that, if in one language two concepts are placed in a hierarchy, than in another language, above the concept that is assigned to the lowest one in the first language there should be at least another concept (see figure). This can be proved by *reductio ad absurdum* (if there is no concept above C_1 in language L2, than C_1 should be a top concept and according to note C above it is impossible that a top concept have above another concept – as happens in L1).


language L1
language L2
Here there should be at least another concept.
C2
C1
C1

The contribution of Internet Service Providers (ISPs) in applying BalkaNet for Information Retrieval

OTENET, subcontractor of Computer Technology Institute will validate BalkaNet semantic resource for their Information Retrieval system, namely a web search engine they are currently developing. The web search engine against which BalkaNet will be tested by performing conceptual indexing tasks comprises of the following basic components:

- Indexing of documents
- Morphological Query enhancement (i.e. normalization of both document and query terms)
- Fact and efficient search system
- Tracing of search keywords in the retrieved documents (i.e. highlighting query terms within the retrieved documents)
- Relevancy ranking of the search results
- Natural Language Interface
- Advanced search facilities

More specifically, the search engine, in which BalkaNet is going to be incorporated, indexes the full text of 700,000 Web pages with continuous update frequencies. It supports wildcards, Boolean searching, term as well as phrase searching; field searching (e.g. title: governmental, url:home.html, keywords), and case insensitive searching. The engine provides two search options: plain and normalized, thus there are two different indices kept, each corresponding to an option. The plain version of the engine indexes the pages fetched by the spider (including stop words), while the normalized one passes the pages through the normalizer, where stop words are excluded and the remaining tokens are induced to their first inflected form. Since the engine is case insensitive all tokens prior to indexing are turned to lower case. For each search request end users can either use the default (plain) mode of the engine or use the normalization mode. Whenever the normalization mode is adopted the query is normalized in order to match with terms from the normalized index. In both cases the query is converted to lower case so that the query terms can be matched towards the respective index. The display order or ranking of search results is determined by the engine from the location of matching words and occurrence of their frequencies.

Some modifications will need to be done to the search engine ion order to be able to perform a search request by seeking for all possible related terms. In addition, the design and functionality of the system has to be further improved so that multilingual documents are indexed and multilingual queries are supported. During the first phases of the project OTENET subcontractor is going to incorporate the English Inter-Lingual-Index in the search engine in an attempt to provide the main data source to users so as to perform search requests by using a hierarchical thesaurus of concepts. Later on the domain labels of the Inter-Lingual-Index are going to be incorporated in the engine as well and a preliminary evaluation will be conducted both by the system's developers and end users. Based on the evaluation feedback OTENET is

going to make all the necessary improvements to the system in order to meet the users' needs. If necessary and in case the so far existing domain labels are not sufficient for performance if conceptual retrieval tasks new domain labels and possible sub-domains are going to be added as well in the semantic network. Quality and quantity measures are going to be used during the evaluation and some awareness activities will be run in order to disseminate the project's results and inform end users of the system's capabilities.

The contributions that OTENET is expecting from the BalkaNet consortium are basically the following:

- The possibility of indexing and classifying documents on the basis of the domain labels of the Inter-Lingual-Index
- The possibility of performing conceptual text retrieval by using domain labels as a kind of web directories in which relevant documents will be found.

There are still other possibilities where BalkaNet could contribute but the step forward that the project implies is for the moment restricted to conceptual classification of indexed documents. A close exploitation plan for future applications of the project will be released at a later stage of the work.

Evaluation of BalkaNet's contribution in IR

Given the specific application of the BalkaNet project in an Information Retrieval environment different tests and methodologies for the evaluation of its contribution will have to be developed. The test sets will involve various sets of queries issued by different user groups (e.g. experienced users, inexperienced users, professionals in IR evaluation etc.) in an attempt to illustrate the effect of semantic classification in relevance of the retrieved results. Tests will be differentiated for different levels in the hierarchy and by making use of different kind of lexical information (ambiguous, polysemous terms etc.). Furthermore, we will investigate to what extend the general vocabulary is complementary to conceptually-based texts classifications and to what extend different information retrieval tasks have any effect on these.

The performance of the tests will be based as a measurement of the additional functionality and quality of the monolingual WordNets. In addition, the queries will be selected and designed in such as way to elicit potential problems while using WordNets in IR such as lexical ambiguity problems etc.

The main criteria adopted for the evaluation of the system's performance are summarized below:

- Precision scores obtained by the engine and relevance scores provided by end users and evaluators
- User involvement in query enhancement by using the domain labels
- Integration with other NLP techniques already present in search engines
- Integration with other document classification techniques
- Recall scores

Moreover, for the evaluation of the abovementioned criteria the following tests will be applied:

- Application of a set of queries without using the BalkaNet domain labels

- Application of the same set of queries with the adoption of the BalkaNet domain labels
- Application of the same set of queries against directory services provided by other search engines
- Application of the same set of queries with the adoption of sub-domain labels
- Application of the same set of queries with the adoption of both domain and sub-domain label
- Assigning weights to keywords for an efficient retrieval
- Examination of the engine's log files to see how users interact with it
- Issuing as a query a keyword which also forms a domain label
- Assessing ability to use domain labels by non expert users

Most of the tests will be performed using the search engine provided by OTENET. However, in order to be able to compare the acquired results with the performance of other systems we are also going to test performance of other systems that support documents or query classification and web directories in order to have a qualitative overview of BalkaNet's performance.

However, even if BalkaNet semantic network proves to be quite beneficiary for semantic classification tasks there might be some areas that will need further enhancement such as the handling of multi-term expressions issues by end users. Thus, the project's application is mainly targeted towards handing single term queries since after all those are the most frequent types of queries issued in IR systems especially by inexperienced end users.

BalkaNet cannot provide functionalities for every type of query (e.g. Boolean logic operators, wildcards etc.) and within the framework of the project no such techniques will be supported despite the fact that the underlying search engine that will be uses already supports such modules. The only extended facility that BalkaNet can provide with respect to query handling is some multi-word data (lexical items for certain meanings that are multi-word expressions) just in order to demonstrate whether they can or not be supported by the IR system.

EVALUATION OF BALKANET'S PERFORMANCE

Internal testing of BalkaNet's results consists of several generic tests that will evaluate the integrity and consistency of BalkaNet multilingual database. The most important test concern:

1. **Checking statistics, syntax and content of the resource**
2. **Assessment of the general interest and added value of the multilingual semantic network**
3. **Perform experiments on the usefulness of the results without taking into consideration statistics.**

1.1 Statistics

Since among users of the intermediate results of the project are not Balkan speakers testing of syntax and content will mainly include the following:

- Lexicon: number of words
- Lexicon: POS distribution
- Synset: statistical distribution (number of synset with one word, tow words etc.)
- Lexicon: corpus frequencies (from available frequency lists)
- Number if internal relations (e.g. from a Romanian synset to another Romanian synset)

1.2 Syntax

The user of the project, namely Memodata, will make an attempt to incorporate the intermediate results of the project in the Lexidion or the Semiograph tools in order to check the syntax of data included in the multilingual database. Unlike the EuroWordNet project, where the Polaris editor was used for automatically checking data before its incorporation in the database in our case BalkaNet's content will be checked against the aforementioned tools in order to ensure that the correctness of data prior to its incorporation in the database.

1.3 Content

The main objective of this testing phase is to detect errors on data that might be introduced by the lexicographers. Such errors might appear locally to a term or a synset and will be of the type:

- Spelling errors in literals
- Formalization of lexical items

Whereas other might be structural ones such as:

- Unexpected loops in the hierarchies
- Forbidden links
- Infinite loops
- Symmetry between inheritances etc.

2.1 Number of links between the different languages

BalkaNet lexical database can be incorporated in or be helpful to many applications varying from machine translation to comparisons of linguistic phenomena across languages.

However, in order to assure its usefulness there must be a qualitative checking of the data it incorporates. More specifically, one type of checking this concerns testing of the different links between the languages involved. This is an apparent need since in EWN database monolingual WordNets have no synonymy links to the ILI apart from their Base Concepts. If this is the case in BalkaNet too monolingual WordNets will have no comparison to 5, 6, 7 or 10 languages that have a lot of links. Thus, by testing the number of links across languages we attempt to assign some grades to each Balkan language to measure the effort required to cover the ILI. Within this test links like "near synonym" will have a grade of zero in cases where there is a difference between the sense given by WordNet 1.5 glosses and the sense given by term of BalkaNet. In addition, such differences have to be explicit. Commonly, these differences originate from:

- Level language: synonym_but_popular_while_English_Is_Not_popular,
- Morpho/syntax: noun_to_verb_synonym,
- Syntax: synonym_in_the_case_of_gentive

In the case where the difference is correctly given, the relation will have a grade of one.

To facilitate this operation, and to optimise our strategy (of making the choice of the lexicon), it is possible to provide a table with each of the WordNet 1.5 synsets and 6 Boolean fields that shows the EWN synonymy link result.

Example: a row like "*59683, n: 010101* " will say that the noun 1.5 synset number 59683 has 0 synonymy link to Italian, 1 synonymy link to German, 0 to Estonian, 1 to French etc.

From this kind of data, we will calculate complete grades of the interest of use of each WordNet. This grade WILL BE NOT a qualitative grade, since it might be possible that a particular Balkan language has no possibility to merge a lot of WordNet 1.5 synset. But, in cases where this notation will be low and well justified a more complex referential semantic resource will be suggested by the users to the consortium. One possible solution or suggestion towards the above direction might be the presence of more links between syntax and the semantic lexicon.

2.2 Usage of the Internal Relations

With respect to the internal relations it should be noted that a language with no coherent links with WordNet 1,5 will not be of any direct interest as far as the application of the project is concerned. This is mainly due to the fact that no gloss will be available for terms that are not linked so it will be hard to check the meaning of the underlying terms and thus make sense discriminations.

2.3 Concrete Experiences of Use without taking statistics into consideration

During this phase what will be tested is the successful import of each monolingual Balkan WordNet in the Lexidiom or the Semiograph tools, the performance of the test results and the amount of work performed within the time limits of the project. In addition, during this phase several experiments will be run in order to test how the BalkaNet database can enhance obtained results of Information Retrieval (IR) systems with emphasis given on Word-Sense-Disambiguation (WSD) tasks.

2.3.1 Information Retrieval (IR) Tasks

Memodata will attempt to provide a retrieval query language once BalkaNet is incorporated in IR systems by expanding already available software with some filter control services. However, before this task is performed a feasibility study will be conducted first.

2.3.2 Word-Sense-Disambiguation (WSD)

By loading the BalkaNet data to the Semiograph tool, an attempt will be made to perform sense disambiguation tasks without a prior training for some small pieces of pre-selected text. Some of the tasks envisaged with in this framework concern the incorporation of the conceptual structure of the Internal Dictionary (a database that can calculate ‘florist’ from ‘a person who sells lilac’) to WordNet(s) we will try the same sense disambiguation. We will note down the results and compare this result to previous ones obtained within the framework of other projects. At this point it should be noted that this experiment will be tried against languages with good formal test results and with a many homonymous terms.

TOOL REQUIREMENTS FOR INTEGRATION IN IR APPLICATIONS

In order for this architecture to work efficiently in an Information Retrieval Application there are a lot of issues that need to be addressed carefully during the design and the implementation of the project.

Most Information Retrieval Applications consist of an indexing mechanism which is responsible for storing the available data in an effective way and providing a way of accessing them effectively by providing one or more keys, and a ranking mechanism which evaluates the data returned from the indexing mechanism in order to find the most relevant among them and present them to the user first.

Since the purpose of this project is not the construction of a new Indexing Mechanism or an Information Retrieval application, in order to test the feasibility of the proposed application (i.e. Conceptual Indexing) already existing mechanisms can be employed, either free source on the web (e.g. htdig -- www.htdig.org) or provided by the partners (e.g. CTI' s Search Engine).

As far as the indexing mechanism one could follow two approaches. The first one is to try and alter the documents prior to indexing so that the terms already found in the constructed WordNets are replaced by their respective concepts and these are stored in the database. With that approach when a query is issued from the user, this query is also transformed in a similar way, and the new modified query is issued to the document database for matching.

The second approach has to do with the so-called query expansion. In that case the documents are stored in the indexing structure of the application verbatim, but when the user issues a query it is expanded by appending to it the keywords in the WordNet that exist between a specific term in the query and the concept that this query keyword belongs to. In this case one is trying to improve the recall of the indexing/searching mechanism by asking for more information at the same time.

What both of these approaches require is an effective way of searching the constructed WordNets for a specified keyword/concept. Since there is no way of knowing whether a term exists in this WordNet unless somebody searches for it this operation will have to be executed a great amount of times for each transaction either indexing or user query. Consequently the main challenge of the IR application is a really quick searching mechanism that is able to represent the whole WordNet structure and data. This structure should be easily updateable and expandable and should be able to incorporate more than one language at the same time (e.g. for a multilingual IR Application).

As far as the Ranking Mechanism is concerned, again this is not a goal of this project. A lot of work has been conducted and a lot of various mechanisms have been proposed (e.g. Google). For this cause we plan to use some ranking algorithm that is suitable for our purposes and that performs relatively well so that we are in position to compare the returned results and whether the conceptual indexing works. At this point we do not think that it really matters what kind of ranking algorithm one uses because our main intention is to evaluate whether the resources acquired during the project are suitable for such a cause i.e. IR Application.

Of course in all of the above we considered that the collection of the indexing structure is static and already acquired. We are aware that almost always this is not the case, but we plan to create a small document collection on which we are going to test this application.

DATA MANAGEMENT AND REPRESENTATION

In this section we will outline how the BalkaNet project work will be coordinated in a convenient and efficient way. We aim at a platform independent distributed environment where information can be obtained and managed, either locally or remotely using web technologies. By these means, an application will be developed, offering through a user-friendly interface full navigational and management capabilities. In order to provide such functionality, diverse groups should be created, assigning different roles and permissions. The application acts towards the users accordingly to their group policy, giving various levels of service awareness, such as hierarchy browsing, statistical information retrieval, word net updating, etc.

