



**Design and Development of a Multilingual Balkan  
Wordnet, (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](mailto:dxri@cti.gr)

Project Web site: <http://www.ceid.upatras.gr/Balkanet>

# Final Report

## List of Contributors

<i>Person</i>	<i>Affiliation</i>	<i>Country</i>
<b>Dimitris Christodoulakis</b>	Databases Laboratory, Patras University	Greece
<b>Natassa Kapatsoulia</b>	Databases Laboratory, Patras University	Greece
<b>Manolis Tzagarakis</b>	Research Academic Computer Technology Institute	Greece
<b>Sofia Stamou</b>	Databases Laboratory, Patras University	Greece
<b>Ioannis Dimitrios Koutsoubos</b>	Research Academic Computer Technology Institute	Greece
<b>Sofia Raikou</b>	Databases Laboratory, Patras University	Greece
<b>Pavlos Kokosis</b>	Research Academic Computer Technology Institute	Greece
<b>Vlassis Krikos</b>	Research Academic Computer Technology Institute	Greece
<b>Baso Aggelopoulou</b>	Databases Laboratory, Patras University	Greece
<b>Mata Anastasiou</b>	Databases Laboratory, Patras University	Greece
<b>Ioanna Ontambasidou</b>	Databases Laboratory, Patras University	Greece
<b>Vassilis Andrikopoulos</b>	Databases Laboratory, Patras University	Greece
<b>Dan Tufiş</b>	Institute for Artificial Intelligence, Bucharest	Romania
<b>Verginica Mititelu</b>	Institute for Artificial Intelligence, Bucharest	Romania
<b>Radu Ion</b>	Institute for Artificial Intelligence, Bucharest	Romania
<b>Eduard Barbu</b>	Institute for Artificial Intelligence, Bucharest	Romania
<b>Luigi Bozianu</b>	Institute for Artificial Intelligence, Bucharest	Romania
<b>Orhan Bilgin</b>	Sabancı University	Turkey
<b>Ozlem Cetinoglu</b>	Sabancı University	Turkey
<b>Cvetana Krstev</b>	Faculty of Mathematics, University of Belgrade	Serbia & Montenegro
<b>Gordana Pavlović-Lazetić</b>	Faculty of Mathematics, University of Belgrade	Serbia & Montenegro
<b>Ivan Obradović</b>	Faculty of Mathematics, University of Belgrade	Serbia & Montenegro
<b>Duško Vitas</b>	Faculty of Mathematics, University of Belgrade	Serbia & Montenegro
<b>Karel Pala</b>	Faculty of Informatics, Masaryk University	Czech Republic
<b>Pavel Smrz</b>	Faculty of Informatics, Masaryk University	Czech Republic
<b>Ales Horak</b>	Faculty of Informatics, Masaryk University	Czech Republic
<b>Dan Cristea</b>	University Alexandru Ioan Cuza	Romania
<b>Oana-Diana Postolache</b>	University Alexandru Ioan Cuza	Romania
<b>Georgiana Puscasu</b>	University Alexandru Ioan Cuza	Romania
<b>Corina Forascu</b>	University Alexandru Ioan Cuza	Romania
<b>Catalin Mihaila</b>	University Alexandru Ioan Cuza	Romania
<b>Gabriela-Eugenia Dima</b>	University Alexandru Ioan Cuza	Romania
<b>Svetla Koeva</b>	Department for Computer modeling of Bulgarian, IBL, Bulgarian Academy of Sciences	Bulgaria
<b>Tinko Tinchev</b>	Department for Computer modeling of Bulgarian, IBL, Bulgarian Academy of Sciences	Bulgaria
<b>Svetlozara Lesseva</b>	Department for Computer modeling of Bulgarian, IBL, Bulgarian Academy of Sciences	Bulgaria
<b>Angel Genov</b>	Department for Computer modeling of Bulgarian, IBL, Bulgarian Academy of Sciences	Bulgaria
<b>George Totkov</b>	Computer Science Dept., Plovdiv University	Bulgaria
<b>Rositza Doneva</b>	Computer Science Dept., Plovdiv University	Bulgaria
<b>Pavlina Ivanova</b>	Computer Science Dept., Plovdiv University	Bulgaria
<b>Dimitar Blagoev</b>	Computer Science Dept., Plovdiv University	Bulgaria
<b>Luben Milev</b>	Computer Science Dept., Plovdiv University	Bulgaria
<b>Tatiana Kalcheva</b>	Computer Science Dept., Plovdiv University	Bulgaria
<b>Petia Nesterova</b>	Computer Science Dept., Plovdiv University	Bulgaria

## Summary

BalkaNet is an EC funded project (IST-2000-29388) that started in September 2001 and finished in August 2004. It aimed at developing (Stamou et al., 2002 (b)) aligned wordnets for the following Balkan languages: Bulgarian, Greek, Romanian, Serbian, Turkish and to extend the Czech wordnet previously developed in the EuroWordNet project. BalkaNet project has insofar delivered many useful results in the fields of both Computational Lexicography and Natural Language Processing (NLP). This report attempts to provide an overall description of the findings, methodologies and results of the project as well as a detailed account on each monolingual wordnet. We also present the freeware multilingual tools designed for the development, maintenance and efficient exploitation of the aligned BalkaNet wordnets. Last but not least a preliminary approach on BalkaNet's application towards IR is described, following the consideration that semantic networks are valuable in the context of real world systems and user communities. The ultimate objective of this contribution is to spread the knowledge and experience that we have acquired, to the benefit of the research and industrial communities. We hope that our shared experience will be helpful for other wordnet-builders.

## Table of Contents

List of Contributors.....	2
Summary.....	3
Table of Contents.....	4
Multilingual Architecture Requirements .....	9
Design Strategies .....	9
Architecture.....	10
Wordnet Management System.....	14
Motivation.....	14
Wordnet Management System Services .....	15
Wordnet Management System Clients and Applications .....	16
Wordnet Management System Interface.....	17
VisDic Editor .....	19
The Experience and the Recommendation .....	20
An example of the DTD.....	20
VisDic XML Representation .....	22
Polaris format.....	23
Final XML format.....	23
VisDic Assessment .....	24
Objectives and Data Requirements.....	25
Lexical Data Acquisition .....	27
Selecting BalkaNet Base Concepts.....	28
Extending BalkaNet Base Concepts .....	30
Restructuring BalkaNet Interlingual Index.....	30
Re-linking BalkaNet ILI to PWN 2.0 .....	31
BalkaNet Specific-Concepts .....	32
The BalkaNet Inter-Lingual-Index .....	33
Adding Domain-Specific Concepts .....	34
Relations to SUMO and MILO Ontologies in VisDic.....	35
Selecting the SUMO Domain-Specific Concepts.....	35
Qualitative BalkaNet evaluation.....	37
Tests for Monolingual Wordnets and Quantitative Comparisons.....	38
Evaluating the Well-Formedness of Balkan Wordnets.....	38
Structure of a Wordnet File.....	39
Cross-Lingual Validation Based on a Parallel Corpus .....	40
Qualitative Evaluation Results.....	45
Experimenting with Valence Frames.....	48
Czech - Adding Verb Valency Frames .....	48
Bulgarian - Adding Verb Valence Frames.....	52
Bulgarian - Verb Net.....	54
Bulgarian Verb Net: Methods and Tools.....	54
Software Tools .....	55
▪ Frame construction for a target language <i>wordnet</i> ( <i>TLWN</i> ) using already developed frames of a source language <i>wordnet</i> ( <i>SLWN</i> );.....	57
Using Balkanet as a training environment for students – the Romanian experience...59	59
Current Status of the Balkan Wordnets .....	61
Status of the Greek Wordnet.....	61
Status of the Turkish Wordnet .....	61
Status of the Romanian Wordnet .....	62

Status of the Serbian Wordnet .....	63
Status of the Czech Wordnet.....	64
Status of the Bulgarian Wordnet.....	65
BalkaNet's Applications .....	67
Objectives and Current Status.....	67
Introducing Conceptual Domains in BalkaNet ILI.....	68
Conceptual Indexing Using Domain Taxonomies.....	68
Challenges.....	69
Steps for Web Documents Pre-processing.....	70
Pre-process web pages .....	70
Lexical chains .....	70
Find the topic category of a web page .....	70
Compiling and Processing a corpus of the BalkanTimes Web Archive.....	71
Impact .....	74
Testing Specifications .....	74
DISSEMINATING BALKANET.....	76
Conferences, Workshops, Special Sessions.....	76
Joining Global Wordnet Association.....	77
User Groups /Promotion and awareness .....	77
References.....	78

## Introduction

Semantic networks (Quillian, 1968) are among the most popular Artificial Intelligence formalisms for knowledge representation that have been widely used in the 70's and 80's to represent structured knowledge. Like other networks, they consist of nodes and links. Nodes represent concepts, i.e., abstract classes whose members are grouped together on the basis of their common features and/or properties, while arcs between these nodes represent relations between concepts and are labelled so as to indicate the relation they represent. In a semantic network, usually, the concepts' labels are mnemonics, informative for the knowledge engineer developer. The semantics of the concepts resides not in the name of the associated labels, but in the concepts' properties and relations to other concepts of the semantic network. The last 20 years or so have seen a tremendous resurrection of interest in semantic networks formalisms boosted, among others, by CYC, the impressive work of Lenat and his colleagues (Lenat, 1995). The ontological representation of general and domain specific knowledge is now claimed to be a sine-qua-non support to any attempt to intelligently solve the hard problems faced by the modern information technology. A special form of the traditional semantic networks came out from the pioneering work of George Miller and his co-workers (Miller, 1990) at Princeton University. They developed the concept of a lexical semantic network, the nodes of which represented sets of actual words of English sharing (in certain contexts) a common meaning. These sets of words, called synsets (synonymy sets), constitute the building blocks for representing the lexical knowledge reflected in WordNet, the first implementation of lexical semantic networks. As in the semantic networks formalisms, the semantics of the lexical nodes (the synsets) is given by the properties of the nodes (implicitly, by the synonymy relation that holds between the literals of the synset and explicitly, by the gloss attached to the synset and, sometimes, by specific examples of usage) and the relations to the other nodes of the network. These relations are either of a semantic nature, similar to those to be found in the inheritance hierarchies of the semantic networks, and/or of a lexical nature, specific to lexical semantics representation domains. The convergence of the representational principles promoted both by the domain-oriented semantic networks and ontologies, and by WordNet's philosophy in representing general lexical knowledge, is nowadays an apparent trend, motivated not by fashion, but by the significant improvements in performance and by the naturalness of interaction displayed by the systems that have adopted this integration. Several NLP systems based on semantic networks initially (80's) relied on (limited) domain specific semantic lexicons for mapping synonymic words used in the input to the same concept of the underlying semantic net. The IURES system (Tufiş and Cristea, 1985a, b) is just one such example.

However, the tremendous technological advancement of the recent years in computers' speed and storage capacity, the unforeseen Web evolution, the widespread of understanding and usage of WordNet, as well as the maturity of the ontology-based technologies, made possible up-scaling the integration of domain knowledge and lexical knowledge at an unprecedented level. This interdependency, which is not always explicit, motivated several researchers' doubts on language independent ontologies (Quine, 1960; Hovy&Nirenburg, Hirst, 2003, etc.)

The public release of the Princeton WordNet (PWN), encoding lexical knowledge about American English, gave an impetus to world-wide research in developing similar knowledge representation resources for other languages. As a distinctive sign of recognition of this impact, the name of the Princeton's semantic network became a common noun – *wordnet* – defining a similarly organized lexical knowledge base for a different language. More than 50 wordnets (for a partial list cf. [http://www.globalwordnet.org/gwa/wordnet\\_table.htm](http://www.globalwordnet.org/gwa/wordnet_table.htm)) are nowadays under construction, all over the world, for more than 40 languages.

The EuroWordNet (EWN) project (LE-2 4003 & LE-4 8328), which started in March 1996 and ended in June 1999, extended the PWN approach with the multilingual dimension adding an Inter-Lingual Index to which all the monolingual wordnets for the languages represented in the project were aligned. The Inter-Lingual Index (ILI) was based on the PWN 1.5, the syn-

sets of which played the role of language independent concepts. The interlingual index was further extended with some language (other than English) specific concepts. Another major extension was the association with each of the so-called Base Concepts of an ontological description subject to be inherited by all more specific concepts in the ILI. For the monolingual wordnets the same structuring as in PWN (Miller et al., 1990) was preserved and via the ILI (interconnecting the languages) it is possible to go from the words in one language to semantically close words in any other language. The index also gives access to a shared top-ontology of 63 semantic distinctions. This top-ontology provides a common semantic framework for all the languages, while language specific properties are maintained within the individual wordnets. The languages represented in EWN were Dutch, Italian, Spanish, German, French, Czech, Estonian and obviously English. The alignment of the monolingual wordnets on the basis of the interlingual index as well as the shared top-ontology turned the EWN multilingual lexicalized semantic network into a multilingual lexical ontology. A detailed presentation of the principles, methodology and results of the EWN project is given in (Vossen, 1999) and on the EWN website (<http://www.ilc.uva.nl/EuroWordNet/>).

Although PWN's coverage does not compare yet with any of the existent wordnets, the latter are continuously extended so that a balanced multilingual wordnet is foreseen in the future. Most of the wordnet projects are affiliated to a recently established professional association, Global Wordnet Association (<http://www.globalwordnet.org/>), which already organized two very successful international conferences (in Mysore, India and in Brno, Czech Republic).

A major contribution to the furthering of the EWN principles (Rodriguez et al., 1998) is the ongoing European project BalkaNet (IST-2000-29388) which initially aimed at extending the pool of the EWN languages with five South-Eastern European languages from the Balkan area: Bulgarian, Greek, Romanian, Serbian and Turkish. In the consortium have been included the Czech and the French teams that participated in the EWN, to liaise towards a perfect compatibility with previously developed wordnets. Also the coordinator of the EWN project, Dr. Piek Vossen, was solicited and accepted to be a consultant for the BalkaNet project. Besides compatibility with the other aligned wordnets, the BalkaNet project ambitioned to a better quality and to a much wider cross-lingual coverage than in EWN. Therefore, the quality control was much stricter.

This report gives an overview of the project in terms of objectives, approaches, methodologies and general development issues. It also presents ongoing research and development activities towards building intelligent applications and exploiting the aligned wordnets of BalkaNet. We report on the challenges associated with building multilingual lexicalized semantic networks. Despite the advances of many recent attempts in building wordnets for a plethora of natural languages, a significant amount of difficulties needs to be tackled every time a new wordnet starts being developed. Such difficulties emerge from languages' properties and lexical resources completeness and deal with the representation of conceptual knowledge. Our incentive is to provide semantic network and lexical semantics communities with valuable insights on the experience and the knowledge we have accumulated while building BalkaNet, so as to contribute in the improvement of their work in as much as possible.

Before going into further details, let us define three terms relevant for the discussions to follow: "sense", "meaning" and "concept". Although closely related, and sometimes interchangeably used, these notions are slightly different distinguishing the perspective from which the encoded knowledge is considered. The notion of *sense* is strictly referring to a word. The polysemy degree of a word is given by the number of senses the respective word has. A traditional explanatory dictionary provides definitions for each sense of a headword. The notion of *meaning* generalizes the notion of *sense* and it could be regarded as a set-theoretic equivalence relation over the set of senses in a given language. In colloquial speech one says *this word has the same meaning with that word* while a more precise (but less natural) statement would be *the  $M^{\text{th}}$  sense of this word has the same meaning with the  $N^{\text{th}}$  sense of that word*. Synonymy, as this equivalence relation is called, is a lexical relation that represents the formal device for clustering the word senses into groups of lexicalized meanings. The

meaning is the building block in wordnet-like knowledge representations. In PWN and all its followers the meanings in the respective languages are represented as *synsets* (synonymy set) and they are implemented as sets of word senses. Each synset is associated with a gloss that covers all word senses in the synonymy set. The meaning is thus a language specific realization of a *conceptualization* which might be very similar to conceptualizations in several other languages. Similar conceptualizations are generalized in a language independent way, by what we call *interlingual concepts* or simply *concepts*. The meanings in two languages that correspond to the same concept are said to be translation equivalent. One could arguably say that the interlingual concepts cannot entirely reflect the meanings in different languages (be it only for the historical and cultural differences), however, concepts are very useful generalizations that enable communication across speakers of different natural languages. In multilingual semantic networks the interlingual level ensures the cross-lingual navigation from words in one language to words in the other languages. Both EWN and BalkaNet adopted as their interlingual concepts the meanings of PWN. This choice was obviously a matter of technological development and a working compromise: the PWN displayed the greatest lexical coverage and is still unparalleled by any other language. To remedy this Interlingua status of English, both EWN and BalkaNet considered the possibility of adding in the inter-lingual index concepts which represent language specific meanings (or meanings specific to a group of languages).



## Multilingual Architecture Requirements

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

With respect to the multilingual architectural requirements the following processes are 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 are integrated in the final database without any prior structural changes. Moreover, the final database is 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 provided to end users for the use of the final multilingual BalkaNet database.

All the abovementioned requirements issued by developers of the project have been fulfilled so that the final resource is easily applicable to any kind of task. The following sections give a detailed overview of the aforementioned criteria and explain who these are met in the methodology followed for the implementation of the project.

## Design Strategies

Following the principles adopted in EWN (Vossen et al., 1997b) and PWN (Miller, 1990), producing a multilingual semantic network fully compatible with EWN (and its extensions) was a general commandment. Thus, it was envisaged an unprecedented multilingual semantic network, covering 15 European languages and creating incentives for other ongoing monolingual wordnets to join it. The benefits of such a multilingual knowledge resource are huge and not only for the less studied languages involved in BalkaNet.

To guarantee monolingual wordnets' compatibility of the approaches followed by the EWN consortium were adopted, the most important of which are: EWN's ILI, EWN's lexico-semantic relations, and EWN's Top-Ontology and Base Concepts (BCs) (Vossen et al., 1997 (a)). However, besides being in line with EWN it was desirable to keep up with the continuous improvements made in the PWN. To account for that we have performed updates to the BalkaNet's ILI every time a new PWN version was released. Thus, having initially employed PWN 1.5 as BalkaNet's ILI, we switched to PWN 1.7.1 and then to PWN 2.0, which is the latest PWN release and the current Interlingua of BalkaNet. To warrant a significant concep-

tual overlap among the BalkaNet wordnets a common set of 8,000 concepts was selected to be linguistically realized in all six languages of the project. Starting off with a common set of concepts ensures a satisfactory degree of conceptual intersection across wordnets and facilitates the cross-lingual evaluation and comparison of the monolingual repositories. The adopted development methodology was supposed to ensure that further independent extensions of the monolingual wordnets would not weaken the conceptual inter-lingual coverage.