Information retrieval and management will be on top of distributed system architecture, using a proper protocol for intra-communication. This is the so-called 'On-line' mode, in which an authorized user can access and/or modify all interconnected information (e.g. connections of synsets among Balkan languages). An 'Off-line' mode also exists, where services operate only on local word network (e.g. viewing and building language hierarchies). In any of the above cases the application should be capable of providing the user with the manipulated data in a predefined XML format. However the storage method of data is still an issue under consideration, and thorough research will be done.

PART B

BALKANET'S FUNCTIONAL REQUIREMENTS – DEVELOPERS REQUIREMENTS

SPECIFICATION OF THE DATA REQUIREMENTS

➢ Linguistic Requirements, Vocabulary Coverage and Selection Criteria

As a primary goal, the BalkaNet project aims at covering general vocabulary at a percentage of 80%, while it is the consortium's secondary goal to include a certain proportion of domain-specific vocabulary as well, so as to enable testing the resulting multilingual resource in various ways and NLP applications. Since the BalkaNet project is going to be incorporated in an IR application it has to contain a sufficient number of words and meanings. Thus, as far as the multilingual content is concerned, the project aims at a total set of 15.000 comparable synsets in each language, corresponding to more or less 30.000 literals. A literal will be either the orthographic representation of an individual word or string or individual words jointed with the underscore character. Such strings of words are called *compounds* or *phrases* and represent a single concept. However, in some cases as for example in the case of Turkish collocations might be lexicalized in the sense that they are orthographically a single token or they just appear together. Such words will be handled separately as single terms after deciding first whether a collocation is lexicalized or not.

A word form might be augmented with information necessary for the correct processing and interpretation of the data. Towards this direction an "integer" (offset) or other kind of identifier will be added to each term so that lexicographers can distinguish duplicates in a single lexical file. Thus, a term in a synset will be represented by its orthographic word form, its semantic field and an identification number.

The part of speech distribution among synsets will be 65% nouns, 25% verbs, 5% adjectives and 5% adverbs and the coverage of the vocabulary has to reflect frequencies extracted from both corpora and lexicons. A synset will consist of synonymous words, relational pointers and a gloss. Comments for use by the lexicographers can be entered in a lexical file outside the synset but they will be discarded when the database will be built.

The lexical resources that could help terminologists decide on the set of terms to be incorporated in the multilingual semantic network vary from dictionaries to corpora and mostly depend on the resources available for each language in electronic format so that they are easily processed. However, in case the electronic lexical resources are not sufficient, printed resources will be used as well with the only drawback that their processing will take much longer and will require a lot of manual work. A thorough process of the lexical resources is required in order to ensure that all concepts important and representative of the participating languages are traced. Importance at this stage is determined by the occurrence frequencies of terms in the lexical resources. In addition, it is very important that consistency across languages is reassured. In order to achieve that extraction of the full set of base concepts along with their equivalences in other language is necessary. Finally, during the extension of

the monolingual WordNets new entries will be added on the basis on their ranking in frequency lists extracted from various corpora.

More specifically, the kind of resources that are going to help terminologists along with the kind of data that could be extracted out of them are listed below:

- Explanatory dictionaries: the information that can be extracted from explanatory dictionaries ranges from a POS tag for each lemma, to the number of senses each lemma holds, definitions / descriptions of each sense etc. Specifically, selection criteria will be based on the number of senses each headword has in explanatory dictionaries and the number of definitions a headword participates in. At cases, lexical productivity (the number of derived words/lexical items from a base lexeme) will also be taken under consideration, since in some cases and for certain of the participating languages, derivational processes produce predictable senses that are expressed by stereotyped definitions in explanatory dictionaries and consequently, separation of senses on such a derivational axis could serve as a criterion for sense distinction. Multiword expressions are going to be dealt with to the extent that they are necessary for the language-specific representation of concepts in each language, also taking under consideration the considerable amount of work required for their extraction and processing.

- Monolingual and Multilingual Corpora: corpora might comprise of various kinds of texts (e.g. literature, manuals, newspaper texts etc.) and will have to be lemmatized and tagged prior to applying frequency metrics on them in order to extract the most representative terms for a particular language. In addition, concordance lines of terms should be kept since we target towards the development of a semantic network and thus we are mostly interested in the frequencies of meanings and not the frequencies of word forms. In cases where bilingual corpora are being used particular attention should be paid so that there is a semantic alignment of the texts comprising the corpus and if necessary semantic annotation of the texts should be made. Due to the lack of large multilingual semantically aligned corpora for the Balkan languages one possible resource for obtaining this kind of information might be some official documents taken from the Commission's web site. Of course such a sample corpus will only be used for testing the conceptual alignment among terms where possible. In the case of bilingual corpora concordance lines should be kept as well in order to facilitate terminologists decide on the correct translation of a term by taking into consideration its context. However, it should be taken into consideration that occurrence frequencies of terms in running texts are considered by many lexicographers to be a very subjective criterion. Among the strongest arguments with respect to subjectivity is the volume and representativeness of the texts included in the corpus subject to the quantitative analysis. With more and more texts available on the Web the size of the data is not anymore a significant issue, but the representativeness remains a systematic complain. The exact

definition of what kind of texts should be included in a corpus for data analysis is a long-standing debate and our aim is not to solve such problems. Due to the aforementioned reasons monolingual and multilingual corpora will be used to the extend that they are useful for checking terms importance and perform tests again terms already extracted from the corpora. The main objective behind using a corpus is to verify meanings of terms extracted from explanatory dictionaries and to check terms that might be ambiguous or polysemous. Finally, occurrence frequencies of terms will be complementary with other kinds of information such as terms' importance in dictionaries, thesauri etc.

- Bilingual Dictionaries: bilingual dictionaries are going to be a rather useful resource of information during the development of the project since linking of translational equivalents of monolingual WordNets against the Inter-Lingual-Index will be mostly based on translations provided by bilingual dictionaries. However, for some of the participating languages (e.g. Serbian) there are no reliable bilingual dictionaries available in electronic form. In such cases printed dictionaries will be used but a lot of manual work will be needed. In general, it would be preferable that for each language more than one bilingual dictionary is used in order to reassure the correct translation of a term in the target language. However, in some cases bilingual dictionaries might be inadequate in the sense that a more specific word is translated with a more general word when there is no direct equivalent. Especially in the case of denotational gaps a bilingual dictionary may give a more general term, which would be more appropriate than the full phrase as a translation. Another problem that might come up is that for each term of the source language bilingual dictionaries might give more than one translations for the target language and the order of the translated terms might not be the appropriate. In particular, the first term given as the translational equivalent of another one might just be the most frequent term for the respective language and not conceptual closer to the term of the source language. However in our case we are interested in translating a term with the one that is conceptually closest. One possible solution for overcoming the above problem is to check the translational equivalent of a term against monolingual corpora in order to detect whether the correct sense of the translated term is assigned. At a latter stage and once the core monolingual WordNets are developed we can check translation mismatches in the following way: if the translation relation between words of two languages are differentiated in terms of the same relations that are distinguished in the WordNets we can thus use this information to check the lexical semantic configurations in each WordNet and derive more information. In addition, in cases where translational equivalents of terms are no available in bilingual dictionaries manual translations will have to take place. Moreover, in some cases translational equivalences in lexica might be based on sense equivalencies in context and in such cases transferring the

ontological description from one English term to its equivalent translation will be the legitimate option.

- Thesauri: another useful resource for obtaining terminology to be incorporated in the monolingual semantic networks are thesauri, which form actually a kind of dictionary where entries along with their synonyms are stored. For some of the participating languages thesauri are already available and they are going to be used by terminologists in order to extract synonymic relations among candidate terms. However, the quality and coverage of thesauri has to be tested in order to conclude on their usefulness for the project. Testing is required since in some cases synonymy among terms is either direct or explicit, i.e. terms are not always interchangeable in context but under some circumstances one can replace the other. Due to the nature of thesauri and the way in which information is stored and organized in them terminologists cannot always decide on which meaning (sense) of a terms is the exact or near synonym to the other term. One way of testing the quality of a thesaurus could be against corpora by replacing each term with its synonyms. Since a testing phase would be rather time-consuming a small sample of the thesaurus could be tested in order to conclude on its quality and usefulness.

- Domain-Specific Glossaries: for some of the participating languages domain-specific glossaries are available comprising of terminology restricted to a particular conceptual domain, i.e. medicine, science, environment etc. Such glossaries are quite useful for terminology acquisition due to the reduced degree of polysemy and lexical ambiguity these terms hold. In particular, terms belonging to a particular conceptual domain tend to share a single meaning overcoming thus polysemy problems. However, since the BalkaNet semantic network is going to comprise mainly of generic vocabulary of the languages involved the use of domain-specific glossaries will be rather limited. Domain-specific terminology will be incorporated just in order to demonstrate the possibility of including domain-specific concepts in our network and as a consequence only a few such glossaries will be used.

- Morphological analyzers: even though monolingual WordNets are going to contain semantic information of the underlying languages, for some of the participating languages (e.g. Serbian, Czech, Bulgarian) due to their rich morpho-syntactic characteristics morphological processing of their lexical resources might be required so that useful information can be extracted out of these. In particular, for the Czech language algorithms implementing word derivational processes might be applied during adding concepts of the Inter-Lingual-Index to the Czech Wordnet and in this case a morphological analyzer that had already been developed, namely AJKA and tool I_Par are going to be used.

The above lexical resources are going to be rather helpful not only during the selection process of the candidate terms for the monolingual WordNets but during

WordNets enrichment and correction as well. Especially in cases where synsets will be developed semi-automatically from (bilingual) dictionaries some wrong translations might cause inconsistencies, thus additional lexical resources will be necessary for the detection and correction of such mistakes. Moreover, in cases where the EWN or WordNet 1,5 are followed some hypero/hyponymy trees might not be consistently connected.

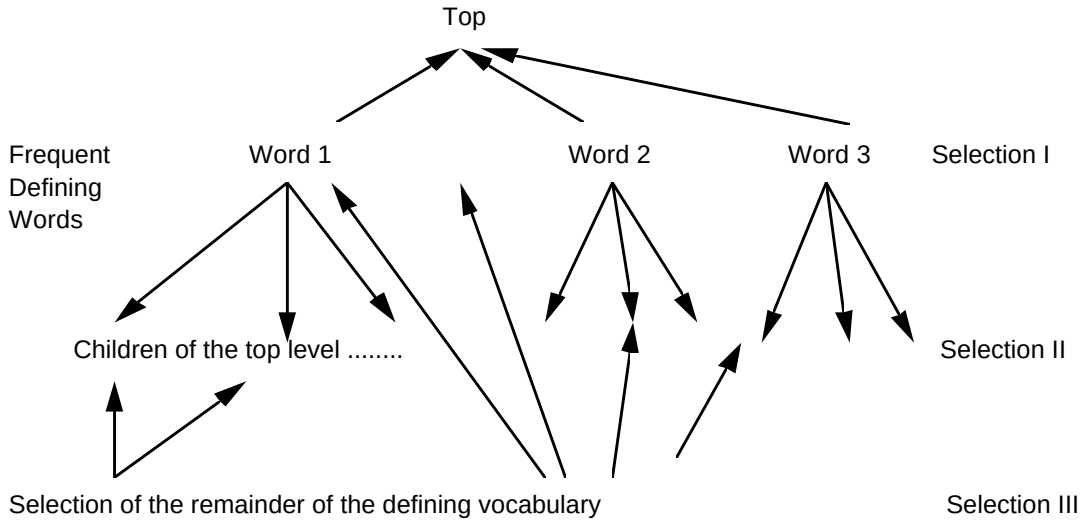
- ✓ Data Requirements: Once each contractor processes the monolingual lexical resources the actual development of synsets will begin. The BalkaNet data will comprise nouns, verbs, adjectives and adverbs in Greek, Turkish, Bulgarian, Romanian, Czech and Serbian. We aim at a total of 8,000 comparable synsets, correlating with about 15,000 most frequent literals for the Balkan language involved. The selection process will have the following characteristics:

    - There should be maximal overlap of the covered concepts across the monolingual WordNets
    - The covered subset has to be generic i.e. all frequent words of the language with their most frequent and common senses should be present
    - Every concept must have a parent concept so that the introduction of new items does not require the enrichment or the modification of the top-concepts.
    - Domain-specific terminology is going to be added in a limited extend just in order to demonstrate the possibility to augment the data with domain-specific vocabularies and to be able to perform domain-specific information retrieval tests with the integrated results.

The actual selection of the candidate concepts will take place in two phases. The first subset of terms will be extracted from the defining vocabulary of each dictionary or lexical resource from which the monolingual WordNets will be derived. The advantage of this methodology is that all terms will be linked with other terms in the lexicon and they are also going to be rather representative (frequent) of the underlying language. Another benefit of the first selection described above is the fact that terms correspond with top concepts in the EuroWordNet database. This is important to ensure compatibility of the top-level across the WordNets.

Once the first set of terms is extracted then the development of WordNets will follow a bottom-up strategy i.e. terms included in the definitions of the extracted terms will be linked to them. The main criteria adopted while the second selection process are the frequency of the defining terms, their occurrence as entries in monolingual dictionaries, and their relation to the first subset by any of the pre-defined semantic internal relations. Furthermore, any extension of the vocabulary in the WordNets will involve the linking of more specific words to well-defined and delineated concepts in the WordNets. That means that extensions will not introduce new hierarchical top-concepts.

The methodology described so far for processing of the lexical resources and concluding on the final set of terms to be included in the monolingual semantic networks is pretty much the same as the one followed for the implementation of the EuroWordNet project and it is called the expand model approach for semantic networks construction. The way in which terminology acquisition and selection will be achieved is illustrated in the following schema as taken from the EWN project.



To achieve sufficient overlap and compatibility we will also make use of the consistency checking and Wordnet-comparison mechanisms that will be implemented in the BalkaNet database with the contribution of the project's user, namely Memodata. To achieve sufficient generality of the concepts included in the monolingual WordNets frequency lists will be extracted from available corpora. Other criteria applies will be the length of the terms, morphological complexity and the degree of polysemy. However, we expect that the defining vocabulary extracted during the first selection process will mostly coincide with the more frequent words in everyday language use and as such the information extracted from the corpora will be probably minimal. On the other hand the contribution of corpora might be rather useful in the determination of the senses of the terms that are going to be included. In cases where there are no available corpora for some of the participating languages the selection of senses will be performed on the basis of labeling in the dictionaries. Such a task will take place in order to avoid obscure and rare senses, which however might be added later on and can be linked to other senses as specific variants.

- Monolingual: the monolingual data requirements that will be taken into consideration while starting the actual development of the core monolingual WordNets are summarized below:

  - **Number of terms and distribution per Part-Of-Speech (POS)**: each monolingual WordNets has to comprise a sufficient number of terms and meanings in order to improve the performance of IR and other NLP applications. More specifically, it has to comprise of approximately ~15,000 word meanings and ~8,000 terms. The POS distribution has to be around 65% nouns, 25% verbs, 5% adjectives and 5% adverbs. The vocabulary coverage has to reflect dictionary or corpus frequencies and the selection process has been described above and it should be possible to handle multiword

expressions. In addition, each monolingual WordNet has to contain at least 85% of the general vocabulary for each language resource and a proportion of sub-language to show the capability to incorporate domain-specific vocabulary. In addition, lexical items and glosses included in the monolingual networks have to be correctly spelled and every meaning of a term has to have a gloss attached that will explain its sense. This gloss will be stored in the Inter-Lingual-Index.

- Multilingual: the multilingual data requirements that will be taken into consideration while developing the monolingual WordNets and prior to linking terms with the Inter-Lingual-Index concepts via multilingual links are summarized below:

  - **Coverage of the vocabulary and overlap between the different languages:** the covered vocabulary of the different languages need to have the highest possible degree of overlap and all the languages represented in the semantic network will have to be 100% to the Inter-Lingual-Index. With respect to the ILI each record will have to have an ID attached compatible with the ID identifiers used in the EuroWordNet project. Changes to the ILI will involve relations to existing ILI records that do not currently exist or the addition of new ILI records along with their IDs. The Inter-Lingual-Index will have to be fully automatic and easy to maintain. The Czech partners are going to be responsible for updating and managing the already exiting new ILI files as well as the new files that will be incorporated in it within the framework of the project. Moreover, the Inter-Lingual-Index is going to have links to structured top-concepts and the Top-Ontology of the EuroWordNet project as well as to the domain labels that will be incorporated in the ILI once they are available. The motivation behind structuring the ILI lies behind the performance of conceptual indexing tasks and is described in detail in a separate of the present document. However, what should be mentioned at this point is that the data structure has to take into consideration possible extensions of the information stored in the ILI with new / additional conceptual features (e.g. inclusion of sub-domains, semantic features etc.)

- ✓ Spelling of Lexical Items and Glosses

It is very important that the lexical items are correctly spelled. However, different orthographic forms of the same terms might appear in a language due to dialect variations, two-formed terms etc. Thus, in cases there is more than one spelling / orthographic form for the same term in a given language, one standard form has to be followed. The same applies for the representation of terms in the Inter-Lingual-Index where a standard spelling has to be followed as well, e.g. USA English vs. UK English.

- ✓ Domains

Since the application of the project's results is mainly going to focus on conceptual indexing tasks the incorporation of conceptual domains in BalkaNet is going to be a rather important task. For deciding on the final set of domain labels that will be included in the multilingual database and the Inter-Lingual-Index a close investigation of already available domains should take place. More specifically, the consortium has discussed two possibilities so far. The first one concerns that we define our own conceptual domains that will come up after an extensive processing of the available lexical resources and the second one implies the use of already existing domain such as the Yahoo directories. Despite the fact that the first proposal would results in more representative domains for the underlying languages however it is a rather time-consuming option and would require extra working hours. Taking into consideration the above and by having in mind that we wish our network and the conceptual areas it covers are applicable to all languages independent of their origins we decided to follow the second approach and use already available conceptual domains for the organization of our concepts in the network. In addition, domains that are available are general-purpose domains meaning that they are widely applicable by all languages and for almost any task. In light of the above the consortium concluded that such domains could be obtained from digital libraries or search engines or even corpora. In order to conclude on which particular domains will be used members of the consortium are going to closely examine available domains and select those that better meet the project's objectives. Towards this direction is was also discussed the possibility that terminologists start working having in mind 2-3 conceptual domain which are rather generic and are met in every language and later on new domains are added to them. Other domains that have been proposed so far by the contractors are: journalese, legalese and national versions of ISO standards

Domains in the BalkaNet project are going to function pretty much like the Base Concepts of the EuroWordNet project with the only differentiation that they are going to be more generic ones and applicable to any language. Despite the above domains are going to be rather generic concepts that will be linked with the Top-Ontology of the network via hyponymic relations. Each synset might belong to one or more conceptual domains depending on the meaning (glosses) of its terms. In addition, the possibility of including sub-domains as well has been also briefly examined and it was decided that sub-domains might be added at a latter stage of the project and once domains are pretty much defined and fixed. Finally, it should be noted that conceptual domain labels are going to be common across the participating languages and they are going to be incorporated in the Inter-Lingual-Index of the network under which the English concepts will be organized in the way described above. Domains can be seen as another way of grouping concepts in larger schemes or scripts and may comprise a diverge range of concepts. Domains are very useful for information retrieval applications since indexing of documents will be performed on the basis of the conceptual domain labels attached to each concept. Since the domains can be organized as a semantic hierarchy as well they are considered as a separate type of objects in the database between which relations can be stored. Despite the fact that domains were initially foreseen in the EuroWordNet database they were not extensively used or tested since such a thing was not within the main objectives of the project. Thus, our differentiation with EuroWordNet lies on the fact that domain labels will be included in our semantic network and it will also be possible for end users to customize them in the future for their own applications.

The purpose of domain information is twofold. First and foremost domain labels can be directly used in information retrieval tasks to group concepts based on scripts rather than classification. In addition, using domains we can separate the generic from the domain-specific vocabularies. Such a thing is important to control the ambiguity problem in Natural Language Processing and to customize the general resource for specific applications. Since the domains are going to be language-independent (language-neutral) they are also going to be stored at the ILI concepts as language independent top concepts with a *belongs-to* relation. Thus, a user who wants to customize the WordNets does not have to store its domain information in every separate WordNet but only fro the relevant ILI records. The link relations from the ILI records to the WordNet synsets will then match the domain information with the language-specific WordNets.

- ✓ Instances – Proper Names

Instances are distinguished form meanings to express the semantic difference between denotational lexical items and referential lexical items. Typical instances are cities and country names e.g. Paris, Belgrade etc. this information cannot be derived from the available dictionaries but the object types have to be distinguished so that users can customize the general and generic WordNet resources for their applications. In the design of the database some of the semantic relations between synsets can also occur between instances.

The possibility of incorporating proper names in the monolingual semantic network has been extensively discussed among the consortium and it was decided that some instances would be included in the individual WordNets but only to a limited extend just to indicate the possibility of incorporating such kind of information in a semantic database. After all there are so many instances that could be included in such a resource since they are used in everyday language that their inclusion would be beyond the objectives and scope of the project.

However, as language-neutral information the instances to be incorporated in the BalkaNet could be directly linked to the individual monolingual WordNets even though they could be linked to the ILI concepts as well, but this would cause many ILI maintenance problems. Finally, in cases where instances are difficult to the extracted from the given resources then they will be minimally added by hand to illustrate the possibilities of customization and extension.

- ✓ Glosses

Each meaning included in the multilingual WordNet needs to have a gloss that explains the sense, which will be stored in the interlingual index and will be in English. However, the option of providing monolingual glosses (original language glosses), in parallel with the English ones for the meanings included in each of the monolingual WordNets (extracted from reference dictionaries for the language in case) will also be considered and applied by the consortium, to the extent this is possible. This is expected to create difficulties in mapping the synsets, but on the other hand it should be representative of each language's specificities, which is among the project's primary goals.

In addition, the idea of including glosses of the respective languages apart from the English ones has been agreed among contractors and a unified formal structure has to be identified so that data is represented in a coherent way.

- ✓ Word forms and Senses

A word form in a semantic network (WordNet) is either the orthographic representation of an individual word or a string or individual words joined with underscore characters, called *compound* or *phrase* (e.g. acid_rain). An integer identification number should be attached to each word in order to distinguish between word forms that appear moiré than once in a lexical file but with different glosses attached since each occurrence of a word form indicates a different sense of the particular word. By using identifiers we will be able to uniquely identify each sense of a term within a lexical file.

A word in a synset is represented by its orthographic wordform and identification number, in addition, a local identifier is attached to each monolingual synsets indicating the source language of the term as described above (i.e. TR identifiers denotes that the synsets has been issues and developed within the Turkish WordNet). Together these items form a "sense key" that uniquely identifies each word/sense pair in the lexical database. Finally, a synset consists of synonymous words, relational pointers, a gloss, local and integer identifiers and example sentences if necessary. Lexicographers can enter comments in a lexical file of use of the synset, which might be discarded when the multilingual database will be built.

- ✓ Hierarchical links

From what has been discussed so far among the consortium the existing EuroWordNet hierarchical links will be adopted. With regard to the organization of synsets, the links in a WordNet are by definition bi-directional, meaning that if word-sense A is related to word-sense B, then word-sense B is related to word-sense A. Of course, the type of link relating the two word-senses is not necessarily identical in both cases (this is only the case for synonymy). It goes without saying that all word-senses should be inter-related in some way and to be attached to a global backbone, so that "orphan-meanings" are eliminated. Moreover, all relationship types need to be clearly defined and the number of items per level should be designed and implemented in such a way that navigation of the end user through the network is feasible and practically possible.

- ✓ Number and types of links (the language internal relations)

Following the type of data organization established by EWN, the average number of different links per meaning would be three:

- ❖ Monolingual, i.e. links of the meaning with other meanings of the individual WordNet)
- ❖ Multilingual, i.e. link from the monolingual WordNet to the ILI
- ❖ Domain link i.e. link with a domain label placed in the ILI

With regard to the type of links, the language internal relations will include not only relations within a part of speech (POS), but also relations across parts of speech (XPOS). The latter type of relation is expected to assist in handling morphology-induced relations between synsets, which are quite frequent in some of the participating languages, or in relating words with different syntactic behavior but comparable semantic content –an important asset from an IR point of view. Moreover, given the fact that each monolingual WordNet will be organized in levels, a consensus among the consortium regarding the number of items included per level is desirable, so that navigation through a thematic area (e.g. "plants") is possible for the end-user.

Since the EWN links will be adopted in this project as well, attributes will be allowed in the links to differentiate behaviors; e.g. conjunctive hyperonym and disjunctive hyperonym can be the same link with two different attributes.

In addition, the Czech partners have proposed the possibility of using the semantic verb classes worked out by B. Levin and a preliminary attempt has been conducted towards this direction, which resulted in a list of approximately 3.000 Czech verbs along with their WordNet 1,5 equivalents. The main motivation behind their attempt was the observation that the verbal H/H trees in WordNet 1,5 and EuroWordNet were mostly based on the troponymy relation.

Finally, with respect to the language internal relations it should be noted that two meanings couldn't share more than one link type i.e. a term being at the same time a hyponym and holonym of another term. Such rules could be used for consistency checking and quality assurance across the individual WordNets. However the final constraints that will be used among link types will strongly depend on the definitions of the links and the criteria applied while processing the resources and extracting the first subset of terms, a task that will be worked out later on.

- ✓ Properties assigned to meanings and monolingual synsets: the information that should be stored in each monolingual semantic network has to address the following:

  - A unique identification number must be attached to each meaning of a term. In addition part from the numeral ID an identifier will be attached to indicate the source language of the term. For example if a concepts is added by the Turkish partners in their monolingual WordNet it should hold the following identified: TR 12345566 denoting that this terms is a Turkish one and has one numeral ID also. Once all such terms are collected and investigated by the Czech partners in order to decide which of them are common for all the Balkan languages and thus they should be included in the Intern-Lingual-Index then it is going to be much easier for all contractors to check the underlying terms and decide on whether they are conceptually equivalent terms in their languages as well.

  - A Part-Of-Speech tag should accompany each term

  - Domain or sub-domain labels should be attached to each concept

  - Morphological or grammatical information should be present in cases where this is necessary to better describe the meaning of the underlying

concept or to express the relation of the concept in question with other concepts to which it is linked.

- A sense number for each POS tag should be attached
- A usage label should accompany each term in order to indicate whether it is a formal term, a dialect term, argot etc.

✓ EWN language internal relations adopted in BalkaNet

The language internal relations will be mostly based on the relations adopted in the Princeton WordNet 1.5 and the EuroWordNet projects. However, some major changes might come up during processing lexical resources, which will be thoroughly discussed in the respective Workpackage. Nevertheless, even though the project is at its early stages some processing of the resources has already been conducted and some problems have been already traced mainly due to the fact that the languages we are dealing with are extremely rich in morphology and thus have a high degree of morphological complexity.

Before describing the abovementioned problems it would be useful to briefly refer to and describe the language internal relations that have been used so far in various existing semantic networks (WordNets). The motivation for deciding to adopt already existing relations is twofold. First and foremost we wish to keep compatibility with existing WordNets in terms of structure of the lexical items and secondly due to the fact that most of the existing relations are capable of declaring in an efficient way semantic relationships between terms and concepts. However, main differentiations might come up due to language particularities and the quality and structure of the available lexical resources for the Balkan languages.

Finally, from an Information Retrieval (IR) point of view and with respect to the envisaged application of BalkaNet's results some of the already existing relations (e.g. synonymy, hyponymy and hypernymy) are rather useful for retrieval of information. Consequently, as already mentioned in another section of the present report the three-abovementioned language internal relations are going to be the common ones in all individual WordNets. If the need for the introduction of a new relationship comes up by one or more partners then such a possibility will be discussed and common, coherent decisions will be made. Nevertheless, it should be underlined at this point that relations across Part-Of-Speech (X_POS)[3] would be allowed in our network the same way X_POS relations were stored and represented in the EuroWordNet semantic network. This is due to the fact that we want to keep maximal compatibility with the EWN and secondly since from an information retrieval point of view the same information can be coded in an NP or on a sentence. Thus by using higher-order noun and verbs in the same ontology it will be feasible to match expressions with different syntactic structures but comparable content (ref. Shift project).