A great challenge of BalkaNet was to deliver lexical resources and NLP tools that would be flexible and re-usable across different applications and user communities. Given the apparent lack of available free-source wordnet building tools it was decided to develop BalkaNet's technical infrastructure in a way so that it is easily adaptable to other tasks. Besides VisDic and Wordnet Management System (WMS), several tools have been built that enable the efficient exploitation of the monolingual lexical resources (i.e., explanatory dictionaries, corpora, thesauri etc.). Those tools have been developed on the basis of the structure and the content of the various lexical resources available and enable the autonomous development of each monolingual wordnet. A significant amount of work has been also devoted in checking the quality of the delivered wordnets and several tools have been implemented towards this task. The specifications behind our methodology for data acquisition and processing were defined on the grounds of modularity, robustness and re-usability. This way we aspire to provide the wordnet-community some missing pieces to the understanding of the evolution of semantic networks.

## Architecture

The data of TID are stored in a relational database: Firebird 1.5. This free RDBMS has all the needed characteristics for such a project:

- Capacity (in terms of number of records, columns, tables, indexes...)
- Support of Unicode: this requirement was of course mandatory to store so many different languages with their different character sets.
- Simplicity
- Speed
- Multi platform (Windows, Linux)
- Stored procedures

Basically, the architecture of TID is very simple because relying mainly in two tables: the LEAF table and the RELATION table.

### The LEAF table:

This table contains all of the nodes of the graph: the concepts, the word senses (or the literals, using the wordnet terminology), the glosses, the ILIs.

Structure of this table

Field	Type	Length	description
LANGUAGE	Char	1	language of the Word Sense or the gloss. The different values are : E English F French I Italian D German H Dutch S Espagnol P Portuguese G Greek

			T Turkish R Rumanian C Czech Y Serbian B Bulgarian W Swedish For a concept or an ILI, we use the character 'M' as metalanguage
SITE	Char	1	Code which indicates where the node has been created (this information is mandatory for the unicity of the key)
NUMBER	Integer		Numerical element of the key
WORDING	Varchar	400	Wording of a word sense or a gloss. In the case of a concept or an ILI, this field can be blank: the wording of these nodes would then be assured by other nodes of type Gloss, linked to them.
GRAMMAR	Integer		This field contains the code of the POS for a Word Sense or the type of a Concept. These information are stored in an additional table.
DATE	Date		Date of creation or modification of the node
MODEL	Integer		Number of inflection model for the nouns, adjectives or verbs
ARTICLE	Varchar	80	Contains the 80 first characters of the wording in uppercase. This allows to retrieve a node by it's wording. In some cases this field may be different of the beginning of the wording. This field is indexed.

The three first fields compose the primary key of the node (ex MA15224, EW566711...)

#### The RELATION table:

Each row of this table contains the relation between two nodes. We call, conventionally, the first node the child node and the second node the parent node.

Structure of this table:

Field	Type	Length	description
LANGUAGE_CHILD	Char	1	Language of the child node
SITE_CHILD	Char	1	Site of the child node
NUMBER_CHILD	Integer		Number of the child node
LANGUAGE_PARENT	Char	1	Language of the parent node
SITE_PARENT	Char	1	Site of the parent node
NUMBER_PARENT	Integer		Number of the parent node
DATE	Date		Date of creation of the relation
TYPE	Integer		Type of the relation (specific, generic, etc). These informations are stored in an additional table.
LANGUAGE_CONTEXT	Char	1	Language of the context node (see below)
SITE_CONTEXT	Char	1	Site of the context node
NUMBER_CONTEXT	Integer		Number of the context node

**Figure 1:** the RELATION Table

The context is a node which allows to precise the context of a relation. The figure below shows the initial data format that TID used to represent it.

Child	Parent	KindOfRel
Author (n)	\author of a lit...	Generic
\author of play	\author of a lit...	Specific
etc.		

**Figure 2:** A general record in the table RELATION in TID.

Although this format was satisfactory for hierarchical data, it reached its limits when we introduced syntactical relations. Let's consider the syntactic definition in **Error! Reference source not found.**:

```
\author of a literary work (List)  SV  \write
                                   VO  \texts
```

Figure 2 shows the table in TID using the same formalism.

Child	Parent	KindOfRel
\author of a literary work (List)	\write	SV
\write	\texts	VO
etc.		

**Figure 3:** A part of TID.

However, it not possible to consider that \author of a literary work (List) is the child of \write and the grandchild of \text in Figure in the same way it is the child of \author of a lit... in the above figure. In addition, in terms of graph, the syntactic paths cannot be recorded without ambiguity, for example if write exists in many different assertions.

Syntactic patterns and lexical ontology represent two different viewpoints that are not necessarily related. To represent them with a relational database, we must take into account that these two dimensions (syntactic/paradigmatic) are different. Figure shows the integration results where

OntoTID means ontology of TID and SyntTID means Syntactical Pattern of TID. The index (1) is the key of the complete pattern. The two last records indicate that OntoTID and SyntTID are parts of TID. This format is more flexible and provides rich new possibilities. Firstly, the format can record any kind of hypergraph in a relational database. Secondly, it enables us to extend the group theory approach to a more general mereology.

Child	Parent	KindOfRel	Context
Author (n)	\author of a lit...	Generic	OntoTID
\author of play	\author of a lit...	Specific	OntoTID
etc.			
\author ...(List)	\write	SV (1)	SyntTID
\write	\texts	VO (1)	SyntTID
etc.			
OntoTID	PartOfTID	part of	TID
SyntTID	PartOfTID	part of	TID

**Figure 4:** A part of TID.

We have used this format to integrate a set of ontological resources. Concerning EuroWordNet and BalkaNet, the format allows us to upload data from xml files to a relational database. Figure shows an excerpt of records where (1) is a key identifying a synset.

Since a synset has its gloss and literal, we have the English gloss {writes (books or stories or articles or the like) professionally (for pay)...} and the English literal *author* located in the *English WordNet*. We notice that in this case, *auteur (n)* is placed in the synset (1) in the

*French wordnet*. In the end, it's also possible to generate the complete list of InterLingua index (ILI).

<i>Child</i>	<i>Parent</i>	<i>KindOfRel</i>	<i>Context</i>
<i>Author (n)</i> {writes (books or stories or articles or the like) professionally (for pay)...}	<i>(ILI 1)</i>	<i>Literal</i>	<i>EnWordNet</i>
<i>auteur (n)</i> (ILI 1)	<i>(ILI 1)</i>	<i>Gloss</i>	<i>EnWordNet</i>
	<i>(ILI 1)</i>	<i>Litteral</i>	<i>FrWordNet</i>
	<i>Inter-lingua</i>	<i>Elementof</i>	<i>ILIs</i>

**Figure 5:** The wordnets.

We will see in part four that the relations between synsets may occur in some wordnets and not in other ones. Context will allow representing that.

#### **Indexes:**

The primary key is made up of the following fields:

LANGUAGE\_CHILD

SITE\_CHILD

NUMBER\_CHILD

LANGUAGE\_PARENT

SITE\_PARENT

NUMBER\_PARENT

TYPE

LANGUAGE\_CONTEXT

SITE\_CONTEXT

NUMBER\_CONTEXT

There are two secondary indexes. One on the fields LANGUAGE\_CHILD, SITE\_CHILD, NUMBER\_CHILD and one on the fields LANGUAGE\_PARENT, SITE\_PARENT, NUMBER\_PARENT. These two indexes allow getting the children or the parents of a node.

## Wordnet Management System

### **Motivation**

- *Monolingual Wordnet Independency*: One of the major principles during the design of the WMS was the independency in the development and manipulation of each wordnet, regardless of its context, i.e. the environment created by the wordnets that this one is connected to. This approach complements in a way the merge approach that was adopted for the BalkaNet project but isn't limited to that, allowing the management of semantic resources and a local level, independently of whether they are inter-connected to others or not.
- *Web access*: An almost de facto requirement in a community like BalkaNet consisting of many different members, the need for access to the system via the Web is made imperative by the size of the data in case they had to be installed and the diversity of access methods that the applications that use them require.
- *Flexible access to semantic data by applications*: The design and development of WMS was mostly motivated by the need for the existence of a system that could be used not only by users but also (and mainly) by applications. But this need for machine readable information also requires a certain degree of interoperability among the system and the applications or other systems that use its services. For this purpose, the system must be able to provide information in a format that can be easily manipulated and transformed into other formats or results.
- *Unified Platform of Wordnet Structure related services*: A critical element in designing a wordnet management infrastructure is the efficient utilization of the wordnet's inherent hierarchical structures under a coherent platform. This would translate into exploiting relations that link the synsets and navigating within the relation trees (or networks in some cases). In this way, the information provided by the position of the synset in a hierarchy can be further used to provide semantic data on tree level or to allow the calculation of structural information like the semantic distance of two synsets that can be necessary in applications like Word Sense Disambiguation and Information Retrieval.
- *Distributed information sources*: The need for the distribution of the information sources stems from the nature of the sources themselves. Since the independence in development and manipulation (and therefore retrieval) was to be maintained, then the information has to be distributed among the wordnet developers. Furthermore, the current trends in system design that are mostly influenced by the Peer-to-Peer paradigm call for the location of the information to 'hidden' to the user, enabling an abstraction between the data and their actual location that can facilitate the development both of new applications and information sources.
- *Platform Independency*: WMS has been envisaged from the beginning as a platform-independent tool that could be used under the majority of the operating systems with the minimum effort possible.

The main advantages of Wordnet Management System are:

- Open-Ended Platform
- State-of-the-Art technologies
- Distributed management and control
- Flexible access to provided services and data.
- Data Storage Independent.

Wordnet Management System's future directions include:

- Versioning of Datasources
- Full Ontology Support on CWMS
- Wordnet Authoring
- More Multilingual services
- Incorporation of other Lexical resources
- Standardization of Data Representation

## ***Wordnet Management System Services***

The basic services provided by the WMS are described below:

### **Monolingual Services**

These services handle the retrieval of either semantic (i.e. content-related) or statistic (i.e. structure-related) data from each monolingual Wordnet. Every wordnet can be accessed by a unique identifier within the context of the WMS network.

#### **getProviders()**

Description: Returns information about wordnets that are hosted into WMS network. Input: None / Output: An array of hosted wordnets, containing each one's (unique) id, server, name, version and natural language.

#### **getSynsetIds(wordnet)**

Description: This service provides all synset identifiers that exist into the requested wordnet. Input: The unique identifier of the wordnet to be queried. / Output: An array of strings, each one containing a synset identifier.

#### **getBaseConcepts(wordnet)**

Description: Provides all synset identifiers that belong to the requested base concept set. Input: The unique identifier of the wordnet to be queried, the Base Concept group. / Output: An array of strings, each one containing a synset identifier of a Base Concept.

#### **getSynsetById(wordnet, synsetid)**

Description: Provides information about the synset identified by the specified id in the selected wordnet. Information contains POS (Part-of-Speech), gloss, Base Concept group and senses of the synset. Input: The unique identifier of the wordnet to be queried, the synset identifier. / Output: The synset that corresponds to the given identifier.

#### **getSynsetByLiteral(wordnet, literal)**

Description: Provides information about synsets containing the specified literal in the selected wordnet. Information contains POS, gloss, Base Concept group and senses of the synset. Input: The unique identifier of the wordnet to be queried, the literal to find. / Output: An array of synsets that meet the query criteria.

#### **getSynsetRelations(wordnet, synsetid)**

Description: Provides information about the semantic relations of the synset identified by the specified id in the selected wordnet. Input: The unique identifier of the Wordnet to be queried, the synset identifier. / Output: An array of synset's semantic relations.

#### **getSynsetRelationsByRelation(wordnet, synsetid, relation)**

Description: Provides information about the semantic relations of the synset identified by the specified id in the selected wordnet. Input: The unique identifier of the wordnet to be queried, the synset identifier, the relation to find. / Output: An array of synset's semantic relations.

**getSynsetTree(wordnet, synsetid, relation)**

Description: Provides a tree structure for the requested synset, according to the requested relation, placing the requested synset as root of the tree. Input: The unique identifier of the Wordnet to be queried, the synset identifier, the semantic relation to find. / Output: The tree structure that is formed.

**getNodeCount(wordnet, synsetid, relation)**

Description: Provides the number of nodes contained in the tree of the requested synset. Input: The unique identifier of the wordnet to be queried, the synset identifier, the semantic relation to find. / Output: An integer which represents the number of nodes in the tree.

**getTreeDepth(wordnet, synsetid, relation)**

Description: Provides the number of levels contained in the tree of the requested synset. Input: The unique identifier of the wordnet to be queried, the synset identifier, the hierarchical relation to find (HYPERNYM/HYPONYM). / Output: An integer which represents the number of levels contained in the tree.

**getDistance(wordnet, synsetid1, synsetid2, relation)**

Description: Provides the distance between two synsets as the difference of their levels in the tree that is formed if for the specified hierarchical relation these synsets share the same root. Input: The unique identifier of the wordnet to be queried, the synset identifier, the hierarchical relation to find. / Output: An integer representing the calculated distance.

**Multilingual Services**

These services can provide the same content-related retrieval of information for a given synset, but this time on a multi-wordnet level by utilizing the common point of reference that provides the BalkaNet ILI or another language-independent structure. The output of these services utilizes a mapping structure called the Hashtable which maps synset information to a (unique) Wordnet identifier.

**getMSynsetById(wordnet[], synsetid)**

Description: Provides information about the synset identified by the specified id for each specified wordnet. Input: The Wordnet identifiers to search, the synset identifier. / Output: A Hashtable containing the synset information for each Wordnet.

**getMSynsetByLiteral(wordnet[], literal)**

Description: Provides information about the synsets that contain the specified literal. By the specified id for each specified wordnet. Input: The wordnet identifiers to search, the literal to query. / Output: A Hashtable containing the synset information for each wordnet.

**Wordnet Management System Clients and Applications**

On top of the WMS API various clients have been realized, including a graph browser for wordnet trees, an MS .NET client, a plug-in to Microsoft Office allowing thesaurus-style access to WMS semantic data. The graph browser is a custom application that uses the tree-like structure inherent to the wordnet, due to the interconnection created by the different kind of relations among synsets. WMS in this case is used as the provider of the relational data, leaving the actual representation of the structure to the application itself.

The Microsoft .NET Client for WMS was built as a demonstration tool, using the WSDL document that describes services that are provided by a standard WMS Server. It performs standard wordnet browsing operations, such as search by literal name and synset id, and re-



trieval of synset information like relations. By these means, it can be used as an ad hoc wordnet browser, but with the additional feature that it can be set to access more than one wordnet (local or remote) at a time.

The purpose of the development of the Microsoft Office plug-in was to provide access to the linguistic data inherent in the multilingual database that is formed by the wordnets of the BalkaNet project to every day applications like Microsoft Word. For this purpose, the plug-in utilizes the services provided by WMS to retrieve data like the synonyms of a given word and provide them to the user as thesaurus-like information. In this way, WMS provides the opportunity for wordnets to be used as a repository of multilingual linguistic information that is available to a multitude of every day applications like text editors, word processors and even internet browsers.

## Wordnet Management System Interface

The following screenshots illustrate the WMS interface.

The screenshot displays the BalkaNet Wordnet Management System interface. At the top, there is a navigation bar with the following menu items: Home, Synset Search, Advanced Search, Preferences, and Logout. Below the navigation bar, the main content area is titled "View HYPERNYMS tree".

Metadata for the current view is shown:

- Wordnet: GrWn (local) version: 0.6
- Synset id: ENG20-02244530-n
- Synset Name: προντίκι:1,προντικός:1;

The hypernymy tree is displayed in a list format:

- [ENG20-02243671-n]τρωκτικό
  - [ENG20-01805729-n]πλακουνοφόρα
    - [ENG20-01780968-n]θηλαστικό:1
      - [ENG20-01394664-n]σπονδυλόζωο:1
        - [ENG20-01389442-n]χορδωτό:1
          - [ENG20-00012748-n]ζώο:1
            - [ENG20-00003226-n]οργανισμός:2
              - [ENG20-00003009-n]ζωντανός οργανισμός:ον
                - [ENG20-00016236-n]άψιχο αντικείμενο:1
                  - [ENG20-00001740-n]οντότητα:1

Figure 6: Hypernymy tree representation via WMS interface

**Wordnet: RomanianWn (local) version: 2.0**

|                    |   |                                 |                         |
|--------------------|---|---------------------------------|-------------------------|
| <b>Synset ID</b>   | ENG20-02244530-n  | <a href="#">View Synset</a>     | <a href="#">new win</a> |
| <b>Synset Name</b> | șoarece:1;  | <i>View Tree:</i>               |                         |
| <b>Pos</b>         | n   | <a href="#">View Hyponyms</a>   | <a href="#">new win</a> |
| <b>Gloss</b>       | Animal mic din ordinul rozătoarelor, de culoare cenușiu-închis, cu botul ascuțit și cu coada lungă și subțire | <a href="#">View Hypernyms</a>  | <a href="#">new win</a> |
|                    |   | <a href="#">Advanced Search</a> | <a href="#">new win</a> |


**Wordnet: TurkishWn (local) version: 2.0**

|                    |  |                                 |                         |
|--------------------|--|---------------------------------|-------------------------|
| <b>Synset ID</b>   | ENG20-02244530-n   | <a href="#">View Synset</a>     | <a href="#">new win</a> |
| <b>Synset Name</b> | fare:1;  | <i>View Tree:</i>               |                         |
| <b>Pos</b>         | n  | <a href="#">View Hyponyms</a>   | <a href="#">new win</a> |
| <b>Gloss</b>       | Sıçangillerden, küçük vücutlu, kemirgen, memeli hayvan (Mus) | <a href="#">View Hypernyms</a>  | <a href="#">new win</a> |
|                    |  | <a href="#">Advanced Search</a> | <a href="#">new win</a> |

**Wordnet: BulgWn (local) version: 2.0**

|                    |   |                                 |                         |
|--------------------|---|---------------------------------|-------------------------|
| <b>Synset ID</b>   | ENG20-02244530-n                            | <a href="#">View Synset</a>     | <a href="#">new win</a> |
| <b>Synset Name</b> | мишка:1;                                    | <i>View Tree:</i>               |                         |
| <b>Pos</b>         | n   | <a href="#">View Hyponyms</a>   | <a href="#">new win</a> |
| <b>Gloss</b>       | дробен гризач с остра муцуна и дълга опашка | <a href="#">View Hypernyms</a>  | <a href="#">new win</a> |
|                    |   | <a href="#">Advanced Search</a> | <a href="#">new win</a> |

**Figure 7:** Cross-lingual synset browsing WMS interface



Balkanet Wordnet Management System

Home
Synset Search
Advanced Search
Preferences
Logout

**Advanced search for synset with synset ID: ENG20-02244530-n**

Synset Name: ποντίκι:1;ποντικός:1; [\[help\]](#)

Synset ID:

**Select wordnets**

Local:

- SrpWn version: 2.0
- PrincetonWn version: 2.0
- GrWn version: 0.6
- RomanianWn version: 2.0
- TurkishWn version: 2.0
- BulgWn version: 2.0
- CzezWn version: 2.0

[search](#)

**Figure 8:** Cross-lingual search main interface

## VisDic Editor

VisDic has been developed mainly for browsing and editing wordnet databases when it was clear that the development of Polaris (EuroWordNet 1, 2) would not continue. However, from the beginning it has also been designed to view and edit any other lexical data in XML format – in this respect it essentially differs from all previous wordnet tools. Thus, VisDic is able to work with XML format, which is regarded as a standard and is readable by many other applications. It should be also remarked that VisDic has been developed as a local tool only.