The three relations as used in EuroWordNet that will be adopted in BalkaNet as well are described below:

[3] X_POS relations are: noun-to-verb-hyponym, verb-to-noun-hyponym, noun-to-verb-synonym and verb-to-noun-synonym

- **Synonymy**: synonymy is by default the basic semantic relation that will be used in BalkaNet and has been used in all-semantic networks since the structure of synsets is based on synonymic relations. After all the advantage of the synset structure is that equivalent meanings of terms are explicitly encoded in the entry structure. A term is a synonym of another term if the former can replace the latter in any context without altering the meaning of the sentence. In EWN project a major distinction was made across synonymic relations which derived from the observation that even though terms sometimes hold the same meaning with others and thus can replace them in any context whereas others can replace them only in particular context or under particular circumstances (e.g. particular types of texts, speech etc.). Consequently, the two kinds of synonymy used in EWN are: exact synonymy and near-synonymy. Both synonymic relations are going to be used in BalkaNet and the second one is going to be extensively used while linking of terms from the monolingual WordNets to the ILI concepts.

- **Hyponymy / Hypernymy**: hyponymy is a fundamental relation around which the WordNets are constructed since hyponymy along with its complementary relation (i.e. hypernymy) link synsets as mixed conjunctive and disjunctive sets creating thus semantic chains in the lexical hierarchy. Hypernymy and hyponymy are inverse, asymmetric and transitive relations, i.e. if Y is a kind of X, then X is hypernym of Y and Y is a hyponym of X. An example taken from Greek language:

  E.g. "*ερπετό*" has hypernym "*ζώο*", ("*reptile"* has hypernym "*animal"*)
  "*ζώο*" has hyponym "*ερπετό*", ("*animal* has hyponym "*reptile"*)

  A hyponymy relation implies that the hypernym (the more general term) may substitute the hyponym (the more specific subtype) in a referential context but not the other way around. A referential context is a context where only the set of discourse entities is considered, whereas grammatical; register, pragmatic and other non-semantic properties of the considered words or context are neglected.

  Various lexicosyntaxctic patterns have been reported[4] for tracing hyponymic relations such as:

  - *such NP as {NP,} * {or/and} NP*
  - *NP {,NP} * {,} or other NP*
  - *NP {,NP} * {,} and other NP*
  - *NP {,} including {NP,} * {or/and} NP*
  - *NP {,} especially {NP,} * {or/and} NP*

  In order to elicit the implicational relation between the hyponym and the hypernym described above, different diagnostic tests with specific phrases were used such as the following:

[4] for further information with respect to the hyponymic patterns please refer to M.A Hearst "Automated Discovery of WordNet Relations" pp.134

Test 1: for Hyponymy-relation between nouns

- A/an X is a/an Y with certain properties
- It is a X and therefore also a Y
- If it is a X then it must be a Y

? the converse of any of the above sentences?

Test 2: for Hyponymy-relation between nouns denoting species and classes

- A/an X is a kind/species/race of Y(s)

? the converse of this sentence?

Even though so far there have been described only three basic language internal relations, however much more semantic relationships are going to be used for linking lexical items in the monolingual WordNets. Such relations will be specified in detail later on but some preliminary conclusions are also reported in other sections of this report.

- ✓ Structure of the network's entries: the entry structure of WordNets is rather different from the way traditional dictionaries tend to organize their entries according to orthographic forms of terms, POS distinction, morphological information etc. in WordNets entries are organized around the notion of a synset, which comprises of one or more word senses of a term which are considered to be identical in meaning. A gloss is attached to each synset which defines its meaning. In the BalkaNet the synset structure is going to be maintained since we do not only wish to keep maximal compatibility with the already existing WordNets but also due to the fact that the synset structure is the most suitable for the application of our project. The latter comes after the observation that the structure of a synset resembles much the structure of the tree and that all synsets are then linked with other concepts in a hierarchical way. This type of organization is going to facilitate checking and testing the data stored in monolingual WordNets whereas at the same time it will facilitate navigation of end users in the semantic network.

  Thus, in order to be able to organize our terms in a synset structure it will be necessary that lexical items extracted from dictionaries will have to be converted in a synset structure. To achieve that we need to determine which senses are going to function as synsets and which synonyms of the above senses are going to function as variants. More specifically, a particular sense of a term found in a dictionary has to be selected in order to be attached to the base form forming our synset. Once this sense is selected synonym terms to the underlying sense and not to the base form in general should be attached and linked to it by means of the synonymy relation, which is going to be the basic relation of our network.

  In addition, each synset is related to other synsets by one or more of the defined language internal relations whereas at the same time each synset will be linked to an English concept of the Inter-Lingual-Index. It should be noted here that linking of the synsets with the ILI concepts will be performed on a conceptual basis rather than on exact translational equivalencies in order to ensure that monolingual synsets of the Balkan languages linked to the same

English concept are exact conceptual equivalencies (i.e. are used with the same sense and in the same context) and not simple translations of terms. In order to achieve that terminologists, who will perform the linking of terms will have to study closely the glosses accompanying each English concept in order to reassure that the appropriate sense a concept of the monolingual WordNet is linked via synonymy to the respective concept in the ILI (i.e. is linked to the term holding exactly the same sense). It is clearly understood from the above that linking of terms might not always be easy since some concepts of the Balkan languages might be non-lexicalized English concepts. In such cases a complex ILI records might be created to express such a relation or new ILI records might be added in the Inter-Lingual-Index.

Moreover, each synset is going to have a "***belong to"*** link to each of the domain labels of the BalkaNet semantic network in order to indicate the specific domain or sub-domain to which it conceptually belongs. Thus, each variant apart from a gloss and an optional sage label that will or might have attached it is also going to be linked to one or more of the conceptual domains in order to facilitate navigation of the end user not only across different synsets but also across different concepts. In particular, the user will be able to view not only terms that are conceptual equivalents, i.e. terms that belong to the same synset but also he/she will be able to view terms that belong conceptually to the same domain (s). The latter is rather important for the performance of conceptual indexing or classification tasks.

The overall synset structure that has been used so far within the framework of the EuroWordNet project and will be adopted for the BalkaNet database is the one illustrated in the following scheme. However, some slight modifications from this structure might come up and will be defined at a later stage of the project.

	Obligatory	Optional	Optional	Optional	Non-Public	Optional
Synset ID	<record no.>					
Variants	**Variant**	**Usage Label**	**Central**	**Corpus Info**	**Source code**	**Status**
	file	---	Yes	Corpus <freq.> Corpus <freq>	WN1.5 sense 1 <record no.> sense... <record no.>	revised
	data file	formal	No	Corpus <freq.>	WN1.5 sense 1	not-revised
	Obligatory				Obligatory Non-Zero	
Language Internal Relations	**Relation Type**				**Related Synset**	
	Hyperonym Hyponym				<record no.>	
	Holonym Meronym				<record no.>	
Multilingual Relations	**Equivalence Type**				**Related English Synset**	
	Eq-Synonym Eq-Hyperonym Eq-Hyponym				<record no. for WN1.5synset)>	
	Eq-Holonym Eq-Meronym				<record no. for WN1.5synset)>	

Finally, in each monolingual entries apart from the information and structure that is going to be common there is also the possibility of incorporating additional information such as corpus frequencies of each variant, morphological information, usage labels etc. such kind of information is going to be optional and its incorporation is going to be decided at a later phase of the project.

Finally, one last thing that should be added and has been stated in a previous section of the present report is the fact that each synset (entry) is going to have an ID attached in order to avoid duplicates and reassure an efficient linking of terms. A local identified in each monolingual WordNet indicating the source language of the synset will accompany each ID. The latter is particularly important in cases where more that one contractor is involved in the development of the same monolingual WordNet (e.g. there are two Romania, two Bulgarian and two Greek partners actively involved in the development of the Romania, Bulgarian and Greek WordNets respectively.)

- ✓ Synset Structure

A common synset structure will be decided among the consortium. For some languages (e.g. Serbian) the inclusion of morpho-syntactic information in synset is vital. Although, the synset structure will much resemble the respective structure of the EuroWordNet synsets it has not been finalized yet since it will greatly depend on the VisDic tool that will be used for the actual development of synsets. Slight modifications to the EWN synset structure might come up, which however cannot reported at this stage since this is a task to follow.

- ✓ Common Language Internal Relations: the relations and the coverage to be represented in the BalkaNet semantic network are described having in mind what is required for the specific end users of the project's results and for the application of the BalkaNet database, given the state of the art in semantics (Princeton WordNets 1.5 and EuroWordNet), the quality of the lexical resources and what is feasible given the available resources, tools and time. When relations and coverage are discussed it will not necessarily imply that all these relations will be represented in the final multilingual database. Towards this direction the consortium has decided that a minimum set of relations has to be common and reflected in all monolingual lexical networks and the Inter-Lingual-Index. These relations are hypernymy, hyponymy and synonymy. Apart from having a coherent and common set of language internal relations among all WordNets another reason for deciding to represent the aforementioned relations in the final database was the fact that we wanted to achieve a minimun degree of compatibility with the EuroWordNet lexical database since once the project is finished an attempt will be made in order to unify both semantic networks in one common European WordNet. Finally, the reason for concluding on the abovementioned types of relations is that they are common across all languages and they can be extracted from the lexical resources that are already available to the consortium. Furthermore there is coherence among the linguistic community of what these relations stand for and what type of link they denote between two concepts.

  However, since we wish to develop WordNets that are representatives of the underlying languages and that include as many terms of the generic vocabulary as possible it is easily understood that in the monolingual WordNet the abovementioned relation types might not be always sufficient to denote the link among terms and that new relations will have to be introduced, which might be common with other languages or not. In light of the above the consortium decided that the three language internal relations i.e. synonymy, hypernymy and hyponymy are going to be present in all monolingual WordNets and in the Inter-Lingual-Index whereas new relation types will be included in the monolingual WordNets. In principle only these relations will be expressed between lexical items which are linguistically salient and which are extractable from the given lexical resources. The resources as such differ considerably in structure and content. We therefore cannot expect that the richness of the results is the same for every monolingual WordNet. The general approach is that if anticipated information is present in a given resource and if it easily extractable by semi-automatic means then it will be stored in the final multilingual database. In this respect the design will provide maximum flexibility in order to store semantic information without deviating too much from the EuroWordNet resource while keeping at the same time the monolingual networks individual and autonomous, that is without making too many commitments for building of the resources.

  Furthermore, during the implementation of the project and the actual development of synsets it might turn out that some of the language internal relations and data types are not practical for the purpose of the project or will hardly occur from the lexical resources. In addition, it might turn out that new relations might need to be added or that some of the EuroWordNet relations

might not need to be expressed but the idea of the functional specification is that all potential problems, aspects and relations are as much as possible anticipated.

- ✓ Possible Inclusion of Morphological Relations: for some of the participating languages with regular derivational phenomena (e.g. Turkish) it might be necessary that morphological relations be declared among terms. However, such relations will be defined and handled within each monolingual WordNet separately since they will be constructed on the basis of the underlying language phenomena. At this point it should be noted that such relations are between word forms and not between synsets as morphological processes mediate them. It is nevertheless possible to expend these relations to synsets should it be necessary. An example of such a productive relation for the Turkish language is the relationship between "*tas*" (stone) and "*taslas*" (to petrify). It is conceivable that the ILI entries for "*tas*" and "*taslas*" will link directly to the respective synsets represented by these but we will also have a bi-directional link from "*tas*" to "*taslas*" with a label *Become*.

- ✓ Inert-Lingual-Index (ILI records): the Inter-Lingual-Index is actually the main repository that links monolingual WordNets as much as possible. Having a language-neutral intermediary level will facilitate representation of language specific properties. One danger for a language-independent intermediary level is that there is no control on the status of this system when changes are made on the basis of distinctions in any of the languages involved. So far, ILI records of EWN were actually an unstructured list of English concepts along with their glosses with the only purpose to provide an efficient mapping across languages. The ILI started of as a list of WordNet 1.5 synsets and was restructured in order to reflect lexicalizations of the involved languages. Summarizing the EWN ILI is nothing more than a list of records with glosses that form the superset of all the concepts occurring in the separate WordNets. However, within the framework of the BalkaNet project it is anticipated that the ILI records as adopted by the EWN database will be restructured in order not only to reflect lexicalizations of the Balkan languages but also to be used as the basic element for performing conceptual indexing and classification of documents in IR systems. Thus, the main differentiation of the BalkaNet ILI in comparison to the EWN one concerns its structure. The main motivation behind structuring the ILI is the handling of mismatches between monolingual WordNets. One envisaged way of structuring the ILI is that the translation relations among terms in the monolingual WordNets and ILI records are simple but concepts in the ILI have the same semantic relations just as the synsets in the monolingual WordNets. By allowing relations between the ILI concepts we can add a new concept for any language specific synsets that is not present in the already existing records of the ILI. Thus, by linking the newly added concepts to the existing ILI-concepts the relation with the other WordNets can be established indirectly. The advantage in that case is that there can be a single and simple equivalence link for all the synsets with an ILI concept. In addition, there are going to be fewer relations but the ILI concepts themselves will be internally linked. Finally, another benefit of a structured ILI is that synsets that occur in two languages but are not lexicalized in English can somehow be matched when the WordNets are

compared. This would enable other WordNet developers to use this sunset(s) for linking. In addition, by allowing relations between ILI concepts we can add a new concept for any language specific synset that is not present in the ILI. By linking the newly added concept to the existing ILI-concepts (which are linked to other languages) the relation with the other WordNets can be established indirectly. The advantage in this case is that there can be a single and simple equivalence link fro all the synsets with an ILI concept. In any case, the structure of the ILI will be further discussed among the consortium and final decisions about this will be taken in a later stage of the project (as envisaged in the Technical Annex).

Apart from the structure of the Inter-Lingual-Index that has been described above the Interlingua is going to form the common intermediary between the consortium and the WordNets data were progress and status of the monolingual WordNets will be demonstrated. Due to the crucial role ILI plays in BalkaNet it is of great importance that it is maintained by a central authority, holding responsibility for updates, consistency checking, domain labels, restructuring, notification released etc. After some discussions the consortium decided that the Faculty of Informatics of the Masaryk University (Czech Republic) is going to hold responsibility of the ILI maintenance since they have a better knowledge of the ILI in comparison to the other contractors due to their participation in the implementation of the EuroWordNet project. The main reason for deciding on a central authority for the ILI maintenance is in order to overcome the danger that two sites are independently adding a new concept to the intermediary network without knowing from each other that they are adding the same concept. To limit this danger we have decided on a policy for changing the intermediary system after notifying the rest contractors and by simultaneously update the WordNets. Attention should be paid in such cases since added ILI records have to be English glossed.

- ✓ Top-Concepts and 3-order Entities: Top-Concepts of EuroWordNet are going to be adopted as the starting point for the development of the core monolingual WordNets. In particular, all members of the consortium agreed that the Top-Ontology and the Base Concepts of EWN are going to be incorporated into the BalkaNet semantic network once they are translated to the respective languages. Currently terminologists translate the EWN Base Concepts using bilingual dictionaries and aligned corpora where available. The Base Concepts will be the central entities around which the core monolingual WordNets will be developed. Once all monolingual Balkan lexical resources are processed the Base Concepts are going to be checked against the above resources in order to test which of them represent lexicalizations of the Balkan languages as well. In case some concepts, which are important and representative of the Balkan language are missing from EWN Base Concepts these will be added. In addition, if some of the already present Base Concepts might prove to be unnecessary for the data of the project they will be ignored. A top-down approach will be followed for the development of synsets corresponding to the first 1080 or so Base Concepts of EWN. This task will be performed in order to give the consortium a better understanding of the mechanisms required for populating the semantic network with new concepts. It will also help the contractors to identify any additional language internal relations that might be necessary for coding any

additional domain related and any kind of morphological information necessary for the association of synsets and lexical items.

The existing in the EWN ontological descriptions were thought in terms of a set of eclectic features extracted from various semantic theories. Modifying them would mean risking compatibility with EWN (which would prevent the extension of EWN with the Balkan WordNets) and spending too much time in trying to achieve consensus on the BalkaNet's ontological model.

Summarizing, by applying the abovementioned data and lexical requirements and by adopting the ontological hierarchies of the EuroWordNet project slightly modified would endure not only compatibility of your database with EWN but would also result in a much more flexible semantic network applicable to many NLP tasks and products. Based on the above any kind of changes (i.e. changes in the domains, the top-concepts and the instance level) will not have to be specified for each WordNet separately but only for the Inter-Lingual-Index records.

Resources that are already available to the consortium and are currently being processed

The Romanian partners have already started a series of quantitative analysis on a very large corpus of journalistic text, plus a few novels, collected from the Web. The corpus comprising of more than 100 million terms was automatically tagged, lemmatized and the content words of interest (nouns, verbs, adjectives and adverbs) were counted and sorted in terms of their frequency occurrence. Once the above process was finished a list of more than 30,000 Romanian lemmas was extracted. Based on the frequency in the running texts this list was divided in three individual segments, each one corresponding to 10,000 most frequent lemmas.

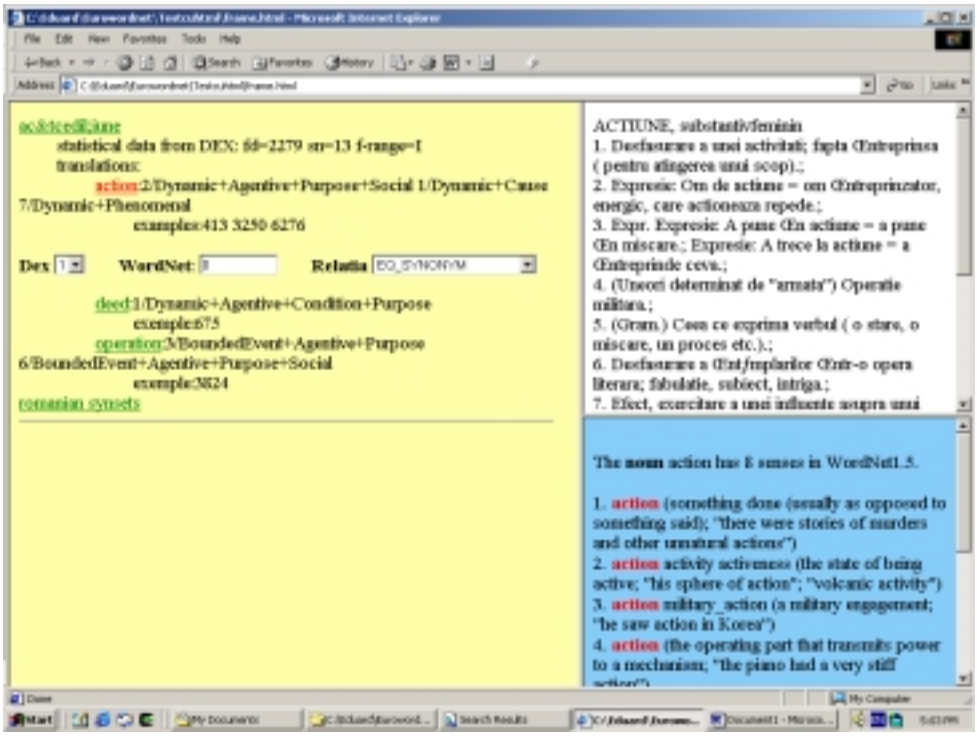
Apart from the corpus an Explanatory Dictionary of Romanian [DE, 1996] is going to be the most authoritative lexicographic resource out of which useful information will be derived for the development of the monolingual Romanian WordNet. The dictionary has been digitized and converted into a lexical database [XML encoded] by RACAI team. For the moment attention will be paid only to nouns. However, since the processing of terms doe not depend on their POS the same process will apply for terms belonging to other POS than nouns. So far 8,000 nouns and nominal compounds (accounting for almost 35,000 senses) have been extracted from the dictionary so that the definitional productivity DP (the number of sense definitions a noun participates in) was at least 3. The list was twice sorted according to the Polysemy degree-PD (the number of senses listed in our dictionary) and according to the definitional productivity. The lists obtained strongly correlated and the conclusion drawn was that both criteria are equally informative and thus received equal weights.

A Romanian Dictionary of Synonyms (RDS) is also available, which is digitized encoded in as an ACCESS database, which is being used for the extraction of synonymic series for the selected Romanian terms. Some of the terms included in the dictionary have usage information attached (e.g. old, regionalism, domain-usage etc.). However, since the main objective of BalkaNet is to develop a semantic dictionary of the generic vocabulary of the languages involved such terms as the ones described above will be used only to extend when their presence is necessary in the network in order to reflect lexicalizations of the underlying languages.

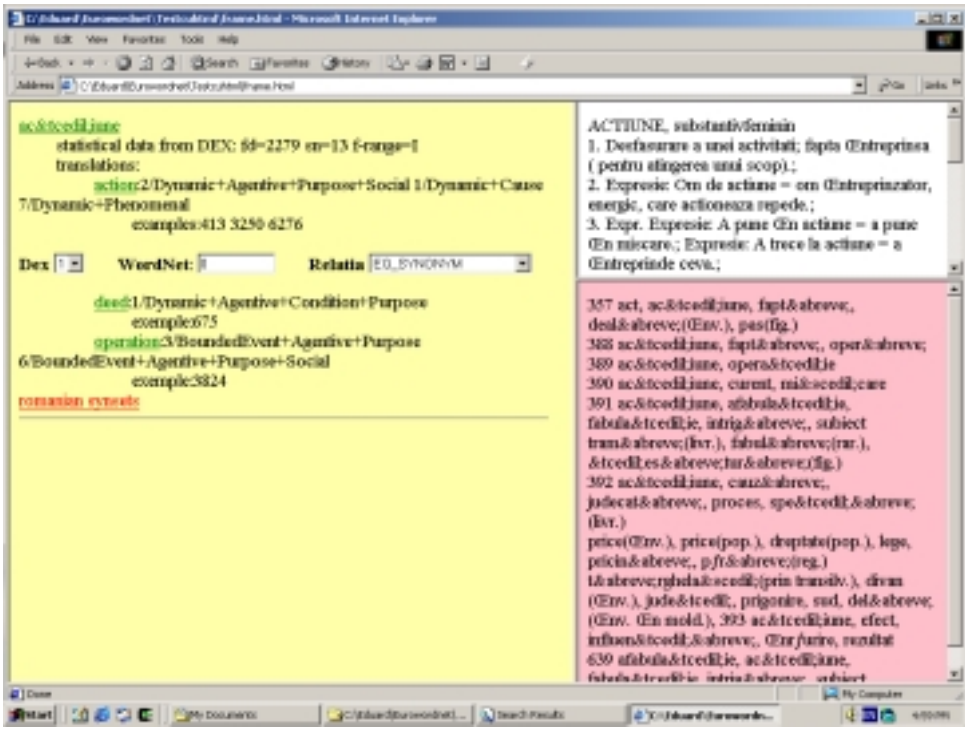
As a result of processing the abovementioned Romanian lexical resources a large html file occurred comprising the following information:

- The number of sense definitions from the Explanatory Dictionary of Romanian
- The translations in English. At this point it should be noted that for these English nouns that are in the base concepts the ontological description is given as provided by WordNet 1,5.
- Association of a sense definition number of the Romanian with the best possible mapping to WordNet 1,5 synsets. It should be noted here that during this stage it is ensured that for any English synset there is a Romanian gloss present

- Assignment of a Romanian synset to the definitions that have been associated in the previous steps



Romanian term and WordNet 1,5 synsets for one of its translations



Romanian term and its synsets

The two screenshots provide an illustration of the html file that has developed so far. The proper management of the ontological structures, as well as their XML encoding will be achieved using the VisDic editor provided by the Czech partners.

For the Serbian language a one-volume explanatory dictionary is available in electronic form comprising about 50,000 entries. Each entry is accompanied with morphological information (POS tags, grammatical information etc.). However, few modifications need to be done such as addition of the number of senses as the latter extracted from printed dictionaries, incorporation of the definitions (description of the senses), addition of register information, addition of the number of senses as the latter will be extracted from a 6-volume Serbian dictionary in printed format.

Apart from the dictionaries a corpus of contemporary Serbian is also available to the consortium comprising of newspaper texts, literature, manuals etc. The corpus however has not been processed yet. With respect to corpora a corpus of parallel texts with Serbian as the source language and English and French as the target languages is also available. The abovementioned texts have not been aligned yet and this is a task that will be performed during the implementation of the project. Finally, as far as the Serbian lexical resources are concerned a Dictionary of Synonyms and a Dictionary of Verb Valences are also available in printed form which will have to be processed and converted to electronic form prior to be directly applied for the project. An alternative solution would be the manual processing of the resources during the expansion of the monolingual Serbian WordNet.

The Czech partners have already developed a monolingual WordNet comprising of approximately 20.000 synsets since they participated in the EWN project. At present they are performing all necessary checks and testing to their network in order to trace errors and correct them. Attention is being paid on the following errors:

- Inconsistent synsets due to false translations. Checking of inconsistencies will be performed mostly manually
- Errors in hyper/hyponymy trees due to the adoption of WordNet 1,5 trees.

Three phenomena identified during the development of Turkish base concepts

- Pseudo base concepts:

This problem emerged when we attempted to find a Turkish equivalent for the Euro WordNet base concept (“BC”) of "condiment". First level hyponyms of "condiment" include "dip", "mustard", "ketchup", "hollandaise", "garlic sauce" etc. There exists no Turkish equivalent for the English BC "condiment" but the hyponyms constitute a fairly well defined and coherent group of words for the Turkish native speaker. Thus we have the hyponyms (and also the hyperonyms), but the concept that “holds them together” is not lexicalized in Turkish.

What we suggest is to define a standard type of entry such as "N/L_TR" (not lexicalized in Turkish, although hyperonyms and hyponyms are a well-defined and meaningful set). These entries will practically be “pseudo-entries” with no associated word or collocation and will probably include only a gloss. The inclusion of such pseudo-entries has important implications for possible future applications. For example, the existence of a pseudo-entry for "condiment" in the Turkish WordNet will link the English word "condiment" to a group of Turkish words and collocations that are the hyponyms of the pseudo-entry. This will allow the editor of an English-Turkish bilingual dictionary, for example, to have access to the Turkish hyponyms

and hyperonyms corresponding to the English entry "condiment", and these hyponyms and hyperonyms will definitely be a part of the definition of "condiment" in Turkish.

- Base concepts having improper connotations

The English BC "ABODE" and the Turkish counterpart "BARINAK" were the source of this problem. In general, this problem involves cases where a neutral word has acquired a certain connotation. In this case, giving the “loaded word” the status of a base concept (under which tens of hyponyms will be listed) becomes inappropriate. For instance, the English BC "ABODE" could have "PALACE" as its hyponym (although this is not the case in Wordnet 1.5 for some reason), while it is not possible to claim that the Turkish word "SARAY" (English "PALACE") is a hyponym of "BARINAK" (English "ABODE"), since the unfortunate candidate BC “BARINAK” has a connotation that implies a poor, simple and small lodging.

In fact, Problem 2 is a weaker version of Problem 1. In both cases, the base concept we are trying to define is the hyperonym of a well-defined, coherent set of hyponyms in Turkish. In Problem 1 there exists not a single candidate, while in Problem 2, the best candidate is an unfortunate one with undesired connotations.

Two possible suggestions are: a) Avoid the “loaded word” and put a “pseudo-base-concept” as defined above; or b) Use the “loaded word” as a BC but include a note that the connotation is to be disregarded.

- Two English BC’s merging into a single BC in Turkish

It is a well-known phenomenon that two or more Base Concepts (BC) in a language merge frequently into a single Base Concept in another language. An example in the Turkish case is the English BC’s “engine” and “motor” and the Turkish BC “motor”. There is a fine semantic distinction between “engine” and “motor” (or at least WordNet 1.5 claims so). This distinction does not exist in Turkish in any way and the only Turkish word that can be used in this context is “motor”. There also exists no justification for defining two senses “motor 1” and “motor 2” in Turkish, since the distinction in English is non-existent in the minds of the average Turkish native speaker (which we think is an important consideration while building the WordNets).