The following reasons led us to the solution based on using XML representation of the wordnet structures:

1. XML formalism definitely comes as a good candidate for a common interchange format that may significantly facilitate sharing of wordnet-like data within and between several languages and this can be done, in fact, independently of the actual implementation of the particular databases. We already converted into XML representations all 8 wordnets from EuroWordNet 1,2 and it can be shown that this conversion helps to correct some inconsistencies in the individual databases, e.g. lost or dangling synsets, missing links to ILI, etc.
2. We also have made a second obvious step – i.e. together with the XML representation we have developed a tool called VisDic (Horak, Smrz, 2004) that can work with it and is intended as a replacement of Polaris tool being used so far. It is implemented under Linux and Windows and after the necessary testing it has become more accessible than Polaris.
3. In comparison with the Italian proposal by Magnini and Girardi (2001) our XML representation is quite closely related to the VisDic tool and because of this it is more specific and not so general as the Italian one. However, the reason is obvious, when developing the tool we had to consider the criteria relevant for the implementation, i.e. features like speed and efficiency. Moreover, this format was developed to hold only the necessary information and not repeat facts that can be derived (for example hyponyms can be derived from hypernyms). This will be demonstrated in the examples below.
4. We would like to stress that on the other hand, however, our ambition was to develop even more general XML representation that would allow to use the tool VisDic not only for the wordnet-like databases but also for any machine readable dictionary that was (or can be) converted into XML format which can be processed by our tool. This result is based on the previous work which is still going on in our NLP Lab. (Karasek, 2000) and includes the conversion of the large *Dictionary of Literary Czech* (SSJC, sec.edition 1989, size approx. 200 000 entries) and smaller *Dictionary of Written Czech* (sec.edition 1994, size approx. 60 000 entries) into XML representation. This solution has proved to be very fruitful – recently the *Dictionary of Czech Synonyms* (sec.edition, 2000, size approx. 21 000 entries and 37 000 synsets) has also been converted into XML representation and can viewed and edited under VisDic.
5. Importing and exporting files: VisDic has been developed as a tool that is able to import and export any XML structured file. The export is performed automatically during the dictionary loading. The XML file is converted to the inner binary representation, which is not immediately readable, but allows the fast searching and editing entries.
6. Journaling (versioning): if we want to modify a wordnet and yet keep the option to restore the original version of the text (as it was before certain changes were made or as of a certain point in time), we need what is called *versioning*. This can be handled by the process called *journaling of changes*: we keep the file containing the original text and create a file of changes where we enter the individual changes made. To obtain the actual picture of the wordnet, we load the original file into memory and gradually carry out all the changes from the file of changes. With each entered change, we note down the time and the originator of the change, regardless of whether it was a user or program. The state before

modification can be restored by skipping a given change. In this way it is possible to keep track of all the changes made by the different people and later decide which one should be kept and which one discarded.

## ***The Experience and the Recommendation***

Our present experience both with wordnet-like databases and the mentioned Czech dictionaries confirms ultimately that XML representations and the respective DTD's can be taken as a good basis for the development of the standards in the area of machine readable dictionaries and lexical databases of various kinds (not necessarily just wordnet-like ones). The main advantages are:

- a) XML representations are general enough and transparent and they can be easily modified and adapted at the same time,
- b) there are tools that make it easy to work with them,
- c) if there is a machine readable dictionary (in any form, even in the form of the typesetting tapes or just having the form of appropriate (e.g. \*.rtf) files containing the typesetting information it is not so difficult to write the respective conversion script which turns the starting dictionary text into XML format,
- d) The experience with the conversion of the large Czech dictionary mentioned above (SSJC) shows that XML representation is suitable also for large dictionaries. In this point we have to add that apart from VisDic tool another dictionary browser is being developed in NLP Lab., called Dictionary Editor and Browser (DEB), which is based on client-server architecture and designed also for classical format of SSJC. It also displays other features, e.g. quite powerful query language, the integration of the morphological analyzer into it as well as the connection to the corpus manager working with our Czech corpora. The purpose for which DEB is being developed comes from the need to turn the rich SSJC data into a regular machine readable dictionary, i.e. to make it more consistent and to check its data where possible. Secondly, we think that DEB should have some important features that will make it more powerful and allow it to be used as a lexicographer's workstation.

### **An example of the DTD**

A `word_meaning` element (Figure 2 and 3) is used to describe both monolingual and ILI synsets. `Word_meaning` attributes include a unique identifier (ID), a part of speech and the synset gloss. There are elements to describe objects related to a word meaning, such as top ontology concepts, domain ontology concepts, variants, internal (i.e. language dependent) relations, and equivalence relations to the ILI interlingua.

```
<!ELEMENT word_meaning (#CDATA | gloss? | concepts? | domains? | variants | internal_links? | eq_links?)>
  <!ATTLIST word_meaning
    id CDATA #REQUIRED
    part_of_speech CDATA #REQUIRED>

<!ELEMENT gloss (#PCDATA)>
<!ELEMENT concepts (#PCDATA)>
<!ELEMENT domains (#PCDATA)>

<!ELEMENT variants (literal+)>
<!ELEMENT literal (#CDATA | examples? | usage_labels? | features? | info?)>
  <!ATTLIST literal
    lemma CDATA #REQUIRED
```

```
sense CDATA #REQUIRED
ewn_sense CDATA #IMPLIED
status CDATA #IMPLIED>

<!ELEMENT examples (#PCDATA)>
<!ELEMENT usage_labels #CDATA>
  <!ATTLIST usage_labels
    date CDATA #IMPLIED
    sub CDATA #IMPLIED
    reg CDATA #IMPLIED>

<!ELEMENT features #CDATA>
  <!ATTLIST features
    connotation CDATA #IMPLIED
    gender CDATA #IMPLIED
    collective CDATA #IMPLIED
    number CDATA #IMPLIED
    unerg CDATA #IMPLIED
    unacc CDATA #IMPLIED
    trans CDATA #IMPLIED
    intrans CDATA #IMPLIED>

<!ELEMENT info #CDATA>
  <!ATTLIST info
    author CDATA #IMPLIED
    date CDATA #IMPLIED
    site CDATA #IMPLIED
    comments CDATA #IMPLIED>

<!ELEMENT internal_links (relation+)>
<!ELEMENT relation (#CDATA | target_wm | features)>
  <!ATTLIST relation
    type CDATA #REQUIRED
    rel_id CDATA #IMPLIED
    inv_id CDATA #IMPLIED>

<!ELEMENT target_wm (#PCDATA)>
  <!ATTLIST target_wm
    id CDATA #REQUIRED
    part_of_speech CDATA #REQUIRED
    lemma CDATA #IMPLIED
    sense CDATA #IMPLIED
    source_variant CDATA #IMPLIED
    target_variant CDATA #IMPLIED>

<!ELEMENT features #CDATA>
  <!ATTLIST features
    conjunctive CDATA #IMPLIED
    disjunctive CDATA #IMPLIED
    reversed CDATA #IMPLIED
    negative CDATA #IMPLIED
    factive CDATA #IMPLIED
    non_factive CDATA #IMPLIED>

<!ELEMENT eq_links (#CDATA | relation+)>
```

**Figure 9:** Word\_Meaning DTD.

---

```

<WORD_MEANING ID="n#8" PART_OF_SPEECH="n">
  <GLOSS> figura geometrica generata da un rettangolo che
    ruota intorno a uno dei suoi lati. </GLOSS>
  <VARIANTS>
    <LITERAL LEMMA="cilindro" SENSE="1" EWN_SENSE="1"
STATUS="new"> </LITERAL>
  </VARIANTS>
  <INTERNAL_LINKS>
    <RELATION TYPE="has_hyperonym" REL_ID="IR000055"
INV_ID="IR000056">
      <TARGET_WM ID="n#12" PART_OF_SPEECH="n" LEMMA="solido"
SENSE="1"> </TARGET_WM>
    </RELATION>
  </INTERNAL_LINKS>
  <EQ_LINKS>
    <RELATION TYPE="eq_synonym" REL_ID="ER000008">
      <TARGET_WM ID="08482581" PART_OF_SPEECH="n"
LEMMA="solid" SENSE="3"> </TARGET_WM>
    </RELATION>
  </EQ_LINKS>
</WORD_MEANING>

```

---

**Figure 10:** Word\_meaning example.

### **VisDic XML Representation**

As mentioned before, VisDic XML representation was developed with regard to speed, efficiency and unique data representation preserving redundancy.

The initial step was to convert Polaris representation in import-export format to some XML representation. It can be done very easily. But resulting XML tree was very deep, and there were too many levels for processing to gain desiderative information. In order to reduce XML tree structure the specialized tool had to be prepared.

The next step is to make searching wordnet relations faster. The relation in Polaris format and also the relations in XML format presented by Italians are easily readable for human, but very complicated for machines. For example, if the machine wants to gain English hypernymical synset corresponding to the synset "being: 1", it can read these information in 4 tags represented internal relation, hypernym, literal and sense. Moreover, the corresponding synset must be found somehow according to the literal and sense. Our approach has only one tag, which is marked as the link tag. It says that this tag contains a key, which points to the corresponding synset. All other information can be retrieved from this link. This makes the search really fast.

This step is also provided by a special script. It can replace corresponding links automatically. Besides, it can find some type of errors in wordnet. The empty links pointing to nowhere and links pointing to the synset itself are reported.

The last two features, the efficiency and the unique data requirement looks slightly contradict at the first sight. For example, the hyponyms are not present in the dictionary, because they can be derived from hypernyms. Also the final size is significantly reduced by this (see Table 1). Searching these hyponyms must be done by means of hypernyms pointing to the corresponding synset. But the inner representation is adapted for this task and this type of search is also fast as the simple search.

It is important that the glosses can be stored only once, and other wordnets can contain external links to glosses. The comparison between original Polaris representation and the VisDic representation is shown in Figure 5.

The VisDic representation can specify lower and upper number of tags in the dictionaries. Its definition differs from classic DTD format. A VisDic definition for wordnet is in Figure 6. Every tag can be understood as a classic tag containing the plain text (N), or it can be signed as a key tag (K), which is unique for every synset, or it can be a link to other synsets (L), the link to specified tag in different dictionary (E) – especially glosses are represented by this technique, and finally the reverse links (R), which specified the relation derived from other one.

### Polaris format

```

0 @3@ WORD_MEANING
  1 PART_OF_SPEECH "n"
  1 VARIANTS
    2 LITERAL "life"
      3 SENSE 1
      3 DEFINITION "living things collectively; "there is no
life on Mars""
      3 EXTERNAL_INFO
        4 SOURCE_ID 1
          5 TEXT_KEY "00003504-n"
  1 INTERNAL_LINKS
    2 RELATION "has_hyperonym"
      3 TARGET_CONCEPT
        4 PART_OF_SPEECH "n"
        4 LITERAL "being"
          5 SENSE 1
    2 RELATION "has_hyponym"
      3 TARGET_CONCEPT
        4 PART_OF_SPEECH "n"
        4 LITERAL "wildlife"
          5 SENSE 1
  1 EQ_LINKS
    2 EQ_RELATION "eq_synonym"
      3 TARGET_ILI
        4 PART_OF_SPEECH "n"
        4 WORDNET_OFFSET 3504

```

### Final XML format

```

<SYNSET>
  <POS>n</POS>
  <SYNONYM>
    <LITERAL>life
      <SENSE>1</SENSE>
    </LITERAL>
  </SYNONYM>
  <ILI>00003504-n</ILI>
  <HYPERONYM>00002728-n</HYPERONYM>
</SYNSET>

```

**Figure 11:** Comparison of Polaris and XML record

### **VisDic Assessment**

As it can be seen from the other parts of this deliverable, VisDic together with its XML representation displays a relevant standardization power which is demonstrated by the unified handling of all Balkanet wordnet. In that respect Balkanet visibly surpasses the quality of the EuroWordNet data and helped to make them consistent and containing fewer errors. No detailed comparisons took place but it is not difficult to see this. At the moment VisDic is a recommended tool for editing and browsing wordnets worldwide – recommendation comes from GWA and can be found on its www pages ([www.globalwordnet.org](http://www.globalwordnet.org)).



## Objectives and Data Requirements

The main goals of the BalkaNet project are to build, in a concerted and harmonized way, aligned wordnets for six languages and to demonstrate their usefulness in real modern applications. A special emphasis was given from the very beginning of the project, on both quality issues and cross-lingual coverage across the monolingual wordnets. Except for the Czech wordnet, all the others started been built from scratch; however they have been supported by many monolingual and bilingual resources. From a certain point of view, this unbiased start-up facilitated the harmonized development of the envisaged wordnets, but on the other side of the spectrum, it raised additional problems imposed by the acquisition of the knowledge pertaining to the target common concepts. Moreover, the core interest in representing within monolingual wordnets a common set of concepts was doubled by the natural requirement that the wordnets should also represent the real language use (both within the monolingual and across the multilingual contexts) of the respective languages.

More precisely, the main goals that were set at the beginning of the project and carefully pursued were the following:

- g1) developing at least 8000 synsets per new language-specific wordnet, commonly selected so that even with this small size, the wordnets should be useful in real applications;
- g2) ensuring maximal interlingual overlap among the BalkaNet wordnets and compatibility with the wordnets developed in the EWN project;
- g3) building free software tools for the efficient management and exploitation of the multilingual semantic lexicon;
- g4) development of application demonstrators such as Word Sense Disambiguation (WSD), intelligent document indexing, cross-lingual Information Retrieval (CLIR), etc. In particular, a system for WSD has already been developed (see [Ion and Tufiş, 2004]), a prototype implementation of an intelligent document indexing formula has been delivered that also performs Multilingual IR.

In order to comply with these goals, the consortium adopted a series of design strategies out of which the most influential were the following:

- s1) the inter-lingual index (PWN 1.5) and the inter-lingual relations were defined the same way as in EWN; based on this decision was possible to select a set of common concepts to be implemented in each BalkaNet wordnet thus maximizing and controlling the cross-lingual coverage;
- s2) since the available language resources, useful in building the monolingual wordnets, were different for each partner, both in format and coverage, each team had to build their own acquisition, development and validation tools so that to make maximum use of the available data; however, because all the wordnets were supposed to be integrated into a single multilingual environment, a common XML format was agreed (see (Horák&Smrž, 2004) and (Smrž, 2004); this format is used by the BalkaNet multilingual viewer and editor VisDic (Pavelek and Pala, 2002). The most recent and more powerful version of VisDic is presented in (Horák and Smrž, 2004).
- s3) because of various improvements apparent in recent versions of PWN we decided to update consequently our inter-lingual index, so that the final BalkaNet multilingual database is based on the PWN 2.0. This is not really a departure from the EWN compatibility (based on PWN 1.5) since more than 90% of the mappings among different versions of PWN (1.5, 1.6, 1.7.2, 2.0) are done automatically.
- s4) to ensure quality control over the monolingual wordnets the consortium decided a set of validation tests, checking the syntactic and structural correctness (Smrž, 2004);

An initial step in the BalkaNet project was bringing the PWN in the same XML format to be followed by all the monolingual wordnets. The synset IDs were given unique values made up from the string "ENG1.5-" (to specify the used version) followed by a sequence of digits representing the offset in the original database of the respective synsets, followed by a tag denoting the part of speech of the literals in the encoded synsets. The BalkaNet initial ILI was represented by the codes representing ID values of the synsets in the XML version of the PWN 1.5<sup>1</sup>. The interlingual alignment is made explicit by assigning the ILI in all languages to the synsets that are equivalent to the PWN with the same ID value. The concepts to be implemented by all the monolingual wordnets, described in a following section, were specified in terms of the ILI codes from which every partner could visualize the associated synset in PWN. The commonly agreed set of concepts (BCS: BalkaNet Concept Set) was obligatory for each monolingual wordnet and it contains 8516 concepts. Besides BCS each partner has the autonomy to extend his/her own wordnet by selecting other ILI codes according to language specific criteria. However, since the monolingual selection criteria were similar, more than 8516 common concepts are found between different pairs of languages.

The actual implementation of the selected concepts was performed by each team according to their own judgements and lexical resources they had at their disposal. The synsets structuring was also left to the latitude of the development teams, the only restriction being that the set of possible semantic relations was the one defined in EWN. The names of lexical relations were sometimes modified either to identify a language specific manifestation of a general lexical pattern or to identify a language characteristic morpho-syntactic relation.

In the vast majority of cases the hierarchical structures in PWN (nouns and verbs) were preserved over the monolingual wordnets following the *principle of hierarchy preservation* (Tufiş and Cristea, 2002). The wordnets hierarchies are inheritance structures (more often than not a synset has only one direct parent) with lower meanings being specialisation of their ancestors.

During the monolingual wordnets development phase it became clear that some of the concepts in BCS selection are not lexicalized in some languages and vice-versa, several synsets created in some wordnets has no obvious ILI equivalent. In the first case, there were created empty synsets (called non-lexicalized synsets) in the wordnets for the languages that do not lexicalize the respective concepts. The non-lexicalized synsets are apparently redundantly preserved in the hierarchy but their purpose is to reflect the proper interlingual relation between the concept and the closest lexicalized synsets in the wordnet. This way, the complex interlingual relations (HAS-EQ-HYPONYM, HAS-EQ-HYPERONYM, etc.) were simulated using only the EQ-SYNONYM interlingual relation (which is the only one handled by VisDic).

The language specific synsets non-lexicalized in English (e.g., meanings describing local kinds of food) were manually added to the ILI with the prefix BILI and from there it could be linked to the synsets of other languages that have a similar lexicalized meaning.

BalkaNet's ILI is meant as a shared conceptual warehouse. To allow a straightforward manipulation of ILI's contents we classified ILI's concepts under broad conceptual domains that have been adopted from the Suggested Upper Merged Ontology (SUMO) thus inducing a conceptual tree-like structure. Such a classification enables the efficient maintenance of the ILI's thematically related concepts and contributes in dealing with the proliferation of ILI's concepts.

After each concerted development phase of the monolingual wordnets, they are normalized (see (Smrž, 2004)) for being loaded into VisDic, the BalkaNet multilingual browser and editor. VisDic has powerful editing facilities by means of which an expand model approach (cf. following section) in the development of a wordnet is strongly supported. Due to its browsing

---

<sup>1</sup> These IDs have been updated as the BalkaNet ILI was upgraded to different versions of the Princeton WordNet. Currently the latest PWN version, i.e., WN 2.0 is being used.

facilities, VisDic supports the merge or combined approaches as well, pinpointing alignment problems which can be easily corrected. As a multi-wordnet viewer it allows for synchronized search: the search is performed in a source wordnet but (via the ILI) the results are, (or can upon request) be displayed for all other target wordnets loaded within the editor. VisDic is an open source tool that currently runs under both Linux and Windows platforms. A detailed presentation of VisDic is given in (Horák and Smrž, 2004).

In addition to the multilingual wordnets editor and viewer, it has been implemented a storage, querying and browsing infrastructure, namely the Wordnet Management System (WMS) (Koutsoubos et al., 2004) which enables the efficient navigation within and across wordnets. A variety of services have been incorporated into the WMS, which enable the efficient navigation within wordnets' hierarchies and the retrieval of their information contents. The latest version of the Wordnet Management System is described in (Koutsoubos et al., 2004).

Besides checking the syntactic and structural correctness of the wordnets as they were developed, a more challenging validation process was conducted when the core monolingual wordnets were in a stable state. This is called the *semantic interlingual validation* (Ion&Tufiş, 2004) and checks, against a multilingual parallel corpus, how the synsets of each monolingual wordnet cover actual use of language and to what degree the established interlingual equivalences among synsets of different wordnets are supported by parallel human translations. This procedure has been already applied to the Romanian-English pairs of wordnets and a detailed presentation is provided in (Ion&Tufiş, 2004) and (Tufiş, Ion, Ide, 2004). The most recent results on semantic interlingual validation for the BalkaNet wordnets are described in the section „Cross-Lingual Validation Based on a Parallel Corpus” of this report. Ensuring an accurate inter-lingual alignment and a large cross-lingual coverage is essential for the performance of the final project's application, which envisages the incorporation of BalkaNet resource in an IR system.

The rationale for employing BalkaNet in IR tasks is that a structured conceptual representation of the domain of interest, linked to multilingual wordnets would contribute in helping users of IR systems to find the required information in a more precise way and, very important, by using in the queries keywords of their own language. Due to the growth of the digital data that is being distributed over the Web, it was chosen to incorporate BalkaNet in a Web search engine in order to enable a more meaningful organization of the data sources that are indexed by Web IR systems. Specifically, the main task that the BalkaNet ontology is called to carry out is to index Web documents on the basis of their conceptual relatedness, i.e., conceptual indexing. Towards the project's application, we have developed a prototype Web search engine that currently indexes approximately 410K multilingual Web documents. These documents are organized into conceptual clusters by means of the conceptual knowledge encoded within BalkaNet's taxonomies and ILI's conceptual domains. It is expected that a semantically structured index of Web data sources will improve the engine's searching mechanisms to retrieve high quality search results. Currently a prototype conceptual indexing infrastructure has been developed that exploits in the most efficient way the information encoded within BalkaNet, in order to conceptually classify Web documents. Besides the indexing framework, a query expansion module has been implemented and it has been incorporated in the search engine. Query expansion enables both monolingual and cross-lingual expansion of query terms with superficially distinct but semantically related words.

## **Lexical Data Acquisition**

In the EWN project there were defined two distinct development models namely, the expand and the merge model.

The expand model essentially is a translation-driven wordnet development approach, in which the literals in each PWN synset are being translated as faithful as possible. The relations of the translated synset are to a large extent automatically imported and the original gloss is translated into the target language. During this process, some new literals could be inserted in

the target synset or some literals in the source synset hard to translate are ignored. In such an approach a high quality bilingual machine readable dictionary (MRD) can speed up dramatically the development in competition with a bilingual human lexicographer. Such an approach was strongly supported by the VisDic multilingual editor and browser.

The merge model assumes availability of monolingual structured language resources in machine readable form. The format of these resources has to be transformed into a wordnet compatible format and the meanings in the target language must be linked to the concepts in the interlingual index. The topology of the target wordnet could be in principle different from the topology of PWN, but the name of the semantic relations should be the same. In such an approach the required resources are either a monolingual thesaurus of comparable granularity to PWN or various MRD dictionaries (explanatory dictionaries, synonym dictionaries, antonym dictionaries, phrasal dictionaries, valency dictionaries, etc.) out of which a wordnet-like structure could be created. Integrating these resources into a coherent acquisition and development environment requires tools aware of the different encoding structures for the supporting resources as well as the output encoding representation.

Depending on the lexical resources the partners had at their disposal, the individual wordnets, development approaches came closer to one or the other development models. However none of the partners adopted a pure expand or merge model.

The papers in this volume describing each monolingual wordnet provide more in-depth details on the language resources used in the respective wordnets development and, where necessary, the tools<sup>2</sup> that were developed for the purpose of this endeavour.

### **Selecting BalkaNet Base Concepts**

There have been reported two main approaches in building multilingual semantic networks, namely, the expand and the merge model approaches (Vossen, 1996). The first one concerns the translation of a core set of synsets to other languages following their equivalence links, whereas the second one implies the independent development of each wordnet on the basis of the available monolingual lexical resources and the subsequent linking of the monolingual synsets<sup>7</sup> to their semantic equivalents in other wordnets. Both models exhibit advantages and disadvantages. To benefit from the advantages offered by both models it was decided to develop BalkaNet by using a combination of both models. This way it is reassured that the monolingual wordnets are richly encoded and comparable across languages (guaranteed by the expand model), while at the same time language-specific properties are reflected into the monolingual wordnets (guaranteed by the merge model).

To achieve the linguistic realisation of the common concepts in all wordnets a three steps procedure was adopted resulting in a series of sets of ILI codes (BCS1, BCS2 and BCS3).

The first BalkaNet Concept Set (BCS1) was identical to the EWN Base Concept set. Base Concepts were selected for reasons convincingly argued in (Rodriguez et al., 1998) and they represent concepts that are lexicalized in all the languages represented in the BalkaNet resource. In the present versions of the BalkaNet ILI, the number of BCS1 concepts is 1218<sup>3</sup>. As in EWN, all concepts in BCS1 have attached a Top Ontology description (cf. Vossen, 1998). For the selection of BCS2 and BCS3 the concepts which were lexicalized in most languages represented in EWN were taken into account. This statistical information was provided by MEMODATA. Also, each partner suggested a list of candidate concepts that would be relevant for their languages. The concepts proposed by at least two partners were also con-

---

<sup>2</sup> A very detailed presentation of the tools can be found in the project's Delivery D3.1 "Design and Development of Tools for the Construction of the Monolingual Wordnets", June 2002. These tools and their user manuals are also available on the project's site.

<sup>3</sup> This is slightly different from the BC in EWN: 1024 (representing 796 nominal concepts and 228 verbal concepts). The difference is due to finer grained synsets of PWN2.0 (BalkaNet ILI) as compared to PWN1.5 (EWN ILI).

sidered as candidates for the common set of concepts. An additional selection criterion was that the concepts in BCS1, BCS2 and BCS3 should correspond to dense sub-networks in PWN. We called this selection restriction the *conceptual density* criterion and it can be stated as follows:

- a) once a nominal or verbal concept (i.e. an ILI concept that in PWN is realized as a synset of nouns or as a synset of verbs) was selected in the BCS, all its direct and indirect ancestors (i.e. all ILI concepts corresponding to the PWN synsets, up to the unique beginners) will be also included into BCS.
- b) all the descriptive adjectival concepts (i.e. ILI concepts that in PWN are realized as synsets of descriptive adjectives) included in BCS should represent values of attributes named by nouns already presented in the chosen set of BCS. This relationship is encoded in PWN by the relation *be-in state*.

A detailed description of the BCS selection process is given in (BKN-D.4.2, 2003) but the figures have changed due to migration from PWN1.7.1 to PWN2.0. Currently, the set of BCS (1,2 and 3) consists of 8,516 concepts implemented in all but one of the monolingual wordnets. The exception is the wordnet of the Serbian subcontractor which implemented 6057 concepts of the BCS, which is 4 times more than planned and a total set of 8059 synsets, which accounts to 5 times more than initially planned. These numbers are really far better than what was initially envisaged given that the Serbian partners joined the consortium in a later phase of the project.

Because the conceptual density criterion operates only on nominal, verbal and (descriptive) adjectival synsets, the BCS includes concepts that correspond neither to adverbial synsets nor to relational adjectives synsets. The selection of these categories of synsets was left in the responsibility of each partner. Each monolingual wordnet has been further extended beyond covering the BCS. In general, the wordnets enrichment process followed a top-down approach starting with the synsets that have been already mapped onto the BCS. The monolingual wordnets extension was mainly guided by language-specific criteria in order to make sure that in spite of the ILI guided development of the first aligned synsets, the lexical distributional properties in each language were not overlooked. Each team responsible for their own language wordnet made various statistical studies on large corpora or used existing lexical resources to identify frequently occurring general words or word senses that should be included into the respective wordnets. During the concerted development of the BalkaNet wordnets, several quality control policies have been adopted and implemented by each wordnet developer. Yet, an overall quality control was performed by one of the partners. The methodology and the evaluation results are largely described in (Smrž, 2004).

Some extended monolingual wordnets included synsets that either represent concepts specific to some Eastern European area, or they automatically derive because of their regular morphological patterns and their easy to predict semantics. The analysis of the latter monolingual synsets in a multilingual context seems to open very interesting pathways. A lexical relation found in one language (by means of a derivative analysis), the semantics of which is predictable might be relevant as a paradigmatic relation in many non-derivative languages. For instance in Turkish two derivationally related words might correspond in Greek to two morphologically unrelated words. However, the semantics of the Turkish derivative affix could be very useful in assigning between the two Greek words a semantic relation, as for instance a case relation, cf., (Bilgin et al, 2004).

While developing language-specific synsets, the need for encoding language-specific relations emerged. Such relations are embedded within monolingual wordnets and they concern inter alia *XX-derivative* (where XX stands for the ISO code of the language), *usage\_domain*, *region\_domain* and so forth.

## **Extending BalkaNet Base Concepts**

One of the main characteristics that hold from very beginning of BalkaNet is the focus on large-scale overlap between monolingual wordnets, in order to maximize the applicability of the created database as a whole. A special set of synsets --- BCs (BalkaNet Common Synsets) has been chosen and all partners agreed on the schedule of the gradual development. Several criteria have been adopted in the BCS selection process, which has taken the following steps:

- Representation of the EWN Base Concepts to maximize the overlap between the two projects.
- Incorporation of consortium's proposal corresponding to the most frequent words in corpora and in various dictionary definitions for their particular languages.
- As an additional criterion, several noun synsets that had many semantic relations in the Princeton WordNet database have been added.
- All the selected synsets based on PWN 1.5 have been automatically mapped to PWN 2.0, which is currently the version BalkaNet is connected to. The synsets that found one-to-one correspondence in the new version have been finally chosen.
- All the hypernyms and holonyms of the chosen synsets have been added to BCS as it was decided to close the set in this respect.

Synsets are formed by true context synonyms as well as variants (typographic, regional, style, register ...) in the BalkaNet wordnets and have all been linked to Princeton WordNet (PWN). Moreover, verb synsets contain literals linked by a rich set of relations, e.g. aspect opposition and iteration.

## **Restructuring BalkaNet Interlingual Index**

An important achievement of the BalkaNet consortium was the upgrading of the ILI records from Princeton WordNet 1.5 to 1.7.1 and eventually to the most recent version WordNet 2.0. While ILI updates several issues have been tackled mainly concerning inconsistent mappings across the different versions. Semi-automatic techniques as well as manual correction of conflicting cases have been carried out to reassure the quality and accuracy of BalkaNet's ILI. The mapping between different versions of PWN was specified in terms of pairs of synsets offsets and was deterministic (one to one) in the vast majority of cases. The few cases where some ambiguities (one to many) persisted, the best choices were tackled manually by the wordnets' developers. Delivering a *fresh* and structured ILI has been one of the most important contributions of the BalkaNet project up to the reporting period. These techniques essentially concerned removal or improvement of any structural elements from monolingual wordnets that were not well formed and showed inconsistencies across wordnets. Some issues pertaining to language particularities have been also adopted by each contractor separately depending on the respective wordnet's structure and language phenomena.

However, in order to make our ILI even more powerful in the context of NLP applications and to facilitate the usage of our resource it was recently decided to further improve ILI's structure by incorporating an additional layer of semantic information to its contents. The additional knowledge added to the ILI concepts is imported from an upper-level ontology, namely the Suggested Upper Merge Ontology (SUMO). SUMO is an ontology that was created by merging publicly available ontological content (Niles and Pease, 2002 (b)) into a single structure and provides definitions for general-purpose terms. SUMO acts as a foundation for more specific domain ontologies and was employed in order to organize ILI's conceptual taxonomies under broad conceptual domains, improving thus the manipulation of the ILI in the context of wordnets' comparison and navigation. At present, BalkaNet's ILI (BILI) is a multilingual structured conceptual ontology that can be employed by a variety of applications without imposing any need for structural changes. Moreover, BalkaNet's structured Interlin-

gua gives the flexibility to incorporate new concepts and/or link new languages to it, whereas it enables the retrieval of meaningful semantic data across different languages

Besides ILI updates, the BalkaNet consortium has also incorporated within BalkaNet ILI domain specific synsets from the thematic categories of Law, Politics and Economy, which have been selected for the purposes of the project's experimental application. i.e., classify Web documents of the above three domains. These domains have been selected from the Balkan-Times website which will be used as the central repository that feeds the engine's index with web documents. Each members of the consortium has incorporated a total set of approximately ~100 common predetermined synsets from each of these three domains. BalkaNet's domain knowledge information originated mainly from the following resources:

- 1) The mapping from WordNet to the SUMO (Suggested Upper Merged) Ontology.
- 2) WordNet Domains 1.0 (Database) developed by Istituto Trentino di Cultura (ITC).

The first resource is in the public domain. It contains SUMO domain labels for 17,453 adjectives, 3,101 adverbs, 65,636 nouns and 11,793 verbs. The second resource is not in the public domain and individual licenses have been obtained from ITC. It assigns every PWN 1.6 synset to one of the 165 domains which are arranged in a special hierarchy. Although all PWN synsets are assigned to one of the domains, 32.154 synsets are assigned to the domain "factotum", which shows that the synset in question does not belong to any special domain. The approach adopted in order for these domain labels to be incorporated within BalkaNet resource concerned the inclusion of domain in the ILI level rather than in the monolingual wordnets. The detailed domains' incorporation approach is the following: Once a synset belonging to one of the three pre-specified domains is traced, the starting and ending nodes of its taxonomy will be marked with the domain label information using the RELATED\_TO lexical relation. All nodes that belong to the path and are between the starting and ending node will inherit the domain information thanks to the transitivity of the IS\_A relation.

### ***Re-linking BalkaNet ILI to PWN 2.0***

One significant achievement of the consortium was moving from Princeton WordNet 1.7.1 to the most recent version WordNet 2.0. As the previous upgrade (from Princeton WordNet1.5 to Princeton WordNet1.7.1) this step assumed applying a set of mapping rules and in some cases, where the mapping was not deterministic, manual mapping. The scripts for these conversions have been developed and applied by all partners. We have prepared a recommended procedure for re-linking synsets previously linked to WordNet 1.5 in VisDic. The aim is to facilitate the process of transition to WordNet 1.7.1.:

1. Start with four VisDic windows, first two containing the same copy of our national wordnet, the third - English WordNet 1.7.1 and the fourth English WordNet 1.5. Click on tab "WN15" in the second window. Push the right mouse button in the second window, choose "AutoLookUp in" and set it to the first window. Identically set "AutoLookUp by MAPHINT in" to English WordNet 1.7.1 and "AutoLookUp by REVMAP in" to English WordNet 1.5 all in the second window. The last step is demonstrated in the following figure.

All synsets that should be re-linked will be listed in the second window all the time. The first will be used for editing. The third will suggest synsets from English WordNet 1.7.1 that should be considered as equivalent to the current synset in the second window. The fourth window presents all synsets from English WordNet 1.5 that could be transformed into those in the third window.

2. Usually, some synsets could not be linked to WN 1.7.1 automatically. The reason is that the English equivalent synset in WN 1.5 - "threshold:1, limit:2" has been splitted to

two different synsets in WN 1.7.1 ("terminus ad quem:1, terminal point:1, limit:2" and "threshold:1").

3. As we want to link the synset in our language to the correct one in English WN 1.7.1 we choose one of the synsets in English WN 1.7.1, use function "take key from 1.7.1" to link the processed synset to its equivalent in English WN 1.7.1.

4. Then we should modify the current synset to exactly correspond to the English equivalent, e.g. delete some literals, modify gloss etc.

5. The best next step is the definition of national language equivalents for all other synsets in the third window (synsets deleted in the previous step will form them usually). Choose one of them, copy it to your wordnet (to the first window) by means of function "Copy entry to", modify what is needed in tag "Edit" of the first window and finally click button "Update" to save changes.

6. Sometimes, the problem with automatic linking was not that one synset from WN 1.5 could be linked to more than one in WN 1.7.1 but, vice versa, that synsets from WN 1.5 has been joined to form one synset in WN 1.7.1. Such a situation is demonstrated in the following figure. Synsets "mogul:1" and "baron:3, big businessman:1, business leader:1, king:3, magnate:1, power:8, top executive:1, tycoon:1" from WN 1.5 has been joined to form "baron:3, big businessman:1, business leader:1, king:3, magnate:1, mogul:2, power:9, top executive:1, tycoon:1" in WN 1.7.1. Usually, only one of such synsets from WN 1.5 has been linked (it can be checked by means of function "Show in") so it is sufficient to take a key from the synset in WN 1.7.1 and perhaps to add other literals to the current synset.

All the synsets in "WN15" tag should be re-link to WN 1.7.1. Finally, tag "WN15" should remain empty (after the restart of VisDic).

## ***BalkaNet Specific-Concepts***

There exist significant historical, social and cultural links between the Balkan languages represented in the BalkaNet project. All languages in the project have significantly influenced each other and there are several concepts that are specific to this region of the world. With these considerations in mind, the BalkaNet consortium made an attempt to incorporate these shared concepts in the wordnets of the respective languages in a systematic way. Below, we summarize the procedure that has been adopted and the results obtained.

### PROCEDURE

- Each partner worked separately to prepare a set of concepts which it thought was specific to that language. One semi-automatic way of doing this is to extract lexical gaps from a machine-readable bilingual dictionary. Dictionary entries that do not have a one-word or two-word equivalent in the target language and are explained by paraphrases can be easily extracted automatically and a high proportion of these are language-specific concepts. Another method is to prepare a survey for native speakers and ask them to write down what they think could be specific to their language, also providing some domains that are rich in language-specific concepts to facilitate the process. In the BalkaNet project both of these methods were used and the results were quite successful.
- Since the BalkaNet project uses Princeton WordNet (PWN) 2.0 synsets as its common pool of concepts, it should be checked before adding a language-specific synset to a Balkan wordnet if it already exists in PWN 2.0. For example, 'baklava' (a kind of pastry) already exists in PWN 2.0 although this was a very likely candidate for a "language-specific" synset for almost all the partners.



- Having confirmed that the concept does not exist in PWN 2.0, the synset was created as a language-specific synset and assigned a Balkan-specific code in the form BUL-xxxxxxx, GRE-xxxxxxx, etc.
- The following were identified as fruitful domains that contain lots of language-specific concepts:
  - Administrative system (institutions, officers)
  - Family relations
  - Religious objects
  - Religious practices
  - Wedding traditions
  - Architecture (buildings, parts of buildings, styles)
  - Food and food ingredients
  - Animals, plants, fish
  - Traditional clothes
  - Traditional occupations
  - Traditional arts, handicrafts
  - Traditional music (genres, dances, instruments)
  - Tools (special types of scissors, knives, cooking utensils, farming equipment etc.)
  - Measurement units
  - Important events (national holidays, wars, treaties, elections)
- Some partners have also provided pictures together with their synsets. Since all language-specific concepts were going to be merged into a single Balkan concepts database with the contribution of each partner, the pictures helped the partners to exactly understand the concepts proposed by the others.
- A gloss in English was provided for every language-specific synset. These glosses allowed the partners to check the other partners' language-specific concepts and to decide if the same concept exists in their language too.
- Every language-specific concept was linked to other synsets via semantic relations. Hypernymy was an obligatory relation, while antonymy and meronymy were also frequently used.