The existence of a finer distinction in another language might improve the value of the Turkish WordNet. Thus, we think that it could be beneficial if the Turkish BC “motor” somehow contains the information that it corresponds to two distinct English BC’s (namely “engine” and “motor”). For example, we could define two distinct “second-level senses” for the Turkish BC “motor” and denote these as “motor 1*” and “motor 2*”. The asterisk is intended to show that this “sense” is not a widely accepted and well-known sense of the entry, but the distinction has been made to take advantage of finer distinctions existing in other languages.

At this stage, we are unable to provide further justification for our suggestion to define “second-level senses”, and would only like to present the issue to your opinion.

Problems Encountered during Selection Process of the Greek Base Concepts

❖ Deciding on a special case of synonyms

Regarding the synonymy relation there is a number of words conventionally referred to as synonyms in Modern Greek dictionaries, which nevertheless seem to need further checking before being used as such in the Greek WordNet.

Standard Modern Greek permits a good deal of vocabulary variation due to the former linguistic situation of bilingualism[5] (diglossia). Word selection varies in such cases according to whether a term is used in spoken or written context or on the types of documents it is found in (i.e. literature texts still tend to make extensive usage of wordforms that do not have all their inflections expressed according to the Modern Greek inflectional system).

The problem has emerged when we possessed lexical resources form tracing the Greek Base Concepts where we encountered the phenomenon that terms listed in dictionaries as exact (direct) synonyms seemed to be related with another kind of synonymy. According to Miller et al. "two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value of it". But what happens when this is a matter of using different registers or of speaking/writing in different pragmatic situations?

On top of that there are a few cases in Greek where terms are widely known and used for their derivational function. In such cases deciding on whether they should be included in a semantic network involves extensive usage of various corpora covering various fields of speech (both spoken and written). However, different kinds of corpora might provide us with different kinds of results, which actually reflect the present state of the language.

At this early stage of the project what we suggest is that the monolingual WordNet of Standard Modern Greek includes, alongside the range of Modern Greek literals elements productive in the present-day language, spoken or written even if they have slightly differentiated inflectional forms from the valid ones. Nevertheless, the kind of synonymy-relation that should link this kind of words is being denoted so far as synonymy* in order to be differentiated from the synonymy relations used in the EWN. We expect future research to shed light on this matter and once various tests on the resources and the terminology extracted are performed we will be able to report more concrete results. For the moment we focus on the language internal relations already mentioned previously since these are going to form the basis for the development of the Greek core WordNet.

[5] The phenomenon of bilingualism refers to two distinct dialects used for quite a long period in Greece which were both widely used with the only differentiation that one dialect used many forms of the Ancient Greek and was through as a more formal one whereas the second one used more forms of Modern Greek and was mostly used in spoken speech.

MULTILINGUAL ARCHITECTURAL REQUIREMENTS

Multilingual architectural requirements should contribute to the easy integration of distributely-developed segments. It is extremely important that the application of the methodology is actually based on the distributed processing and simple integration, having always in mind the relations between words and their senses that resemble those supported by object model. UML, use-case method and diagrams could be useful tools for definition of the architectural requirements.

With respect to the multilingual architectural requirements the following processes should be supported:

- Interlingual mapping and correspondence
- Inheritance of properties and links
- Traversal of links
- Equivalence links
- ***Belong to*** links to the respective conceptual domain labels
- Representation and visualization of the relations
- Integration of tools already available for WordNet construction. The individual WordNets have to be integrated in the final database without any prior structural changes required. Moreover, the final database has to be easily integrated in any product without any changes.
- Integration of the VisDic tool in the final multilingual database
- Efficient querying of WordNets and selection of specific relations
- Traversal of relations between and across WordNets
- Simultaneous view of linked WordNets for two languages with and without the ILI intermediary
- Accurate and clear documentation has to be provided to end users for the use of the final multilingual BalkaNet database.

All the abovementioned requirements issues by developers of the project have to be met in order for the final resource to be easily applicable to any kind of task. Moreover, by meeting the above objectives developers will be able to perform checking tests on the database and further enhance it or enrich it in the future.

The remaining parts of the deliverable give a more detailed overview of the aforementioned criteria and explain who these are met in the methodology that will be followed for the implementation of the project.

Specifications of the VisDic Tool[6]

Shortcomings of the Polaris tool led us to the idea to build another WordNet instrument that can meet requirements of lexicographers and dictionary users in a better way. In January 2001, the project of Natural Language Processing Laboratory at Faculty of Informatics at Masaryk University in Brno called VisDic was started. The abbreviation VisDic means "Visual Dictionary". The main task of this graphical application is to view and edit any lexical data. There is only one restriction: dictionaries must be stored in XML format. This format is well suitable for representing any dictionary entries. It is quite readable, and moreover it can be considered as a standard data format commonly used by many different applications.

In VisDic a user can view and edit up to 10 dictionaries. It can open more copies of one dictionary (useful for WordNet tasks). Every database has its own window frame with query text box, list box containing query results and the window presenting specific view of found entries.

There are different types of view in VisDic. The common one is the text view. User can specify colors, fonts, indentation and profile of every entry part. Another view is naturally the XML form of the entry source. Special type is a tree view arranging more entries to the tree specified by parent and children relation. This appears to be very useful for browsing WordNet hypero-hyponymical (further H/H) trees. Edit view can be also defined by the user and its purpose is to change actual entry (synset) in a dictionary. Finally, list of words accesses all the literals found in the dictionary. It is also helpful for systematic work during the data editing.

Dictionaries can be linked. Any data stored in one database can be accessed from another one. For example, not every WordNet has its own synset descriptions – so called glosses. Then it is useful to use English descriptions. Because these texts are big in relation to the size of the WordNets that contain only literals, it is useful to have glosses stored at one place – in the ILI database. Other WordNets then can access the synset descriptions by a link to the ILI dictionary. Moreover, if the description is modified for some reason, the change will be immediately displayed in other WordNets.

As any common databases, XML databases in VisDic have their keys – unique attributes which identify every entry. WordNet has also its key – ILI number. This number is in the "NNNNNNNN-X" form, where N are digits and X is a part of speech (n = noun, v = verb, a = adjective, b = adverb). During WordNet conversion from Polaris import-export format, this key was taken from EQ_SYNONYM external relation.

The active (opened) dictionaries in VisDic can cooperate by means of keys. A user can grab one entry, drag it to another dictionary and drop it there. If the target dictionary contains an entry with the same key value as the source dictionary, the corresponding data are found.

VisDic has also some functions implemented for WordNet users. It is able to find the topmost entries in a tree. In a H/H tree, result will consist of synsets not having any hypernym. In this way inconsistencies in a database can be easily displayed. Other functions can fully expand a tree and count number of synsets under the specified

[6] The VisDic tool is being developed by the Czech partners (FI MU) with the overall supervision of Prof. Pala.

node. The "Hang under" function can automatically link actual synset to another synset in another dictionary. Finally "Tree copy" function can copy the whole tree from one dictionary to another. For example, during development of a new WordNet, the specified tree from English WordNet can be easily copied. Then it is necessary to translate literals only because all the relations will be preserved.

VisDic is also a highly configurable tool. Most of the VisDic's behavior can be defined in the configuration files. VisDic uses two types of configuration files. One type specifies global information such as fonts, colors, active dictionaries, existing dictionaries, etc. The second type is specific for every dictionary. It holds its name, key tag, definition of views, etc.

At the present moment, VisDic contains all the EuroWordNet databases, ILI database, SSJČ Dictionary (Slovník spisovného jazyka českého), SČS (Slovník českých synonym), Collins Cobuild dictionary and the example of XML formed English corpus.

The application is developed under Linux X-Windows system with GTK support. However, the program code is platform independent and can be re-compiled in any other operating system using GTK libraries, e.g. Windows GTK libraries are accessible on the Internet.

The data format of VisDic XML representation is enclosed into XML tags. Only synonyms and its sense number are fully defined. Internal language relations and external relations are understood as the links to other synsets. These links are represented by the key value, which is the ILI number. Glosses are the external tags, as mentioned above, automatically extracted from ILI database according to an ILI tag.

Example (you can compare it with the previous import-export format):

```
<SYNSET>
  <POS>n</POS>
  <SYNONYM>
    <LITERAL>food
      <SENSE>1</SENSE>
    </LITERAL>
    <LITERAL>nutrient
      <SENSE>1</SENSE>
    </LITERAL>
  </SYNONYM>
  <ILI>00011263-n</ILI>
  <HYPERONYM>00010368-n</HYPERONYM>
  <GLOSS>any substance that can be metabolized
    by an organism to give energy and build tissue
    &03  1stOrderEntity  Comestible  Form  Function  Object  Origin
Substance
  </GLOSS>
  <HYPONYM>00848471-n</HYPONYM>
  <HYPONYM>04830107-n</HYPONYM>
  <HYPONYM>04830190-n</HYPONYM>
  <HYPONYM>04830627-n</HYPONYM>
  <HYPONYM>04834499-n</HYPONYM>
  <HYPONYM>04837708-n</HYPONYM>
  <HYPONYM>04838303-n</HYPONYM>
  <HYPONYM>04838667-n</HYPONYM>
  <HYPONYM>05011422-n</HYPONYM>
  <HYPONYM>05057848-n</HYPONYM>
```

```
  <HYPONYM>05074818-n</HYPONYM>
</SYNSET>
```

Apart from VisDic, there are some other applications, smaller ones that help to maintain dictionaries in general. These tools are able to convert dictionaries between VisDic and XML representations, query dictionaries, find some types of errors in WordNet databases, obtain dictionary statistics, etc.

Communication and Data Exchange

This section deals with the communication problems during WordNet's development.

WordNet databases can be shared by several ways:

1. Every group works independently with their own data. The supervisor who is responsible for the task will synchronize time after time all the work. This way was practiced during EuroWordNet 1,2 development. A disadvantage of this processing is the fact, that all the progress made by one of teams cannot be checked and controlled with other group. This is guaranteed only at the special time point, at which supervisor checks all WordNets and sends results to all participants. This work is also time consuming for the supervisor.
2. Every group works with data stored on the server. Every request is handled immediately. The advantage is obvious at the first sight - the development can be fully synchronized at any time. But there are other arguable points. There is no difference between the current version and final version. Is this a good feature? Certainly not. Imagine that more people are working on one part of WordNet (e.g. special sub tree). One of them systematically adds synsets under one node. During this work the second one browses these synsets, and decides that this node must have more synsets than he actually found and he wants to append them. Obviously, the final work will not be consistent, because one synset can be added twice or the same literal in similar synsets can be duplicated. Moreover, working in this way requires from the users to be constantly on-line which can be quite limiting.
3. Every group works independently with their own data. Every team is responsible for synchronizing data by sending them to the server in the format of what is called a journal. Journals are the text files, which describe how the final database has to be modified - which synsets are added, deleted or updated. In this way it differs from the first one with regard to the supervisor role. Server, not supervisor, is responsible for the automatic control of received data. Thus WordNet can be updated any time. Inconsistencies described in the second way can be prevented by editing the journal file before sending it to the server. Moreover, journal files can be concatenated or sorted. Therefore its parts can be created individually by every participant and separately. The need for on-line presence is no longer relevant.

Taking into consideration the above we conclude that he best way to access data is the compromise between the first and second way, which we have presented as the third way.

The second problem we have to deal with is how to control consistency and what is really necessary to control. There are four types of WordNet synchronization that can be controlled automatically:

1. Invalid links - some synset relations of the specific synset can point to the non-existing synset. These errors are prevented when the WordNet is edited by VisDic application, but after making manual changes in the journal files, these errors can occur.
2. Links to the same synset - the synset relation points to the same synset. Moreover, the hyperonymical or hyponymical relation can make loop which means that the sequence of the synsets joined by this relation will never end.
3. Duplicate literals - WordNet can contain more than one definition of the literal in the specified sense number.
4. Global consistency of trees - synsets have to meet this condition: if A is hyperonym of B in a WordNet then A is hyperonym of B in every WordNet. This is a very strict condition, therefore the parts of hypero-hyponymical trees, which do not meet it, are not treated as an error then they are only logged.

We are able to develop the described tool. The checking of some inconsistencies is now implemented not as a component of the VisDic application, but as the part of conversion from Polaris import-export format to XML format. It is feasible to include these functions to the final application, but it would better to implement the checks in such a way they will be applied after updating WordNets by means of journal files.

DIFFERENTIATION OF THE ARCHITECTURE REQUIREMENTS COMPARED TO THE EWN

Project Engineering Process

1. Software Engineering Methodology

Towards the better completion of the BalkaNet Project there is a need for a methodology to be followed so that we can organize and model the behaviors of the overall system. The characteristics that differentiate a methodology have to do with its ability to predict and, consequently, avoid possible mistakes during the implementation, so that give correct and simultaneously not time-consuming directions and generally to explicitly describe the steps that need to be taken.
According to the afore-mentioned, the methodology that mostly satisfies these requirements and we are going to use it in the BalkaNet Project is very close to the ICONIX Unified Object Modeling Approach. This software engineering methodology is based on the UML notation framework and is proposed by D. Rosenberg on his book *"Use Case Driven Object Modeling with UML" (Addison-Wesley)*. It introduces 5 phases in a software engineering process:

- Domain Modeling
- Use Case Modeling
- Robustness Analysis
- Interaction Modeling
- Detailed Design

The reason why we decided to follow this method derives from the fact that it is very flexible and has worked perfectly when being tested in numerous real environments. Furthermore, it offers three very important features:

1. It is **iterative** and **incremental**. Multiple iterations occur between developing the domain model and identifying and analysing the use cases. Other iterations exist, as well, as the team proceeds through the life cycle. The static model (composed of class diagram) gets refined incrementally during the successive iterations through the dynamic model (composed of use cases, robustness analysis and sequence diagrams). So Apart from the sequential order of the fore-mention steps there is always the ability to go back and iterate a step, updating this way its outcome and documentation.

2. A high degree of **traceability.** At every step along the way, reference back to the requirements in some way. There is never a point at which the process allows to stray too far from the user's needs. Traceability also refers to the ability of tracking objects from step to step as analysis melds into design.

3. **Streamlined usage of UML**, which means that it provides a structural course of building the uml diagrams, from the beginning of a project down to its implementation stage.

1.1 Methodology Approach

As we mentioned above our process is inspired by the ICONIX method. We only change the order of the different steps.

Building a first sample of the Domain (Static) Model first, does not seem very appropriate. For that reason we split the Use Case Modeling phase into two steps. The first one involves the creation of General Use Case Diagram(s) and is applied before the Domain Modeling with the aim to help capturing the static nature of the system with very general terms. The second step follows the Domain Modeling phase and has to do with the Detailed Use Case Diagram(s).
The steps that are going to be followed are presented here in a graphical way and then detailed later in this document. The Coding part has been introduced here only to show that someone can start to code only when he has made all the project engineering related documents. As we can see, the outcome of this process is the Detailed Design, which is the base document for the Coding part.

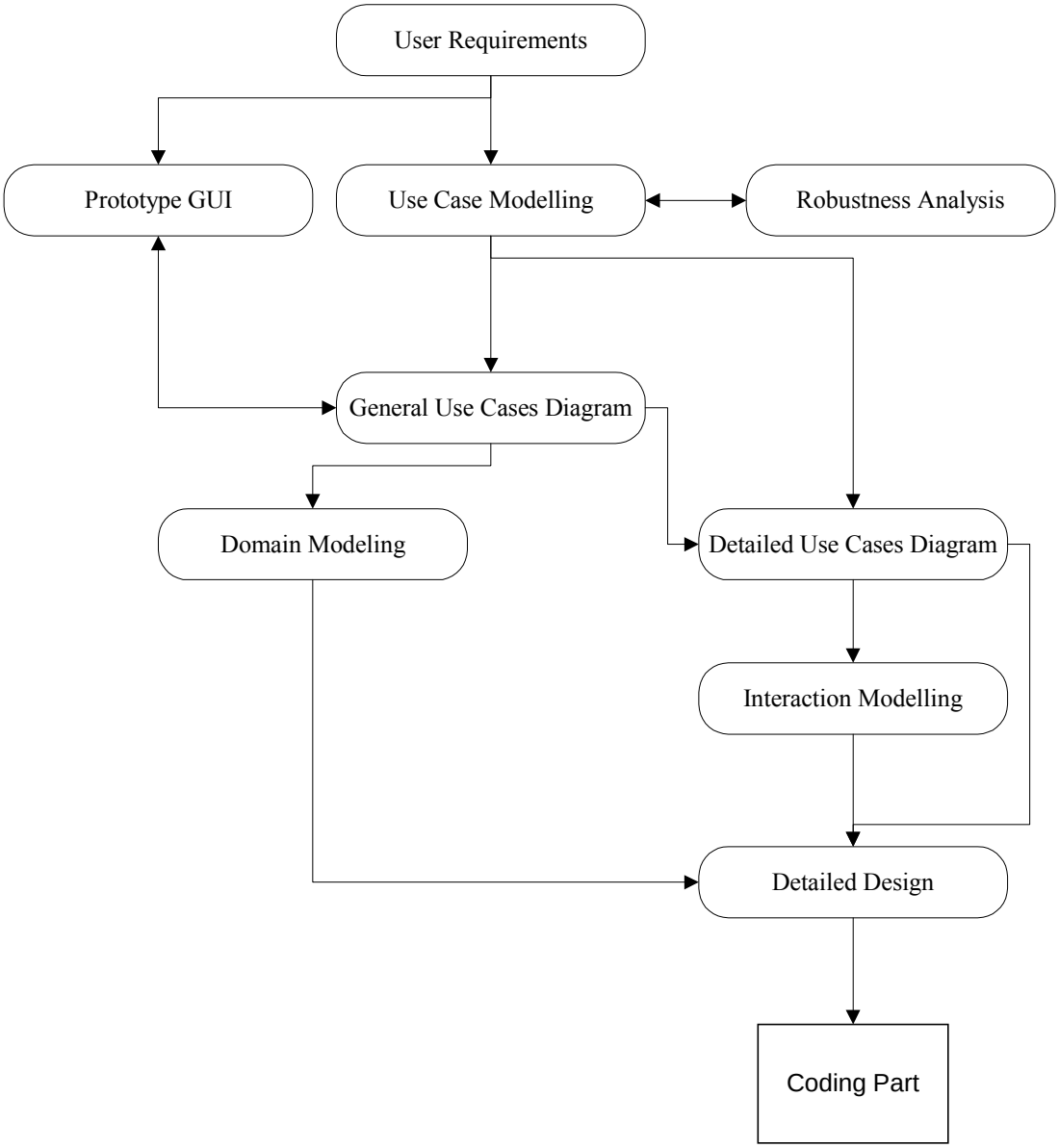


Figure 1. Suggested methodology to be followed.

Of course, at this stage of the BalkaNet project we can not be certain that we are going to stick strictly neither to the specified order of the documents nor even to the mere existence of some of them. However, we will begin by following the steps described above and any necessary changes or modifications needed will be made after considering the feedback given from the partners.

1.1.1 Detailed presentation of the intermediate steps

For each step, the result document produced is presented:
• *its role*: why should I build this document?

• *what does it need*: presents the documents that are needed to build this document.

• *what does it help*: presents the documents that this document helps to build.

1.1.1.1 User Requirements

• **Role of the User Requirements:**
They present user-defined criterion that the system must satisfy and must be the most complete possible in order to facilitate the creation of the next documents.

• **What does it help:**
It helps creating the Prototype GUI as it contains what the system has to do, and even sometimes constraints on how to do it.
It helps building the General Use Case Diagram as it contains the actions that the user will make on the application.

1.1.1.2 Prototype GUI

• **Role of the Prototype GUI:**
The prototype GUI allows the client to see what the system will look like. This is only a graphical object; there is no code behind it.

• **What does it need:**
The User Requirements give the necessary information to build the Prototype GUI as it presents exactly what the system has to do.

• **What does it help:**
It is built at the same time as the General Use Case Diagram because with this GUI you can easily find the possible actions for the user.

1.1.1.3 Use case Modeling

The essence of the Use Case Model is to capture and/or structure user requirements of a new system, by identifying all the big functional blocks (‘use cases’) and by detailing all the scenario’s that users (‘actors’) will be performing (within these big functional blocks).The documentation of this model is made by the Use Case Diagrams, which show the relationship among use cases within a system or other semantic entity and their actors. We build two types of Use Case Diagrams, a general and a more detailed one. After this phase, we organize, if necessary, the use cases into Package Diagrams.

General Use Case Diagram

• Role of the General Use Case Diagram:

The General Use Case Diagram presents both the actions that the user does on the system and the system reactions. In fact, it present more the "what" than the "how", describing what the system will do at a high level, with the a user focus for the purpose of scoping the project and giving the application some structure.
Sometimes it can also help to refine the User Requirements as new questions can raise when creating the use cases.

• What does it need:

The User Requirements give enough information to build the use cases as you can easily extract the actions on the system from it.

• What does it help:

It is built at the same time as the Prototype GUI because it is easier to refine it with these use cases.
It is very useful to build the Domain Model as you can extract the objects from the use cases and place them in relationship in the Domain model.
It is also the base document for the Detailed Use Case Diagram which is a detailed version of the General Use Cases Diagram.

Detailed Use Case Diagram

• Role of the Detailed Use Case Diagram:

The Detailed Use Case Diagram is the General Use Case Diagram in a detailed version. In fact, it presents more the "how" than the "what". Furthermore unlike General Use Case Diagram, it presents every basic courses of action (the main start-to-end path the user will follow under normal circumstances) along with a possible set of alternate courses of action (infrequently used path, such as exceptions, error conditions, etc).
Sometimes it can also help to refine the General Use Case Diagram as new questions can raise when creating the use cases.

• What does it need:

It needs the General Use Case Diagram, as it is a detailed version of it. It also needs the User Requirements to get the necessary information.

• What does it help:

The General Use Case Diagram helps building the Detailed Design as someone can pick up the actions that become methods in the Detailed Design.
It also helps creating the Sequence Diagrams as someone can pick up the order of the actions from it.

1.1.1.4 Domain Modeling

• Role of the Domain Modeling:

Domain Modeling is the task of discovering "objects" (classes) that represent real-world things and concepts, as well as relationships between them. The Domain Model is a static model that shows the relationship between the objects of the application, e.g. captures the static structure of the system. The Class Diagram(s) consists the documentation of this stage. Each class can have attributes but no methods yet.
Building a good Domain Model is important because it is the base of the Detailed Design, which is the last document before coding. Sometimes it can also help to refine the General Use Cases Diagram.

• **What does it need:**

The General Use Case Diagram is the document from which someone can extract the objects to put in the Domain Model.
Sometimes it can also help to refine the User Requirements as new questions can raise while creating the Domain Model.

• **What does it help:**

The Domain model can help to build the Detailed Use Case Diagram (but it's not its role) in conjunction with the General Use Cases Diagram.
It helps to build the Detailed Design as they are the same document, one is only more detailed than the other one.

1.1.1.5 Robustness Analysis

• **Role of the Robustness Analysis:**

Robustness Analysis is not a separate step in modeling. It is an integral part of the use cases writing. It involves analyzing the narrative text of each use case and identifying a first-guess set of objects that will participate in the use case. Afterwards it updates the domain model with these objects.
This kind of analysis is the link between the analysis (what) and the design (how).
Key roles of robustness analysis
1. Sanity check: it helps to make sure that the use case text used is correct and that it does not specify system behavior that is unreasonable (or impossible).
2. Completeness check: it helps to be sure that the use cases match all the alternate course of action.
3. Object identification: we may have missed some objects
4. Preliminary design: robustness diagrams are less complex and easy to read than sequence diagrams.

What does it help:

It helps to finish the analysis Class Diagram and so it helps to build the Detailed Design which is a detailed version of the Domain Model.

1.1.1.6 Interaction Modeling

• **Role of the Interaction Modeling:**

Once finished with Domain Modeling and the Robustness Analysis someone has uncovered most of the problem space objects and assigned some attributes to them. He has defines high-level static relationship (Domain Model) and a few dynamic relationships (Robustness Analysis).

Interaction Modeling is the phase in which someone builds the threads that weave the objects together, enabling this way to start seeing how the system will perform useful behavior. Someone may think of Interaction Modeling as representing the changing of the guard between analysis and design.
The goal of Interaction Modeling is to allocate behavior. That is, for each use case it identifies the messages between different objects (Sequence Diagrams). If needed a Collaboration Diagram is also designed in order to show key transactions between objects, as well as a State Diagram or/and Activity Diagram to show real-time behavior.
Since this phase presents the interactions between the objects of the application, it allows someone to extract the methods of these objects. Sometimes it can also help to refine the Detailed Use Case Diagram.

• **What does it need:**
It needs the Detailed Use Case Diagram as it presents the interactions between the objects of the application.

• **What does it help:**
It helps creating the Detailed Design as its presents the interactions between the objects.

1.1.1.7 Detailed Design

• **Role of the Detailed Design:**

The Detailed Design is in fact a detailed version of the Domain Model that contains now the methods of the objects. It matches exactly the classes someone is going to produce in the coding part and so it finishes the static model. Logically is the last part of the project engineering process before the coding part. Once someone has made it, the coding part is much easier as he only has to write the code and not to think about the organization of the classes. If needed someone can also produce Deployment and Component Diagrams at this step, in order to help him during the implementation phase.
Sometimes the Detailed Design can also help to refine the Domain Model.

• **What does it need:**

The Detailed Design needs the Domain Model, as it is a detailed version of it. It then needs the Detailed Use Case Diagram and the Diagram(s) produced by the interaction-modeling phase, as it is in them that you can extract the methods of the objects.

• **What does it help:**

The Detailed Design is the base document for the coding part.

2. User Requirements

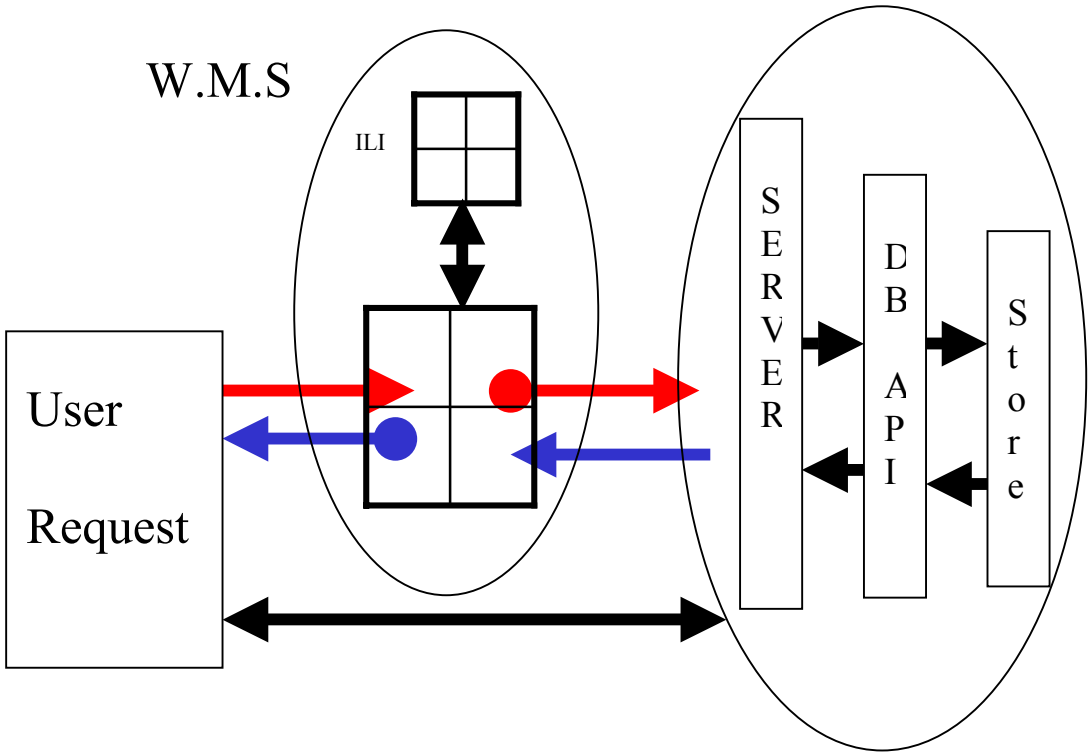
The WordNet Management System (W.M.S.) is an Open Language Engineering System that provides
–Common protocol for federated requests
–Stand alone access
–Connectivity among services through protocol

Federated requests

The user requests should be based on agreed protocol that provides interaction between user and the WordNet Management System. The left hand side circle of the following figure represents the W.M.S., which incorporates the I.L.I structure (see: WordNet Vs BalkaNet). The right hand side represents each partner. It consists of the store (Core WordNet) and the shells above it. DB API is responsible for DB management. The Database may be whatever the partner wants to be (commercial rdbms, file system etc). The server layer is responsible to communicate with the W.M.S. due to defined protocol.