## RESULTS

The numbers of language-specific synsets added by each partner are as follows:

### ***The BalkaNet Inter-Lingual-Index***

After each partner developed its own set of language-specific concepts as described above, each team checked the synsets of all the other teams to see if some of these concepts were lexicalized in their own language too. The aim was to find the intersection between the concepts proposed by the partners and to merge the concepts into a single, unified database to be called the BILI (Balkanet Inter-Lingual Index).

This comparison process also helped the teams include synsets they had forgotten to include in their wordnets in the previous development stage but were included by one or more of the other teams.

### CONFLICTS

Since every team checked the concepts of every other team and made a claim that Concept X in Language A is the same as Concept Y in Language B, there was the possibility of conflicting claims. For example there would be a conflict if the Serbian team claimed that ROM-252 was the same as SRP-133, while the Romanian team claimed that SRP-133 was ROM-155.

Moreover, it was possible to have conflicts that arise from a chain of claims involving more than two languages. This would be the case for instance if the Turkish team claimed that GRE-12 was the same as TUR-17, the Serbian team claimed that TUR-17 was the same as SRP-32 and the Greek team claimed that SRP-32 was the same as GRE-34.

Considering the possibility of conflicts, the claims made by each team were collected by the Turkish team and an algorithm was developed to detect them. 17 conflicting cases were detected as a result, which were sent to the concerned teams who were going to discuss the validity of the claims that caused this conflict and reach an agreement.

In most of the cases, the conflict was caused by an obviously wrong association of two different concepts. In some other cases, the concepts involved in the conflict were very close to each other and it was quite natural for such a situation to occur.

### BILI NUMBERS

After the conflicts were resolved by way of discussion between the partners, the final list of Balkan-specific concepts was prepared and each concept was assigned a unique BILI number in the form BILI-xxxxxxx, starting with BILI-00000001. These BILI numbers have been used by all partners in the final versions of their wordnets.

### RESULTS

The following table shows the distribution of BILI entries in the final BalkaNet wordnets:

## ***Adding Domain-Specific Concepts***

The final application of the Balkanet project is a search engine which will undertake conceptual indexing of web documents and multilingual query expansion. Three domains have been selected for the purposes of this experiment, namely law, politics and economy. These domains have been selected from the BalkanTimes website which will be used as the central repository that will feed the engine's index with web documents. Each team is going to add to its wordnet 100 predetermined synsets from each of these three domains.

The experimental search engine also requires preprocessing of the documents to be queried. Therefore, each team has assigned POS-tags and provided lemmas for a predetermined set of texts.

### ***Relations to SUMO and MILO Ontologies in VisDic***

SUMO (Suggested Upper Merged Ontology) is being created as part of the IEEE Standard Upper Ontology Working Group. An upper ontology is limited to concepts that are meta, generic, abstract or philosophical, and hence are general enough to address (at a high level) a broad range of domain areas. Concepts specific to particular domains are not included in an upper ontology, but such an ontology does provide a structure upon which ontologies for specific domains (e.g. medicine, finance, engineering, etc.) can be constructed. MILO (Mid-Level Ontology) is intended to act as a bridge between the high-level abstractions of the SUMO and the low-level detail of the domain ontologies.

The main part of relations from SUMO and MILO has been converted to a “dictionary” that can be browsed in VisDic. It is a union of SUMO and MILO containing all the concepts together with subclass, instance, subRelation and subAttribute relations.

The respective SUMO concepts are added to all Princeton wordnet synsets in attribute SUMO. Czech team has prepared the mapping from PWN 2.0 to SUMO under VisDic.

It is possible to look up for the concept in the “SUMO & MILO“ dictionary and view the ontology tree of the concept.

### ***Selecting the SUMO Domain-Specific Concepts***

The procedure followed in order to decide which of the SUMO categories are conceptually closer to our domains of interest, i.e. politics, economy and law has as follows:

At first place we run through the ontology tree of SUMO <http://virtual.cvut.cz/kifb/en/toc/all.html> and the concepts that shared the same name with the domain in their labels, were located i.e. political organization. Secondly, it has been examined whether the super classes/subclasses of these concepts had a relevant name to any of the predefined domains or whether they were semantically connected. In cases where the super or sub classes were not helpful, the coordinate terms of the concept in question, were taken into consideration on the basis of name and/or sense as well i.e. corporation, educational organization, religious organization. However, some times there was a need for consulting the related wordnet synsets of each concept; in such a case we run through the wordnet synsets' list as grouped by the SUMO site by examining their literals and glosses.

This procedure has been delivered by different people. At the end, the suggestions of each one have been merged and resulted in the following list.

#### **Domains**

##### **1. POLITICS**

- geopolitical area
- nation
- state or province
- city
- political organization
- government
- political process
- citizen

##### **2. LAW**

- law
- legal action
- regulatory process
- ordering
- obligation
- normative attribute
- contract
- purchase contract
- service contract

### 3.ECONOMY

- corporation
- transaction
- financial transaction
- lending
- borrowing
- increasing
- decreasing
- currency measure (us dollar, united states cent, euro cent, euro dollar)
- monetary value
- advertising
- betting
- buying
- selling

## Qualitative BalkaNet evaluation

The quality control of the BalkaNet wordnets was a major task in this project. Quality control concerned two main issues, namely validating the quality of the contents and structure of each monolingual wordnet on the one hand, and validating the quality and contents of each wordnet in comparison to the other wordnets within BalkaNet. Following this objective, both monolingual and cross-lingual quality control tasks were carried out. Besides the validation tests developed by each partner for their own wordnets, centralized validations and evaluations were also performed. Herein, issues pertaining to cross-lingual wordnets validation are highlighted. Since all the wordnets are XML encoded an obvious general test was conformance with the BalkaNet DTD (Document Type Description). Some other tests, also syntactic in nature, referred to wordnets prescribed structure. Examples of such tests are: identifying duplicated literals in a synset, checking if each literal of any synset has assigned a sense label, checking if all concepts in the BCS have a linked synset in each of the wordnets, checking for conceptual density of each wordnet (no dangling nodes or relations), checking for loops in the wordnets, etc. A web implementation of these tests and several others has been also implemented (by the DCMB team) so that each partner could cross-check his/her own validation.

The results of the centralized validation tests were communicated to all partners for corrective actions. The continuous interaction on the validation issues between partners resulted in a quality control methodology, which was implemented in various versions. Syntactic validation methods say very few about the quality of the synsets and the accuracy of the ILI-based cross-lingual alignment. This is a very thorny issue and there is no generally accepted methodology in the wordnet community. EWN project has also been rather elusive on these aspects.

The approach BalkaNet consortium adopted was to exploit recent developments in the parallel corpus technology. A text translated (by professionals) into several languages should be an ideal test-bed for cross-lingual validation of aligned wordnets. The basic intuition underlying this approach is that words that are used as reciprocal translations in the parallel texts should be also retrieved in the respective wordnets (via ILI) as translation equivalents. In order to transform this intuition into an operational validation method, it was decided to use the “Ninety Eighty Four” parallel corpus, based on the famous novel of George Orwell. This corpus, developed during the European project “Multext-East” contained already four of the seven languages of interest in Balkanet (Bulgarian, Czech, English and Romanian). The Greek, Serbian and Turkish partners prepared the respective language translations in the required format for being included into the parallel corpus, rising to 10 the number of languages represented in this unique multilingual corpus. The corpus is sentence aligned and part-of-speech (POS) tagged in all languages and the tagging of six of the translations has been carefully hand validated. A second step towards semantic validation was to select a bag of English words present in the original version of Orwell’s “Ninety Eighty Four” the senses of which were expected to be retrieved in the BalkaNet wordnets. To this end, there were selected from all the English nouns and verbs occurring in the corpus, only those that belonged to synsets (corresponding to concepts) that were in the BCS selection and therefore presumably aligned to synsets in all the BalkaNet wordnets. The resulted set contained 733 words out of which only 209 had at least two senses. These words occurred altogether 1621 times not always translated in every language present in the parallel corpus. All the partners received the list of the 209 ambiguous English words, to be used in the cross-lingual validation of their Wordnet against the ILI (PWN2.0). One of the Romanian partners developed a Word Sense Disambiguation platform called WSDtool incorporating a highly accurate word aligner in parallel corpora. The results of this validation are shown in the section “Cross-Lingual Validation Based on a Parallel Corpus”.

## **Tests for Monolingual Wordnets and Quantitative Comparisons**

One significant achievement of the consortium since the last report was moving from Princeton WordNet 1.7.1 to the most recent version WordNet 2.0. As the previous upgrade (from Princeton WordNet1.5 to Princeton WordNet1.7.1) this step assumed applying a set of mapping rules and in some cases, where the mapping was not deterministic, manual mapping.

Based on the consortium consultation we designed a set of formal general constraints that every wordnet was expected to observe. The constraints were implemented as a set of tests and each partner applied them and worked towards removing or correcting all the structural elements of their wordnets that did not observe the rules of well-formedness. A couple of other language specific restrictions have been proposed and implemented by some partners.

The first quantitative evaluation, namely the number of the synsets and their part-of-speech distribution as compared with the specifications in the Technical Annex, showed that the consortium achieved more than it was promised.

The quantitative comparisons among the well-formed wordnets were meant to give an overall evaluation of the cross-lingual coverage and to this end we computed intersections among the cross-linked synsets in all languages.

A better indication of the quality and compatibility will be given by comparing the consistency of the interlinked wordnets against a parallel corpus. The comparison of the wordnets will be based on the equivalence relations to the EuroWordNet ILI records and the translation equivalence relations as featured by the parallel corpus.

## **Evaluating the Well-Formedness of Balkan Wordnets**

The following objectives have been set and successfully accomplished while evaluating Balkan wordnets quality:

1. XML well-formedness of the wordnets (compliant with the VISDIC format).
2. Literals and sense ids: this was probably one of the hardest issues so solve. The easy part was to ensure that all the literals in any synset were already assigned a sense identifier. Also it was easy to check that no identical literals (irrespective of the sense labels) belonged to the same synset. The single conceptual restriction was that the combination literal + sense identifier should be unique. Since our implemented wordnets were cantered on a subset of senses in PWN it was unavoidable to have words in the target wordnets for which only some of the senses were considered.
3. IDs validation (the synsets were labelled with valid unique IDs)
4. POS validation: the synsets were tagged only with one of the 4 categories n, v, a, b)
5. Internal relations validation (no duplicates, relations belonging to the standard set of relations, no loops)
6. network density validation (no dangling synsets or relations);
  - i. an existing synset which has no hypernym was mapped to an ILI that in PWN is a topmost synset (such as unique beginners for the noun hierarchy); otherwise it was a dangling node;
  - ii. an existing (binary) relation which misses either of the two synsets it is supposed to connect is considered a dangling relation if the missing synset would correspond to an ILI in the commonly agreed set. Otherwise it is not and it should be deleted.
7. glosses validation (no empty definitions, spellchecking, definition in the own language)

8. senses validation (no literal with the same sense label should appear in more than one synset)

## Structure of a Wordnet File

Each WordNet has been built by the Wordnet Management System and is then stored in one individual XML file. To handle the different languages, these XML files will use the Unicode Charset (UTF8 ). Those are the files used by VisDic.

Here is an extract of the Romanian Wordnet XML file :

```
<SYNSET><ID>ENG20-00004609-n</ID><POS>n</POS>
<SYNONYM><LITERAL>viață<SENSE>1</SENSE></LITERAL></SYNONYM>
<DEF>forme de viață, văzute în mod global; "Nu există viață pe
Marte"</DEF>
<STAMP>Dan Cristea</STAMP>
<BCS>1</BCS>
<ILR>ENG20-00003009-n<TYPE>hypernym</TYPE></ILR>
</SYNSET>
<SYNSET><ID>ENG20-00004824-n</ID><POS>n</POS>
<SYNONYM><LITERAL>celulă<SENSE>1</SENSE></LITERAL></SYNONYM>
<DEF>Element constitutiv fundamental al organismelor vii, alcătuit
din membrană, citoplasmă și nucleu, reprezentând cea mai simplă uni-
tate anatomică.</DEF>
<STAMP>Dan Cristea</STAMP><BCS>1</BCS>
<ILR>ENG20-00003009-n<TYPE>hypernym</TYPE></ILR>
<ILR>ENG20-00003226-n<TYPE>holo_part</TYPE></ILR>
<ILR>ENG20-05681603-n<TYPE>category_domain</TYPE></ILR></SYNSET>
```

Description of the different tags :

- SYNSET : contains all the data relative to Synset.
- ID : identifier of the ILI. The prefix ENG20 means that it had been created by the Princeton WordNet, version 2.0, while the prefix BILI means that the synset is a BalkaNet specific one.
- POS : part of speech. The possible values are :
  - o n : noun
  - o v : verbe
  - o b : adverb
  - o a : adjective
- SYNONYM : list of the literals of this synset. At least one literal is mandatory.
  - o LITERAL : wording of the literal
  - o SENSE : number used for the sense differentiation.
  - o LNOTE : note about this literal
- Def : gloss of the synset. This wording allows to describe the synset. It's not mandatory.
- STAMP : gives some additional information about this synset : author, date...
- USAGE : gives an example of use of the synset
- BCS : number of the base concept associated with this synset. The possible values are 1, 2 or 3.
- ILR : Interlingua relation. Gives a relation between this synset and the specified Ili.

TYPE : type of this relation. The possible values are : be\_in\_state, category\_domain, causes, derived, eng\_derivative, holo\_member, holo\_part, holo\_portion, hypernym, near\_antonym, particle, region\_domain, similar\_to, subevent, usage\_domain, verb\_group

## Cross-Lingual Validation Based on a Parallel Corpus

If we take the position according to which word senses (language specific) represent language independent meanings, abstracted by ILI records, then the evaluation procedure of wordnets interlingual alignment becomes straightforward: in a parallel text, words which are used to translate each other should have among their senses at least one pointing to the same ILI or to closely related ILIs. However, both in EuroWordNet and BalkaNet the ILI records are not structured, so we need to clarify what “closely related ILI” means. In the context of this research, we assume that the *hierarchy preservation* principle (Tufiş & Cristea, 2004) holds true. This principle may be stated as follows:

*if in the language  $L_1$  two synsets  $M_1^{L_1}$  and  $M_2^{L_1}$  are linked by a (transitive) hierarchical relation  $H$ , that is  $M_1^{L_1} H^m M_2^{L_1}$  and if  $M_1^{L_1}$  is aligned to the synset  $N_1^{L_2}$  and  $M_2^{L_1}$  is aligned to  $N_2^{L_2}$  of the language  $L_2$  then  $N_1^{L_2} H^m N_2^{L_2}$  even if  $n \neq m$  (chains of the  $H$  relation in the two languages could be of different lengths). The difference in lengths could be induced by the existence of meanings in the chain of language  $L_1$  which are not lexicalized in language  $L_2$ .*

Under this assumption, we define the *relatedness* of two ILI records  $R_1$  and  $R_2$  as the *semantic similarity* between the synsets  $Syn_1$  and  $Syn_2$  of PWN that correspond to  $R_1$  and  $R_2$ . A semantic similarity function  $SYM(Syn_1, Syn_2)$  could be defined in many ways. We used a very simple and effective one:  $SYM(Syn_1, Syn_2) = \frac{1}{1+N}$  where  $N$  is the number of oriented links

traversed from one synset to the other or from the two synsets up to the closest common ancestor. One should note that every synset is linked (EQ-SYN) to exactly one ILI and that no two different synsets have the same ILI assigned to them. Furthermore, two ILI records  $R_1$  and  $R_2$  will be considered closely related if *semantic-similarity*  $(Syn_1, Syn_2) \geq k$ , where  $k$  is an empirical threshold, depending on the monolingual wordnets and on the measure used for evaluating semantic distance.

Having a parallel corpus, containing texts in  $k+1$  languages ( $T, L_1, L_2 \dots L_k$ ) and having monolingual wordnets for all of them, interlinked via an ILI-like structure, let us call  $T$  the target language and  $L_1, L_2 \dots L_k$  as source languages. The parallel corpus is encoded as a sequence of *translation units* (TU). A translation unit contains aligned sentences from each language, with tokens tagged and lemmatized as exemplified below (for details on encoding see <http://nl.ijs.si/ME/V2/msd/html/>):

```
<tu id="Ozz.113">
  <seg lang="en">
    <s id="Oen.1.1.24.2"><w lemma="Winston" ana="Np">Winston</w>
      <w lemma="be" ana="Vais3s">was</w>      ... </s>
  </seg>
  <seg lang="ro">
    <s id="Oro.1.2.23.2"><w lemma="Winston" ana="Np">Winston</w>
      <w lemma="fi" ana="Vmii3s">era</w>      ... </s>
  </seg>
  <seg lang="cs">
    <s id="Ocs.1.1.24.2"><w lemma="Winston" ana="Np">Winston</w>
      <w lemma="se" ana="Px---d--ypn--n">si</w>      ... </s>
  </seg>
  . . .
</tu>
```

We will refer to the wordnet for the target language as T-wordnet and to the one for the language  $L_i$  as the  $i$ -wordnet. We use the following notations:

$T\_word$  = a target word, say  $w_{TL}$ ;

$T\_word_j$  = the  $j$ -th occurrence of the target word;



$eq_{ij}$  = the translation equivalent (TE) for  $T\_word_i$  in the source language  $L_j$ , say  $w_{SL_j}$ ; a pair  $(w_{TL}, w_{SL})$  so that in a given context (a translation unit)  $w_{TL}$  and  $w_{SL}$  are reciprocal translations is called a translation pair (for the languages considered);

EQ = the matrix containing translations of the  $T\_word$  (n occurrences, k languages):

|        | $L_1$     | $L_2$     | ... | $L_k$     |
|--------|-----------|-----------|-----|-----------|
| Occ #1 | $eq_{11}$ | $eq_{12}$ | ... | $eq_{1k}$ |
| Occ #2 | $eq_{21}$ | $eq_{22}$ | ... | $eq_{2k}$ |
| ...    | ...       | ...       | ... | ...       |
| Occ #n | $eq_{n1}$ | $eq_{n2}$ | ... | $eq_{nk}$ |

Table 1: The translation equivalents matrix (EQ matrix)

$TU_j$  = the translation unit containing  $T\_word_j$ ;

$EQ_i$  = a vector, containing the TEs of  $T\_word$  in language  $L_i$ : ( $eq_{i1} eq_{i2} \dots eq_{in}$ )

More often than not the translation equivalents found for different occurrences of the target word are identical and thus identical words could appear in the  $EQ_i$  vector. If  $T\_word_j$  is not translated in the language  $L_i$ , then  $eq_{ij}$  is represented by the null string. Every non-null element  $eq_{ij}$  of the EQ matrix is subsequently replaced with the set of all ILI identifiers that correspond to the senses of the word  $eq_{ij}$  as described in the wordnet of the  $j$ -language. If this set is named  $IS_{ij}$ , we obtain the matrix  $EQ\_ILI$  which is the same as EQ matrix except that it has an ILI set for every cell (Table 2).

|        | $L_1$   | $L_2$   | ... | $L_k$   |
|--------|---|---|-----|---|
| Occ #1 | $IS_{11} = \{ILI_p   ILI_p$<br>identifies a synset<br>of $eq_{11} \}$ | $IS_{12} = \{ILI_p   ILI_p$<br>identifies a synset<br>of $eq_{12} \}$ | ... | $IS_{1k} = \{ILI_p   ILI_p$<br>identifies a synset<br>of $eq_{1k} \}$ |
| Occ #2 | $IS_{21} = \{ILI_p   ILI_p$<br>identifies a synset<br>of $eq_{21} \}$ | $IS_{22} = \{ILI_p   ILI_p$<br>identifies a synset<br>of $eq_{22} \}$ | ... | $IS_{2k} = \{ILI_p   ILI_p$<br>identifies a synset<br>of $eq_{2k} \}$ |
| ...    | ...   | ...   | ... | ...   |
| Occ #n | $IS_{n1} = \{ILI_p   ILI_p$<br>identifies a synset<br>of $eq_{n1} \}$ | $IS_{n2} = \{ILI_p   ILI_p$<br>identifies a synset<br>of $eq_{n2} \}$ | ... | $IS_{nk} = \{ILI_p   ILI_p$<br>identifies a synset<br>of $eq_{nk} \}$ |

Table 2. The matrix containing the senses for all translation equivalents (EQ\_ILI matrix)

If some cells in EQ contain empty strings, then the corresponding cells in  $EQ\_ILI$  will obviously contain empty sets. Similarly, we have for the  $T\_word$  the list  $T\_ILI = (ILI_{T1} ILI_{T2} \dots ILI_{Tq})$ .

The next step is to define our target data structure. Let us consider a new matrix, called VSA (Validation and Sense Assignment):

|        | $L_1$      | $L_2$      | ... | $L_k$      |
|--------|------------|------------|-----|------------|
| Occ #1 | $VSA_{11}$ | $VSA_{12}$ | ... | $VSA_{1k}$ |
| Occ #2 | $VSA_{21}$ | $VSA_{22}$ | ... | $VSA_{2k}$ |

|        |                   |                   |     |                   |
|--------|-------------------|-------------------|-----|-------------------|
| ...    | ...               | ...               | ... | ...               |
| Occ #n | VSA <sub>n1</sub> | VSA <sub>n2</sub> | ... | VSA <sub>nk</sub> |

**Table 3.** The VSA matrix

with  $VSA_{ij} = T\_ILI \cap IS_{ij}$ , if  $IS_{ij}$  is non-empty and  $\perp$  (undefined) otherwise.

The  $i^{\text{th}}$  column of the VSA matrix provides valuable corpus-based information for the evaluation of the interlingual linking of the the  $i$ -wordnet and T-wordnet.

Ideally, computing for each line  $j$  the set  $SA_j$  (sense assignment) as the intersection  $ILI_{j1} \cap ILI_{j2} \dots \cap ILI_{jk}$  one should get at a single ILI identifier:  $SA_j = (ILI_{T\alpha})$ , that is the  $j^{\text{th}}$  occurrence of the target word was used in all source languages with the same meaning, represented interlingually by  $ILI_{T\alpha}$ . If this happened for any T\_word, then the WSD problem (at least with the parallel corpora) would not exist. But this does not happen, and there are various reasons for it: the wordnets are partial and (even the PWN) are not perfect, the human translators are not perfect, there are lexical gaps between different languages, automatic extraction of translation equivalents is far from being perfect, etc.

Yet, for cross-lingual validation of interlinked wordnets the analysis of VSAs may offer wordnet developers extremely useful hints on senses and/or synsets missing in their wordnets, wrong ILI mappings of synsets, wrong human translation in the parallel corpus and mistakes in word alignment. Once the wordnets have been validated and corrected accordingly, the WSD (in parallel corpora) should be very simple. There are two ways of exploiting VSAs for validation:

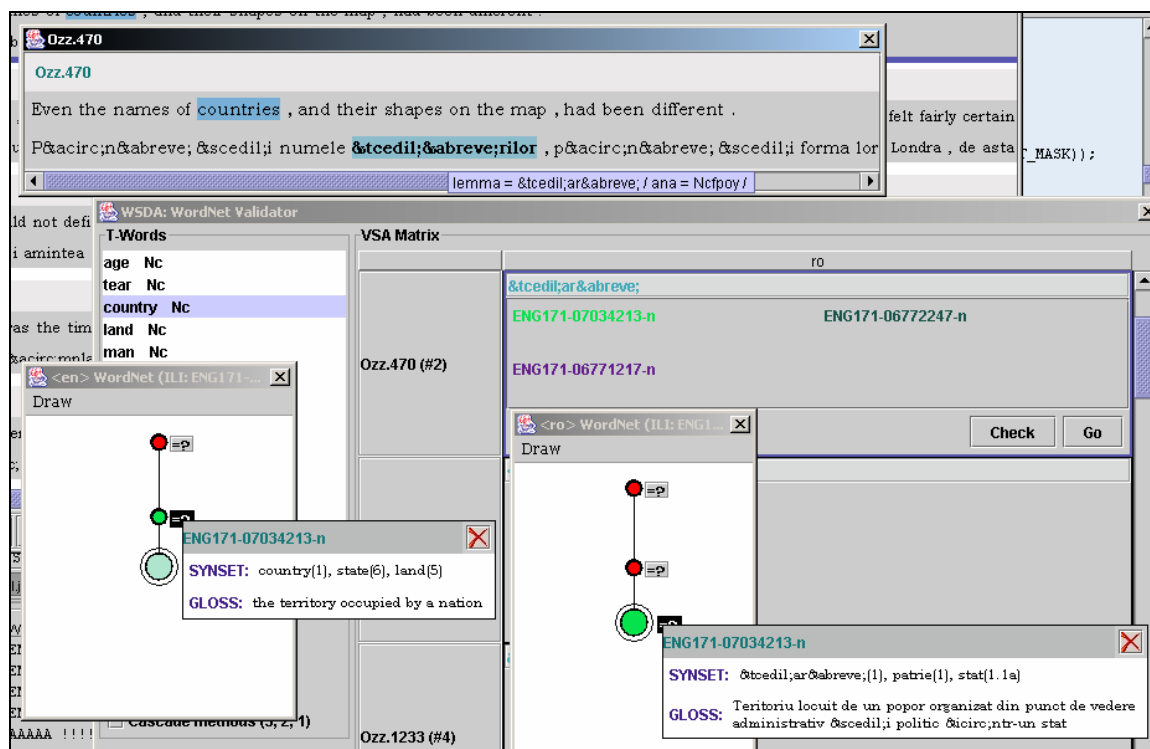
**Vertical validation (VV):** the development team of  $i$ -wordnet (native speakers of the language  $L_i$  with very good command of the target language) will validate their own  $i$ -wordnet with respect to the T-wordnet, that is from all VSA matrixes (one for each target word) they would pay attention only to the  $i^{\text{th}}$  column (the  $VSA(L_i)$  vector).

**Horizontal validation (HV):** for each VSA all SAs will be computed. Empty SAs could be an indication of ILI mapping errors still surviving in one or more wordnets (or could be explained by lexical gaps, wrong translations etc) and as such, the suspicious wordnet(s) might be re-validated in a focused way. The case of an SA containing more than a single ILI identifier could be explained by the possibility of having in all  $i$ -languages words with similar ambiguity.

Our system called WSDtool implements the methodology described above and offers an easy-to-use interface for the task of semantic validation. It incorporates the translation equivalents extraction system (TREQ&TREQ-AL, described in [Tufiş et al., 2003] as well as a graphic visualization of the two wordnets used in the validation process. We exemplify a horizontal WSDtool validation session by considering the En-Ro language pairs. The intersection between ILI sets of  $w_{en}$  and  $w_{ro}$  is presented in a table for every occurrence of  $w_{en}$  in the parallel corpus. The cell at line  $i$  (labeled with the translation unit identifier of the sentence containing the  $i^{\text{th}}$  occurrence of  $w_{en}$ ) and column labeled with the target language name (ro) contains the intersection of ILI sets of literals  $w_{en}$  and  $w_{ro}^i$  where  $w_{ro}^i$  represents the Romanian translation for the  $i$ -th occurrence of  $w_{en}$ . The cell's content ranges over the next three cases:

1. the cell contains an ILI set; this means that each of the literals  $w_{en}$  and  $w_{ro}^i$  are found in synsets which are mapped onto the same ILIs. The user is required to choose the ILI which points to the correct sense in both languages (see below). If such an ILI cannot be found, the user is offered another choice: to indicate the missing sense in the Romanian wordnet for the  $w_{ro}^i$  literal. Finally, if all the senses of

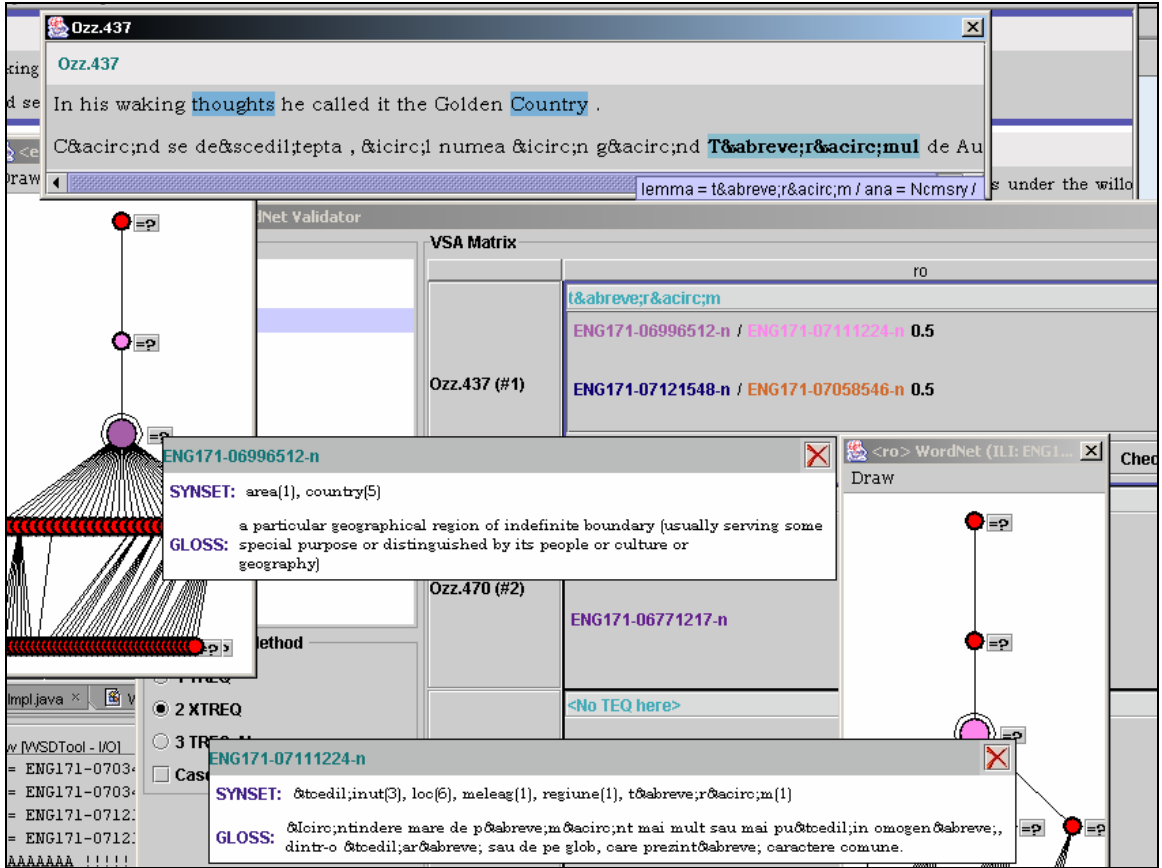
$w_{ro}^i$  are implemented, the user is asked to remap one of  $w_{ro}^i$  synsets to satisfy the translation equivalence pair;



**Figure 12:** The translation unit Ozz.470 contains the second occurrence of  $w_{en}$  'country'

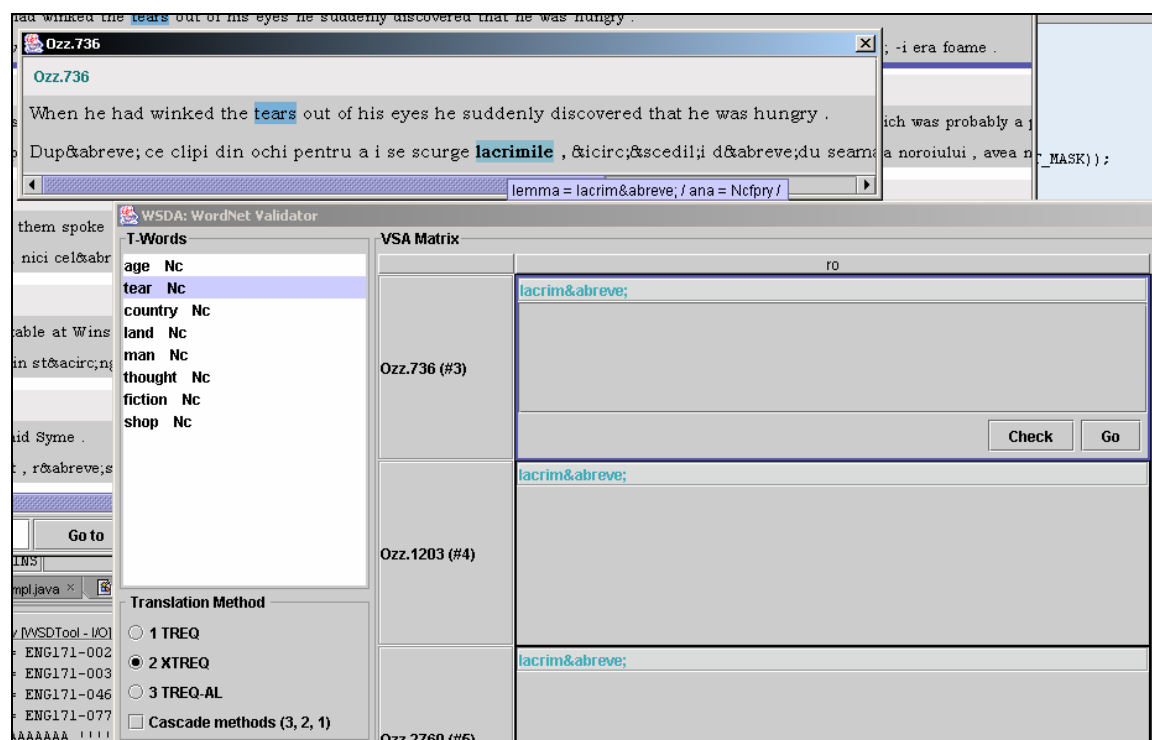
This occurrence is translated in Romanian by  $w_{ro}^2$  'țară' and we can see that the selected table cell contains the ILI set of the intersection. In this case, ILI171-07034213-n is the identifier for the correct sense in both Romanian and English

- the cell contains pairs of ILIs; each pair ends with a real number denoting a similarity measure between the members of the pair; the similarity measure was calculated as  $\delta_N = \frac{1}{1+N}$  where  $N$  is the number of links between the pair members in the PWN hierarchy (it is easily seen that when  $N = 0$ ,  $\delta_0 = 1$  which means that the two ILIs are identical; for  $N = 1$ ,  $\delta_1 = 0.5$  which shows an HH relationship or a coordination between pair members); all pairs in the interval  $[\delta_2, \delta_0]$  were retained. The user is now required to choose the pair which reflects the best HH relation between pair members ('the best' means that the pair member corresponding to  $w_{en}$  should reflect the sense used – see figure 3). If such a pair does not exist, the preceding actions (from 1.) are to be followed;

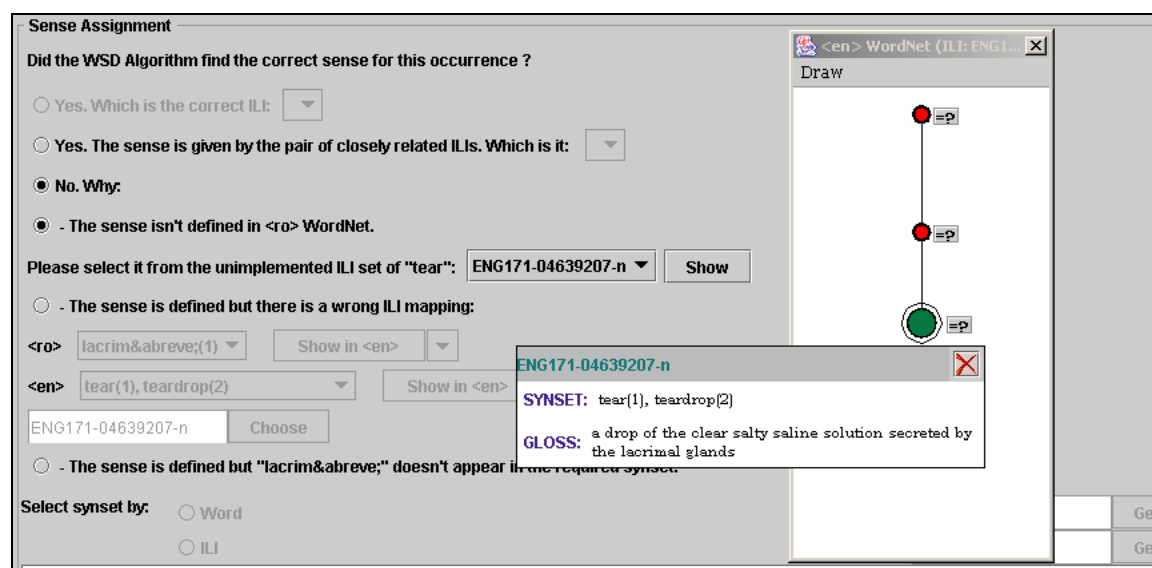


**Figure 13:** The selected cell ( Ozz.437(#1), ro ) reflects the ILI intersection between ‘country’ and ‘tărâm’. As none of the corresponding ILIs are the same, the cell presents two pairs of ILIs between which  $\delta_N$  is maximal (0.5, with  $N = 1$ ). In this case the first pair is correct.

3. The cell is empty; this is a potential alignment error in the Romanian wordnet or an incomplete Romanian synset (see figure 4). If  $(w_{en}, w_{ro}^i)$  is a correct translation pair, then one of the following must hold: the relevant  $w_{ro}^i$  synset is wrongly mapped, the sense of the  $i^{\text{th}}$  occurrence of  $w_{en}$  is not yet implemented for the corresponding translation equivalent literal  $w_{ro}^i$  (see figure 5) or the literal  $w_{ro}^i$  does not belong to the relevant Romanian synset. If the latter case holds, the user is asked to add the literal (with the appropriate sense number) to the correct synset (this way, synset expanding can be achieved in a focused way: context study).