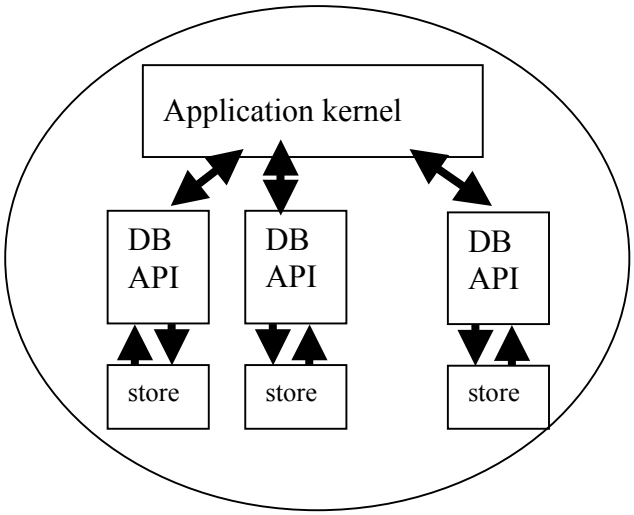
Let us say that user requests the synonym tree of the Greek word run [red arrow on the left hand side]. The W.M.S. knows how to find the Greek Core WordNet [red arrow on the right hand side]. The W.M.S. asks the Greek Server and the request is being forwarded to the database api layer that is responsible for the extraction of the above request. Then, the response is forwarded back to the Greek Server and the Greek Server sends it back to the W.M.S. [blue arrow on the right hand side]. Then, the answer is being forwarded to the user [blue arrow on the left hand side].

The problem that arises with the abovementioned architecture is that if many user requests arrive, the W.M.S. should be able to service all of them in a reasonable time. Since the blue arrows carry data to be shown to user the W.M.S. end up with performance issues. The solution to the abovementioned problem is to return to user part of the answer needed to establish a peer-to-peer communication [black arrow].



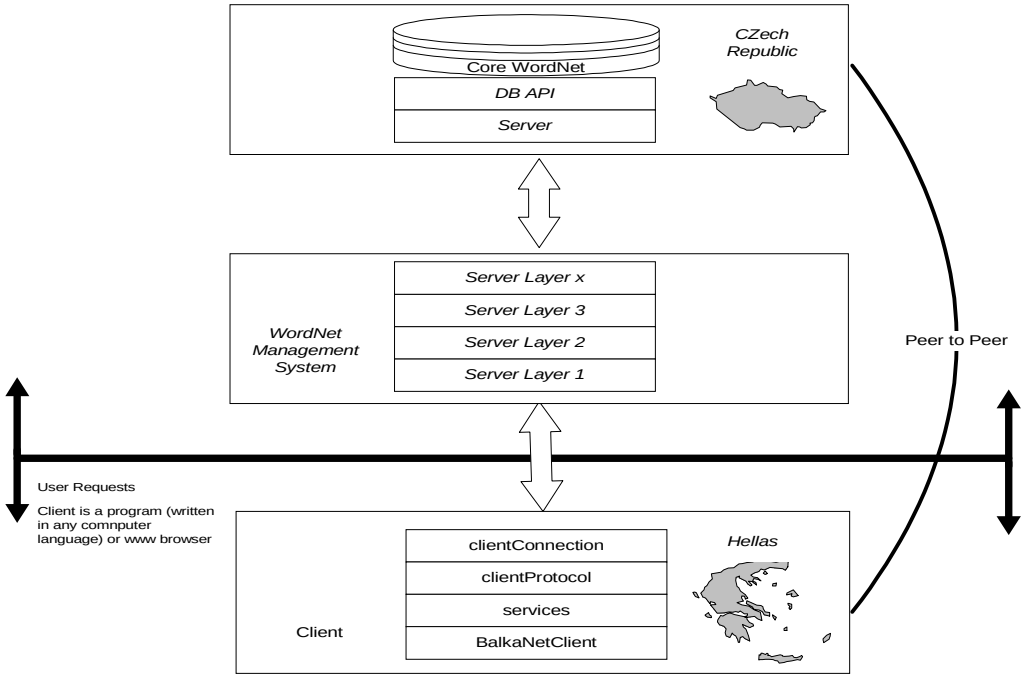
Stand Alone Access

The abovementioned platform can be easily transformed to a stand-alone architecture. Each partner has to contribute his store and DB API as shown at the Federated Architecture. The kernel is based on a protocol that communicates with all stores through DB APIS. It should be mentioned that the architecture of each store has its own structure and is implemented according to each partner. This is a main advantage not only at the Stand Alone Access but also at the Federated Access as well. Since each partner is responsible for interaction between DB API and store, the application kernel should not be aware of what is happening on the bottom side of the architecture.



Connectivity among services through protocol

A protocol should be established so that data exchange through client–W.M.S.–partner and "peer to peer" follow a defined Document Type Definition (D.T.D). The data will follow an XML format according to the D.T.D. The architecture is based on layers that a higher layer hides the complexity of the previous one.


Core WordNet
DB API
Server
CZech Republic
Server Layer x
Server Layer 3
Server Layer 2
Server Layer 1
WordNet Management System
Peer to Peer
User Requests
Client is a program (written in any comnputer language) or www browser
clientConnection
clientProtocol
services
BalkaNetClient
Client
Hellas

CONCLUSIONS AND FUTURE WORK

In this deliverable we have demonstrated the requirements set both by end users and developers of the BalkaNet network and we have described the methodology to be followed for the implementation of the multilingual database. More specifically, the project's objective is not only to develop a multilingual semantic network of Balkan concepts but also to offer European community the infrastructure of performing various NLP tasks in a more efficient way.

Towards this direction, we have performed a limited (due to time considerations) market research in order to trace areas towards which BalkaNet could contribute in a meaningful way. Having in mind that Information Retrieval Systems tend to use extensively linguistic information and infrastructures and by focusing of previous applications of semantic networks we concluded that BalkaNet could contribute towards conceptual classification tasks due to the hierarchical structure of the data it comprises of.

Moreover, classification techniques applied so far in IR tasks tend to use various kinds of thesauri and corpora which most of them are limited to the English language. On top of that web directories tend to classify documents based on pre-defined categories for which origins and updates we have little knowledge. BalkaNet aims at developing conceptual domains independent of the underlying languages, which will form the basis for the performance of conceptual indexing and classification tasks. A continuous update of the domains will be made feasible as the semantic network is glowing and already existing semantic directories will be able to be incorporated in our infrastructure.

Having in mind the final application of our project and by focusing on the users needs we described the requirements set both by users and the consortium and which will be followed during the implementation of the project. In particular, we can group the requirements described so far in the following categories:

- Data requirements (monolingual and multilingual)
- Architectural requirements
- Data management and representation requirements
- Maintenance requirements
- Quality and consistency checking requirements
- Product integration requirements
- Functionality and applicability of the project's results
- Technical infrastructure
- WordNet Management System requirements
- Multilingual architecture requirements

- Quality of the lexical resources
- Language internal relations
- The functional specification of the methodology

From now on based on the methodology specified above the actual development of the monolingual WordNets is going to take place along with the development of the technical infrastructure of the WordNet Management System. From a linguistic point of view lexical resources collected and examined so far are going to be processed in order to extract the first set of defining terms that will form the basis for the development of the core monolingual WordNets whereas from a technical point of view the actual implementation of the WordNet Management system will take place starting from the incorporation of the VisDic tool in it. Last but not least tools for the development of the monolingual WordNets and the processing of the lexical resources are going to be developed from scratch and enhanced in case they are already available in order to facilitate linguists organize their work and integrate data in the lexical database.

BIBLIOGRAPHY

Beckwith R., Miller G.A., (1990) "Implementing a lexical Network" in International Journal of Lexicography, Vol.3, No.4 (winter 1990), 302-312

Bloksma L., Diez-Orzas P., Vossen P., (1996) "User Requirements and Functional Specification of the EuroWordNet Project" LE-4003 EWN Project, Deliverable D001, version 5, final

Buitelaar P., Sacaleanu B. (2001) "Ranking and Selecting Synsets by Domain Relevance" In Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations, NAACL 2001 Workshop, Carnegie Mellon University, Pittsburgh, 3-4 June 2001

Carbonell, Jaime G., Yang, Yiming., Frederking, Robert E., Brown, Ralf, D., Geng, Yibing., Lee, Danny., (1998) in Artificial Intelligence Journal special issue: Best of IJCAI-97, 1998, pp.323-345

Dorr B., Jones D., (1996) "Role of Word Sense Disambiguation in lexical Acquisition: Predicting Semantics from Syntactic Cues" in proceedings of the International Conference on Computational Linguistics

Fellbaum C., (1998) "A Semantic Network of English: the mother of all WordNets" in Computers and the Humanities, Special Issue on EuroWordNet

Fellbaum C., (1990) "English Verbs as a Semantic Net" in International Journal of Lexicography, Vol.3. No.4 (winter 1990), 278-301

Gilarranz J., Gonzalo J., Verdejo F., Stanford C.A., (1997) "An Approach to Conceptual Text Retrieval Using the EuroWordNet Multilingual Semantic Database" in Working Notes of AAAI Spring Symposium of Cross-Language Text and Speech Retrieval

Gonzalo J., Verdejo F., Chugur I., Cigarran J., (1998) "Indexing with WordNet synsets can improve Text Retrieval": in COLING / ACL'98 Workshop on Usage of WordNet in Natural Language Processing Systems

Hearst M.A., (1998) "Automated Discovery of WordNet Relations" in WordNet: An Electronic Lexical Database, editor: Fellbaum C. MIT Press, Cambridge, USA spp.131-151

Jackendoff R., (1992) "Parts and Boundaries", in B. Levin and S. Pinker (eds.) Lexical & Conceptual Semantics, Cambridge University MA: Blackwell: 9-45

Levin B., (1993) "English Verb Classes and Alterations, a Preliminary Investigation" University of Chicago Press, Chicago / London

Miller G.A, Beckwith R., Fellbaum C., Gross D., Miller K., (1990) "Introduction to WordNet: An On-line Lexical Database", in International Journal of Lexicography, Vol.3, No. 4 (winter 1990), pp.235-244

Ntoulas A., Stamou S., Tzagarakis M., (2001) "Using a WWW Search Engine to Evaluate Normalization Performance for a Highly Inflectional Language" in proceedings of the ACL/EACL-2001 Student Research Workshop, 6-11 July 2001, Toulouse, France

Peters W., Peters I., Vossen P., (1998) "Automatic Sense Clustering in EuroWordNet", In Proceedings of the LREC Conference, 1998, Granada, Spain

Peters W., Vossen P., Diez-Orzas P., Adriaens G., (1998) "Cross-Linguistic Alignment of WordNets with an Inter-Lingual-Index" in Computer and the Humanities

Pustejobsky J., (1991) "The generative lexicon" in Computational Linguistics, 17, 4, Cambridge MA: MIT Press: 409-442

Pustejobsky J., Bergler S. (eds), (1992) "Lexical Semantics and Knowledge Representation" Proceedings of the First SIGLEX Workshop Berkeley, USA June 1991. Lecture Notes in Artificial Intelligence nr. 627. New York / Berlin: Springer Verlag

Richardson R., Smeaton A.F., (1995) "Using WordNet in a Knowledge-Based Approach to Information Retrieval" in Proceedings of the BCS-IRSG Colloquium, Crewe

Rodriquez H., Climent S., Vossen P., Bloksma L., Roventini A., Bertagna F., Alonge A., Peters W., (1998) "The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology", in Computer and the Humanities

Rosch E. (1977) "Classification of Real World Objects: Origins and representations in Cognition" in P.N Johnson-Laird and P.C. Wason (eds.) Thinking: readings in cognitive science. Cambridge: Cambridge University Press: 212-222

Smeaton A.F., Kenelly F., O'Donnell R., (1995) "TREC-4 Experiments at Dublin City University: Thresholding Posting Lists, Query Expansion with WordNet and POS Tagging of Spanish" in TREC-4 Proceedings, National Institute of Standards and Technology, Gaithersburg, Md. USA

Tengi R., (1998) "Design and Implementation of the WordNet Lexical Database" in WordNet: An Electronic Lexical Database, editor: Fellbaum C. MIT Press, Cambridge, USA pp.105-127

Vossen P., (1995) "Grammatical and Conceptual Individuation in the Lexicon" PhD Thesis, Universiteit van Amsterdam, Ifott 15

Vossen P. (1996) "Right or Wrong: Combining Lexical Resources in the EuroWordNet Project". In Proceedings of the Euralex Conference, pp. 715-728.

Vossen P., Diez-Orzas P., Peters W., (1997) "Multilingual Design of EuroWordNet" in Proceedings of the ACL-97 Conference, Madrid, Spain

Vossen P., (ed.) (1997) "EuroWordNet: A Multilingual Database for Information Retrieval" Kluwer Academic Publishers

Vossen P., Bloksma L., (1998) "Categories and Classifications in EuroWordNet" in Proceedings of the LREC Conference, 1998, Granada, Spain

Vossen P., Peters W., Gonzalo J., (1999) "Towards a Universal Index of Meaning" in Proceedings of the ACL-99 Siglex Workshop, University of Maryland, USA