**Figure 14:** The cell at ( Ozz.736(#3), ro ) is empty. The third occurrence of ‘tear’ was translated by ‘lacrimă’ (SGML entities notation: ‘lacrim&abreve;’) and this is a correct translation pair.



**Figure 15:** The reason for the void intersection above is that ‘tear’ was used in a sense that was not implemented in the Romanian wordnet. The figure shows a portion of the check window where the user specifies that this sense of ‘tear’ is not implemented in the current version of the Romanian wordnet.

## Qualitative Evaluation Results

The WSDtool helped project partners to detect wrong alignments between their wordnet and the PWN2.0 and also to spot incomplete synsets (that is, synsets that lack the translation equivalent of the target word found by the WSDtool in a translation unit). Since the rationale for the WSDtool methodology has been described at length elsewhere (i.e. see the recently published article “D. Tufis, R. Ion, N. Ide: *Fine-Grained Word Sense Disambiguation Based*

on *Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets*” at the 20th International Conference on Computational Linguistics, COLING2004 held this year in Geneva) we will present only the results of the validation tasks carried out by the teams involved (see the table below):

|                    | <i>GR</i>             | <i>BG</i>             | <i>SR</i>              | <i>RO</i>              |
|--------------------|-----------------------|-----------------------|------------------------|------------------------|
| <b>TOTAL Occs.</b> | 1156                  | 1159                  | 1232                   | 1291                   |
| <b>WSDOK1</b>      | 610 ( <b>52.76%</b> ) | 737 ( <b>63.58%</b> ) | 1000 ( <b>81.16%</b> ) | 1056 ( <b>81.79%</b> ) |
| <b>WSDOK2</b>      | 47 ( <b>4.06%</b> )   | 73 ( <b>6.29%</b> )   | 90 ( <b>7.30%</b> )    | 99 ( <b>7.66%</b> )    |
| <b>SNDEF</b>       | 0                     | 0                     | 22 ( <b>1.78%</b> )    | 0                      |
| <b>SDEFADD</b>     | 167 ( <b>14.44%</b> ) | 127 ( <b>10.95%</b> ) | 59 ( <b>4.78%</b> )    | 2 ( <b>0.15%</b> )     |
| <b>SDEFMAP</b>     | 0                     | 0                     | 0                      | 2 ( <b>0.15%</b> )     |
| <b>BLURRED</b>     | 31 ( <b>2.68%</b> )   | 35 ( <b>3.01%</b> )   | 0                      | 56 ( <b>4.33%</b> )    |
| <b>MACHINE</b>     | 301 ( <b>26.03%</b> ) | 187 ( <b>16.13%</b> ) | 61 ( <b>4.95%</b> )    | 76 ( <b>5.88%</b> )    |
| <b>HUMAN</b>       | 0                     | 0                     | 0                      | 0                      |

**Table 4:** Validation results

In this table, for each language, we considered only the English occurrences of the target words that were translated in the respective language. We showed before that the 209 target words have in the English original 1621 occurrences. However, as one may notice from the first row of the Table 4, an average of 400 occurrences per language were not translated.

The definitions of the entries in the first column of the table are as follows:

- **WSDOK1:** WSDtool found a translation equivalent that has at least one ILI in common with the current target word. This is a good point for the source wordnet;
- **WSDOK2:** WSDtool found a translation equivalent that is semantically closely related with the current target word. This is also a good point for the source wordnet;
- **SNDEF:** the sense that the current occurrence of the target word was used in is not yet implemented in the source wordnet. It cannot be considered a bad point for the source wordnet because we wanted to evaluate only the existing sense inventory;
- **SDEFADD:** the sense of the current occurrence of the target word is defined in the source wordnet but the translation equivalent does not belong to that synset. This means that the synset is incomplete and we considered this case a bad point for the source wordnet;
- **SDEFMAP:** the sense of the current occurrence of the target word is defined in the source wordnet but the relevant synset (that contains the translation equivalent) is wrongly mapped on ILI. That is, it has another correspondence in Princeton wordnet. Also a bad point of the source wordnet;
- **BLURRED:** the translation equivalent is not wrong but the translation itself is rather loose and does not justify adding the translation equivalent to the relevant synset. Not a bad point for source wordnet!;
- **MACHINE:** the translation equivalent was wrongly chosen by the word alignment engine of the WSDtool. Of course, this cannot be a bad point for the source wordnet;
- **HUMAN:** the translation equivalent, although correctly chosen by the system, is wrong due to defective translation. The bad point remark is the same as above.

If we discard the last three rows of the table (they are not really relevant for the evaluation of the wordnets accuracy interlinking), and consider that the first two rows (WSDOK1 and WSDOK2) provide evidence for correct synsets content and alignment, then the figures shown in Table 5 represent the percentages of the number of occurrences of the target words that were successfully disambiguated based on information provided by each individual wordnet.

|              | <i>GR</i>             | <i>BG</i>             | <i>SR</i>              | <i>RO</i>              |
|--------------|-----------------------|-----------------------|------------------------|------------------------|
| TOTAL Occs.  | 824                   | 937                   | 1171                   | 1159                   |
| WSD Accuracy | 657 ( <b>79.73%</b> ) | 810 ( <b>86.44%</b> ) | 1090 ( <b>93.08%</b> ) | 1155 ( <b>99.65%</b> ) |

**Table 5:** Statistics on semantic interlingual validation

The WSDtool precisely pinpointed the errors (most of them caused by incomplete synsets) and each partner corrected them accordingly.

Semantic validation of wordnets alignment is a secondary functionality of the WSDtool which was primarily designed as a word sense disambiguation program. For the WSD task, the program incorporates also a word sense agglomerative clustering module and several heuristics to cope with the occurrences not translated in one or more languages - see for details (Tufiş, Ion, Ide, 2004). One of the greatest advantages of applying such methods to parallel data is that it may be used to automatically sense-tag corpora in not only one language, but rather several at once and with the same sense inventory. If we note that there is a considerably large number of literals with a single sense in PWN (119528 out of 145627 which means approximately 82%), we see that the WSD method as implemented in the WSDtool can almost have a full coverage if we extend it by saying that every translation pair for which there is a single sense in its English part (as extracted from PWN) receives that sense.

Recently we conducted experiments (with the final deliveries of the BalkaNet wordnets) for proper WSD task on the 1984 parallel corpus. The 1621 occurrences of the 209 English target words were hand disambiguated by three independent experts and the disagreements were negotiated at a later moment. This way, resulted a Gold Standard annotation which was used to assess the performance of WSDtool. The evaluation considered three sense inventories available in the BalkaNet wordnets: the synsets IDs (the Princeton Wordnet 2.0 senses), the IRST DOMAIN labels, and the SUMO/MILO categories. As one would expect, as with any classification task, the WSD accuracy is better when the number of semantic classes is smaller.

The results are presented in the following table (1621 occurrences of all target words from the list):

| Sense Inventory       | ILI records   | SUMO categories | Domains       |
|-----------------------|---------------|-----------------|---------------|
| Sense Inventory Size  | <b>115424</b> | <b>2066</b>     | <b>163</b>    |
| WSDtool Precision (P) | <b>79.48%</b> | <b>87.26%</b>   | <b>92.78%</b> |
| WSDtool Recall (R)    | <b>78.16%</b> | <b>85.81%</b>   | <b>91.23%</b> |
| F-Measure (FM)        | <b>78.81%</b> | <b>86.52%</b>   | <b>91.99%</b> |

**Table 6:** Word Sense Disambiguation with three sense inventories

Out of 1621 total occurrences 27 could not receive a sense tag (in any of the sense inventories) mainly because the target literal was wrongly aligned by the translation equivalents extractor module of the WSDtool. In this case we used a simple heuristics assigning the most

frequent sense label of the literal in question (the most frequent sense number 1, the domain of the most frequent sense number and the composed SUMO category of the most frequent sense number). The results changed as shown in Table 7:

| Sense Inventory                  | ILI records | SUMO categories | Domains |
|----------------------------------|-------------|-----------------|---------|
| Sense Inventory Size             | 115424      | 2066            | 163     |
| WSDtool Accuracy<br>(P = R = FM) | 78.74%      | 87.16%          | 92.78%  |

**Table 7:** Word Sense Disambiguation with three sense inventories (using a simple heuristics)

The results are not surprisingly better than most results reported in the WSD literature which in the vast majority try to solve the problem in monolingual texts. Parallel corpora supported by aligned wordnets, as in our experiment, are extremely valuable resources not largely available to the research community.

## ***Experimenting with Valence Frames***

### **Czech - Adding Verb Valency Frames**

Verbs are usually described by means of their **valence frames**. They can contain both the **syntactic information** about the verb construction itself, i.e. what **surface cases** (in Czech and other highly inflected languages) are associated with a particular verb, and the **deep cases** or **semantic roles** that are required by the meaning of the verb.

Our main was to come up with a consistent system of semantic role tags that would form a **base for lexico-semantic constraints** integrated into various NLP modules such as a natural language **parser**. When building Czech verb synsets we have paid a systematic attention to the surface verb valences. This follows from inflectional nature of Czech which displays a rich declension structure – each Czech noun (as well as adjective, pronoun, numeral) can appear in one of seven cases: Nominative - 1, Genitive - 2, Dative - 3, Accusative - 4, Vocative - 5, Locative – 6 and Instrumental - 7. This is indicated in valence frames, i.e. each Czech verb synset contains also its respective valence frame displaying the information about the corresponding morphological cases that are obligatorily (or optionally) associated with it.

The first step is to have the information about surface valences – for Czech we have a list of 15 000 verbs (Pala, Ševeček, 1998), however the links between valences and senses have been systematically prepared for some 5000 items so far, particularly for those being included to Czech wordnet (the estimated number of verbs in Czech is about 36 000 items). The surface valences display the following form:

```
balit:1 (pack:11)
1 kdo1 = co4 do čeho2 ← valence frame together with the respective
sense number,
balit:2 (flirt:3)
kdo V koho ← valence frame together with the respective
sense number,
balit:3 (pack:12)
kdo1 V co4 ← valence frame together with the respective
sense number,
```

### **A Complete Notation**

While EuroWordNet notation for Internal Language Relations including semantic roles (such as ROLE\_AGENT – ROLE\_AGENT\_INVOLVED) is based on binary relations we have de-



cided to opt for the more **complex** and **empirically adequate** notation which comprises both **surface** (morphological) cases required by Czech, and the respective **semantic roles**, e.g.:

(vf1) {*jíst, eat*} kdo1\*AG(person:1|animal:1)=co4\*SUBSTANCE(food:1)

(vf2) {*pít, drink*} kdo1\*AG(person:1|animal:1)=co4\*SUBS(beverage:1)

(vf3) {*obléct si, put on*} kdo1\*AG(person:1|animal:1)=co4\*ART(garment:1) na  
co4\*BODY(body part:1),

(vf4){*vyprávět:1|tell:3*}kdo1\*AG(person:1)=co4\*INFO(message:2),komu3\*ADR(recipient:1)

The morphological cases are indicated as said above. The semantic roles are denoted by the general labels taken mainly from the **EWN TOP Ontology**, together with the **subcategorizing literals** from the set of Base Concepts, and include the numbers of the respective senses. In our opinion, the notation used in (vf21)-(vf4) presents the information about the syntactic and semantic properties of a given verb in a natural way and it describes the real lexical data more adequately.

The comparison of the information contained in the largest Czech dictionaries with lexico-semantic constraints obtained (semi)-automatically shows that the (semi)-automatic technique of semantic role tagging can significantly speed-up the process of building verb valence dictionaries designed as lexicons appropriate for NLP applications. For this purpose the newly built interface linking Czech wordnet with the Czech morphological analyzer AJKA has been implemented (see below).

### 1000 Czech Verbs

As a case study we present the results of our investigation of the 1000 frequent Czech verbs taken from Czech wordnet. The valence frames we are working with come from the list of approx. 1000 Czech and English verbs or, more precisely, from the list of Czech and English verb synsets belonging to the Czech and English WordNet. Our frames differ from others, e.g. the ones used in Vallex [5] in the following aspects:

- inventory of the main semantic roles is based on the EuroWordNet Top Ontology and the set of Base Concepts,
- main roles are **further subcategorized** by means of the particular literals taken from PWN 2.0, and this sub-categorization can be regarded as complementary to the one used in Vallex,
- close relation to the wordnet with its **large hierarchical structure** allows us to get closer to real lexical data.

Take e.g. the verbs *vstoupit* | *to enter* in the following sentences:

(v1) *Ten člověk vstoupil do strany v r. 1968.*

(v2) *Ten člověk vstoupil do budovy před 10 min.*

(v1e) *This person entered the (Communist) party in 1968.*

(v2e) *This person entered the building 10 min. ago.*

If we use an existing inventory of the roles then the constituents *strana* | *party* and *budova* | *building* would be most likely labelled as PAT(iens) but our knowledge of Czech and English tells us that this label does not capture the respective difference in meaning. We obviously are dealing with the two different senses of the verb *vstoupit* | *enter* or, more precisely, with *vstoupit:4|enter:3* and *vstoupit:3|enter:1* if we use the standard WordNet notation (PWN 2.0). Thus *vstoupit:4|enter:3* means that people typically enter political organizations and *vstoupit:3|enter:1* denotes that people enter places like buildings. If we want to express this fact by means of the semantic role tags we need **more specific sub-categorization features** or labels that would express the meaning differences indicated above. A similar observation

can be made about many other verbs, see e.g. the roles associated with verbs like *eat*, *drink*, *wear* or *drive* which in turn require sub-categorization features like FOOD, BEVERAGE, GARMENT and VEHICLE.

The solution we are offering uses **two level semantic role labelling** in our valence frames:

- on the first level – **general labels** like AG, PAT, OBJ, INSTR, LOC, ADDR, ...
- on the second – **sub-categorization level** we take advantage of the rich wordnet hierarchical structure and use **selected literals** occurring in the particular synsets as labels – through them we can access the individual lexical units when we process sentences (v1) or (v1e) on the morphological and syntactic level. It should be stressed that wordnet hierarchical structures capture approx. 100 000 synsets (in Princeton WordNet). No other resource offers such extensive coverage.

The mentioned list of 1000 verbs **was sorted** according to their **deep valence frames** with the assumption to obtain some semantically interesting verb classes. If we have a look at the obtained list we can say that our assumption has been justified with some reservations, namely: the list of 1000 verbs is still not large enough yet, there is a quite large number of the small classes (groups) typically containing 2 items. The results we have arrived at are shown in the following table:

| Verb frame  | Frequency | Sense characterization                    |
|---|-----------|---|
| AG(person:1)=ANY(anything:1)                      | 33        | various verbs                             |
| AG (person:1) = ACT (act:2)                       | 23        | solving tasks, performing activities      |
| AG (person:1) = OBJ (object:1)                    | 21        | manipulating with objects                 |
| AG (person:1) = PAT (person:1)                    | 21        | relations between persons                 |
| AG(person:1)=SOC(person:1)                        | 16        | social interaction                        |
| AG (person:1) = X                                 | 15        | non-personal verbs, without complement    |
| AG (person:1) = \$ (ze)                           | 15        | communication activities, verba dicendi   |
| AG (person:1) = SUBS (food:1)                     | 9         | verbs of eating                           |
| AG (person:1) = LOC (location:1)                  | 8         | motion verbs                              |
| AG (person:1) = ACT (job:1)                       | 7         | Working                                   |
| AG (person:1) = OBJ (object:1) = LOC (position:1) | 7         | motion with objects, positioning in space |
| AG (person:1) = OBJ (object:1) = OBJ (object:1)   | 7         | combining objects                         |
| AG (person:1) = ABS (abstraction:1)               | 6         | keeping rules                             |
| AG (person:1) = ART (garment:1)                   | 6         | verbs of dressing                         |
| AG (person:1) = EVEN (result:3)                   | 6         | making conclusions                        |
| AG (person:1) = ACT (role:1)                      | 5         | being in a position (or losing it)        |

**Table 8:** Valence frames results

Looking at the table the following conclusions can be drawn:

- discrimination power of the frames is reasonable and it is closely related to the selection of the sub-categorization features, i. e. if the sub-categorization features are chosen appropriately a usable semantic classification of the verbs can be developed,
- the obtained classes are not arbitrary and can be confirmed by the corpus data by means of Word Sketches techniques,
- we get independent feedback that frames associated with the respective verbs manually can be confirmed also semi-automatically,
- the classes in some way correspond to Levin's verb classification [Levin, 4].

### A Comparison with Bulgarian and Romanian

The following hypothesis can be formulated: the deep valence frames for approx. 1000 Czech verbs have been taken from Czech wordnet. Those verbs are linked to their English equivalents by means of ILI, which means that the frames prepared for Czech verbs can be applied to their English equivalents as well. It certainly would be premature to claim that the semantic roles associated with Czech verbs strictly apply to their English counterparts, such statement

might be considered too universalistic, but in any case if they are translation equivalents with the same meaning there **has to be** a reasonable agreement.

In BalkaNet a comparison has been made to test whether the indicated agreement would apply also to other languages, particularly to Bulgarian and Romanian. The results of the comparison appear very promising and they can be characterized as obviously confirming the hypothesis mentioned above. The deep valence frames prepared for approx. 1000 Czech verbs have been tentatively associated via ILI with the corresponding English, Bulgarian and Romanian verbs.

### **Romanian - Adding Verb Valence Frames**

In order to provide valence frames for some verbs in our wordnet, we took the following steps:

- For a set of Romanian verbs occurring in the “1984” corpus with a frequency of around 100 occurrences, their concordances were extracted.
- For each concordance the verb was semantically disambiguated: the corresponding Romanian synset containing it was identified using VisDic for visualizing the RoWN, alongside with PWN 2.0 and CzWN.
- The valence corresponding to the equivalent Czech synset is identified (if existent) and is checked against the Romanian data, and modified accordingly if necessary. When a Czech valence is not identified, a valence suggestion for the Romanian verb is provided, following the indication in the file provided by the Czech team.

### **Remarks**

1) The disambiguation process raised some difficulties due to the following facts:

- Sometimes the concordance chosen (a sentence) proved not enough for choosing the right meaning.

Ex.: Se gândi. / SE thought-he

- Some occurrences are difficult to assign a sense. On the one hand, this is due to the fact that senses are too refined in the wordnet, and, on the other, to auto hyponymy.

Ex.: Winston se gândi, apoi spuse:.../Winston SE thought, then said-he:...

The verb in this example can be disambiguated as either belonging to the synset {chibzui:1.2, cugeta:1.2, [se] gândi:1.2.x}(EN: {think:3, cogitate:2, cerebrare:1}, Gloss: use or exercise the mind or one's power of reason in order to make inferences, decisions, or arrive at a solution or judgments) or to the synset {[se] gândi:1.2} (EN: {think:8}, Gloss: decide by pondering, reasoning, or reflecting). However, the two senses are in hyponymy relation, the former being the hypernym of the latter.

2) The valence frames we present were identified for one verb in a synset. So far we have not checked against a corpus if they are valid for the other verbs in the same synset. However, taking a rough look at them, we could say that the frame suggested for one of the verbs stands correct for the others, as well. Still, this needs testing on a corpus.

3) Assigning the valence for each synset raised the following issues:

- Insufficient data. The concordances extracted are not enough for giving a final form to the frame. For instance, for the verb a începe:1 the frame cineva1\*AG(person:1) = \$(să) be augmented with an alternative: \$(să)|(ceva4)\*ACT(act:2). So, a larger corpus would be necessary for giving a definite form of the valence frames.

- Interlingual comparison of the frames. For all the senses identified for the Romanian verbs, only 13 have Czech equivalents for which frames are provided in the CzWN. When comparing them with the ones suggested for the verbs in RoWN, one notices that most of the times (in 9 cases out of 13) the frames are identical. In two situations the identity is prevented by the incompleteness of RoWN frames; the frames suggested for Czech are valid for Romanian, too, but they were not encountered in the concordances analyzed. In one situation the difference is triggered by the different syntactic behavior of the equivalent verbs in the two languages. One example is provided by the Romanian synset {începe:1, porni:7.3}.

In another situation the lack of identity is due to the fact that, while for Czech no adjunct is included in the frame, for Romanian there are some adjuncts in the first two frames proposed: see {chibzui:1.2, cugeta:1.2, [se] gândi:1.2.x}.

- Possibility for clustering. Let us consider the next two synsets: { [se] afla:3.1, [se] găsi:9.1, fi:3.1 } and { [se] afla:3.1.x, fi:3.1.x, [se] găsi:9.1.x } Both have the same frame: (cineva1\*AG(person:1)|ceva1\*OBJ(object1)) = unde\*LOC(location:1). Moreover, they are in hyperonymy relation: the latter is the hyperonym of the former. In such cases, when two synsets have the same valence, the semantic difference between them is rather difficult to perceive, and they are in hyponymy/hyperonymy relation, then we consider that they are worth being clustered.

Romanian is a pro-drop language. So far there is no means for treating the sentences in which the subject is not lexicalized, but is expressed due to the rich verb inflection. To mark the “subject” position in the valence as X is not a solution, as there are also verbs in Romanian that cannot have a subject (e.g. ploua “to rain”) and thus we would not have a clear image of which verbs are impersonal and which simply doesn’t have the subject lexicalized. Another possibility of notation for the pro subject would be to mark it as optional. Place, manner, time, etc. express the circumstances in which an activity, etc. takes place. Thus, each verb expressing an activity, etc. can have such adjuncts. To specify the possible adjuncts as optional in the valence frame for (each verb in) each synset is too time and energy consuming. That is why we consider that a different way of dealing with such situations should be found.

However, if MAN (for instance) is a complement (N.B. not an adjunct), it must appear in the frame. This is the case of the verb a se comporta “to behave” in Romanian, which cannot occur without its manner complement. Moreover, if we choose to mark the optionality of adjuncts, we cannot mark the unlexicalized subject in the same way, since the syntactic phenomena in each case is different.

### **Bulgarian - Adding Verb Valence Frames**

The states of affairs (either concrete or abstract) to which simple sentences refer can be considered (to some extent) as constant in time and language-independent. This presupposes that a certain predicate is associated with a given number of arguments by means of a given type of semantic relations. This statement has two important consequences (Koeva, 2004).

As the argument structures correspond to situations, their description on the semantic level has to be independent from the natural languages. Thus, the argument structure (number of arguments and specific semantic relations) is constant without a direct relation to the natural language in which the arguments receive lexical and syntactic interpretation - the cross-language differences appear only on lexical and syntactical level. The second consequence is that if the predicate is realized with more than one lexical item the argument structure of all items has to be equal. Thus the synonymous verbs always take equal number of arguments with equal semantic relations although there could be differences in the projection of the syntactic phrases, selective restrictions and explicitness of phrases.

In order to provide valency frames for some Bulgarian verbs, we took the following steps:

- Select all Bulgarian verb synsets that correspond to Check synsets assigned with valency frames.
- Find corpora examples for Bulgarian verbs and distinguish different syntactic environment according to different meanings defined in Bulgarian WordNet.
- The valence corresponding to the equivalent Czech synset is identified (if existent) and is checked against the Bulgarian data, and modified accordingly if necessary.

We started with the determination of a uniform theoretical model for the formal representation of the Bulgarian syntactic structures which underlies the architecture of the corresponding software tool developed for data processing (Koeva, 2004). The theoretical model is a linguistic hypothesis itself, although it relies both on the deep studying of the world theories and the Bulgarian linguistic tradition. Several famous theoretical models dedicated to verb semantics, predicate – argument structure and verbal alternations are very influential.

We supposed that during the process of practical work we could encounter new language phenomena that could not be described with the accepted theoretical model. That is why we have provided options for adding new parameters for language description, as well as for modification of the already chosen parameters. This assumption practically made the corresponding web based system for adding, editing and validating data to a great extent language independent as well as theory independent. As a result the system can be easily remobilized in order to be used for languages with different grammars or for one and the same language for different purposes.

It is recommended that the number of arguments be evident from the meaning definition. The definitions, as well as the distinction of the meanings, may be changed if they are not precise enough in the monolingual explanatory dictionaries. The author of the syntactic entry can check the corpora distributions of the target word forms, if they are provided. The examples taken from the corpus as well as the language competence of the authors could lead to the union of two or more meanings, the division of one meaning into several, the addition of the new meaning or the elimination of an old one – we marked the differences but we did not change the WordNet structure.

The development of the Syntactical Frames of the Bulgarian Language is carried out with the help of a web– based system called SynText (Syntactic lexicon Tool), developed at DCMB. The tool allows the developers to work independently from each other and using different operational systems (i.e. Windows or Linux), while using one and the same data base. The architecture of the SynText system is organized in accordance with the dictionary entry structure described above. The SynText application has the following major characteristics:

- **Language independent** – data from different languages could be added (the language dependent parameters and the pertaining values can be easily changed – reformulated, added or deleted, if necessary);
- **Theory independent** – (the theory dependent parameters and the pertaining values can be again easily changed, if necessary, thus the application can be considered to a great extent (not fully of course) theory independent and the verification of the theoretical hypothesis could be obtained);
- **Dynamic** – the system allows fast and easy administration of the linguistic markers (parameters and their values) from the authorized person, it is fully configurable and customizable by the administrator;
- **Web–based** – different operational systems can be used with minimum requirements for the client machine;
- **Functional** – many authors could work simultaneously on one and the same data base, as the system supports for the users user roles, provides authentication and special guest access;

- **Uniform** – the input data are unified within the frame of the current theoretical model;
- **Informational** – the system allows different check ups: to recall all words from one and the same grammatical type, to recall all words that satisfy particular criterion, to recall all words that have equal grammatical features, etc.
- **Open** – the system is based on open source technologies, with open architecture and written in pure Java – so it can be deployed on different platforms.

## **Bulgarian - Verb Net**

The creation of verb valence frames for a given natural language is extremely important for the syntactic and semantic analysis of texts written in this language. The verb frames could be used in the both phases of analysis in natural language processing (NLP). The frame representation allows building of a real intelligent software system that “understands” the meaning of texts in the given language even in cases when those texts include unknown linguistic constructions [Totkov’2003].

### **Bulgarian Verb Net: Methods and Tools**

#### **Prehistory**

There are a lot of examples for the representations of the verb syntactic and semantic combinatorial possibilities. A common feature of those representations is the grouping of verbs into classes: *VerbNet* [Kipper et al.’2000] and *LCS* [Dorr, Olsen’1996] use modifications of Levin’s classes, *FrameNet* uses its own classes [Backer et al.’2002], and *PropBank* – the *VerbNet* classes [Kingsburry et al’2002].

*FrameNet*<sup>4</sup> is based on the so-called frame semantics [Fillmore, Baker’2001’]. Each frame models semantic and syntactic valence (by the frame elements, the grammatical function and phrase type). Frame elements represent different situational roles. Classification of verbs and corresponding frames in *FrameNet* is done completely on semantic principle, in distinction with the other systems where (following the classification of Levin) the participation of a verb in diathesis alternations is the criterion for grouping of verbs into syntactic-semantic classes.

In *PropBank* a list of possible arguments and their labels are given for each annotated verb. About 1400 thematic roles and special labels for specification of adjuncts (as TMP – time, MNR – manner, DIR – direction, PRP – purpose, etc.) are used in the project.

*LCS* is a verb dictionary (11,000 verbs) created in the Laboratory of computational linguistics and information processing at the University of Maryland, USA. It represents the semantics of the lexical units by combination of syntactic structure and semantic contents.

An attempt for representation of the model of verb subcategorization (MVC) for 400 Bulgarian verbs is the work [Popova’1987]. A computer realization of the representation is presented in [Totkov’90, Totkov, Tanev’1999]. The suggested notation of MVC consists of several syntactic-semantic slots (arguments) that the verb creates in the sentence context. Several MVCs can correspond to one and the same Bulgarian verb.

#### **Approach Applied**

The research carried out has the **main goal** to find out an appropriate frame structure that allows incorporation of the primary syntactic and semantic information for the Bulgarian verbs in the *BWN*. In this way, the resulting *BWN* will be an invaluable language resource that could be used by linguists and non-specialists as well as a source for building various kinds of NLP systems. The chosen notation of the verb valence frames allows for each verb synset to be specified several frames, corresponding to different verb valences. Each frame consists of: a list of synset literals; list of arguments, as well as its status (obligatory or not); example sen-

---

<sup>4</sup> FrameNet is a lexicographic research project consisting of two parts – FrameNet 1 and FrameNet 2, leading by Ch. Fillmore and B. Atkins.

tence(s). For each argument the specification includes: corresponding indefinite pronoun (submitting appropriate morpho-syntactic information of the surface verb valency), semantic role (an identifier of the deep verb valencies denoted by the general labels taken from the EWN Top Ontology, e.g. *AG* – agent, *ACT* – act, *OBJ* – object, etc.) together with subcategorisation literals accompanied by numbers of respective senses (e. g. *person:1*, *plant:1*, *artifact:1*, etc.).

The enrichment of *Balkan wordnet* [Pala, Smrz;2004] and the construction of verb valency frames is initiated by the Czech *BalkaNet* team for the *Czech wordnet (CzWN)* and is later prolonged for the Bulgarian one. Since both languages (Czech and Bulgarian) are Slavonic a relatively great part of the verbs should realize their valency in one and the same way.

The following examples prove the last assumption:

produce, make, create – create or manufacture a man-made product

BG: {произвеждам} някой\*AG(person:1) | не-  
що\*ACT(plant:1)=нещо\*OBJ(artifact:1)

CZ: {vyrabet, vyrobit} kdo\*AG(person:1) | co\*ACT(plant:1)=  
co\*OBJ(artifact:1)

uproot, eradicate, extirpate, exterminate – destroy completely, as if down to the roots; "the vestiges of political democracy were soon uprooted"

BG: {изкоренявам, премахвам} някой\*AG(person:1) |  
нещо\*AG(institution:2)= нещо\*ATTR(evil:3) | \*EVEN(terrorism:1)

CZ: {vykorenit, vyhladit, zlikvidovat}  
kdo\*AG(person:1) | co\*AG(institution:2)=  
co\*ATTR(evil:3) | \*EVEN(terrorism:1)

carry, pack, take – have with oneself; have on one's person

BG: {нося, взимам} някой\*AG(person:1)= не що\*OBJ(object:1)

CZ: {vzit si s sebou, brat si s sebou, mit s sebou, mit u sebe}  
kdo\*AG(person:1)= co\*OBJ(object:1)

The above consideration is the motivation for the chosen approach in the construction of the valency frames of the Bulgarian verbs. It is performed in two stages:

**Stage 1.** Construction of the frames for those Bulgarian verb synsets that have corresponding (via ILI number) verb synsets in the CzWN and in addition these CzWN synsets are provided with already developed frames.

**Stage 2.** Creation of frames for verb synsets without analogues in the CzWN.

## Software Tools

*Two software tools* are developed for the construction of the Bulgarian verb frames. The first one (*Verb Example Extractor*) is a subsidiary tool that extracts simple example sentences (along with their syntactic frames) for a given verb from text corpora. The produced examples for a verb serve to orient the frame constructor (supposed to be an expert in linguistics) while he/she is developing the verb frames using the functionality of the *Frame Editor*. The basic tool used for development of the *BVN* is the so called *Frame Editor*. The main purpose of the *Frame Editor* is to automate the construction of valency frames for *wordnet* verb synsets in a given language. It is designed as a universal tool for construction of frames, no matter of the language to which they belong.

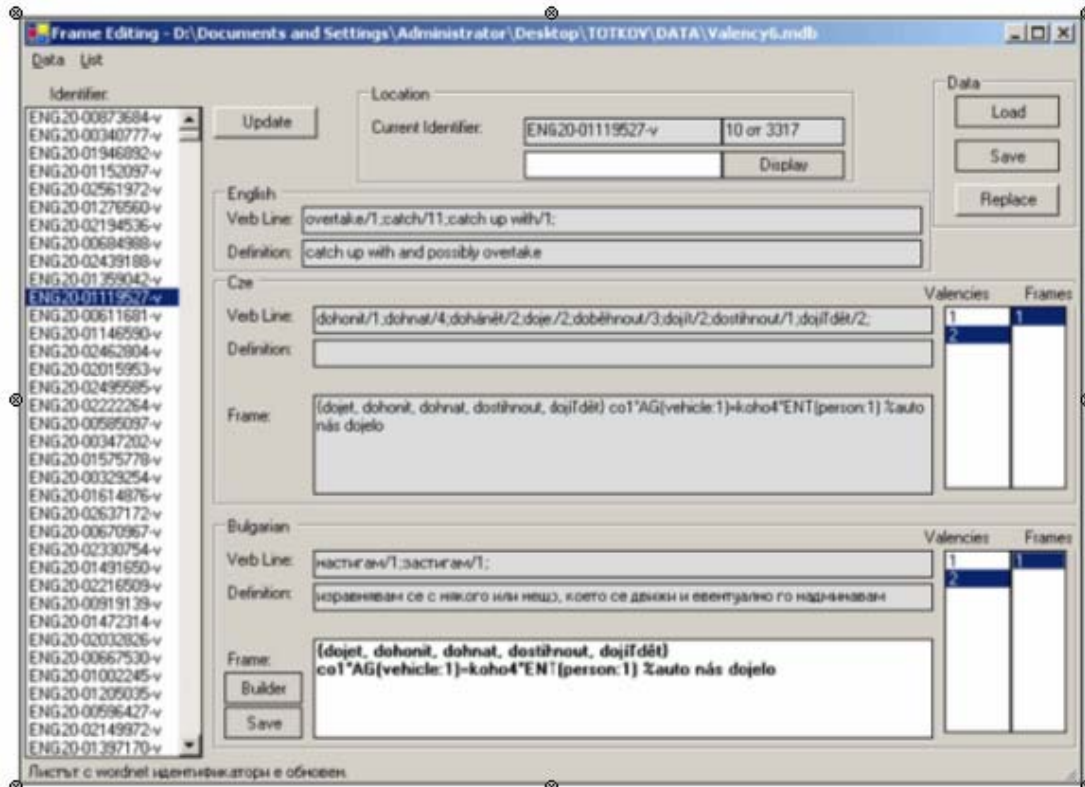


Figure 17. Basic options of the *Frame Editor*

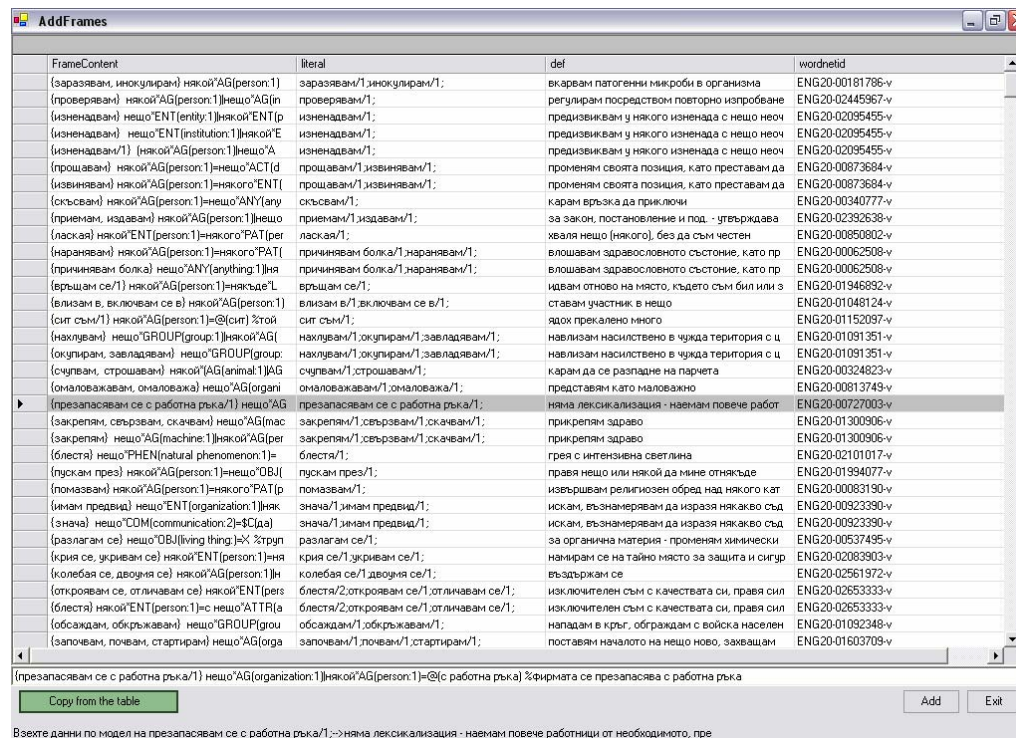


Figure 18. Selection of an existing frame in the *Frame Editor*

Its functionality can be described globally as follows:



- Frame construction for a target language *wordnet* (*TLWN*) using already developed frames of a source language *wordnet* (*SLWN*);
- Construction of new frames for the *TLWN*.

For now, the tool is applied in the case where the *SLWN* is *CzWN* and the *TLWN* is *BWN*.

The basic options of the *Frame Editor* allow (Figure 1.):

- Transformation of the *SLWN* frames into *TLWN* frames (using ILI links between the synsets);
- Automatic replacement of key words in the *TLWN* frames (using a special subsystem *DB TextReplace*);
- For a given ILI number, viewing simultaneously the corresponding synsets from *EWN*, *SLWN* and *TLWN* along with sense definition and frames (where applicable);
- Manual *TLWN* frames editing (using Remove and Edit mode).

For the cases where the frames of a verb synset are not identical in the *SLWN* and the *TLWN*, the advanced options of the system provide various powerful possibilities (Figure 2.): visualization of the list of all (already created) *TLWN* frames; selection and editing of an existing frame as a frame basis for the chosen verb using *Copy* and *Edit* mode, etc.

In addition the advanced options include a special subsystem *Build* that allows (Figure 3.):

- Visualization of frames by their separation in arguments/roles;
- Automated maintenance of a template library of frame argument lists, based on *TLWN* frames;
- Construction of verb frames using all frames for the corresponding ILI from the *TLWN* or *SLWN*; using all different arguments from already developed *TLWN* frames, as well as from any argument list from the template library.

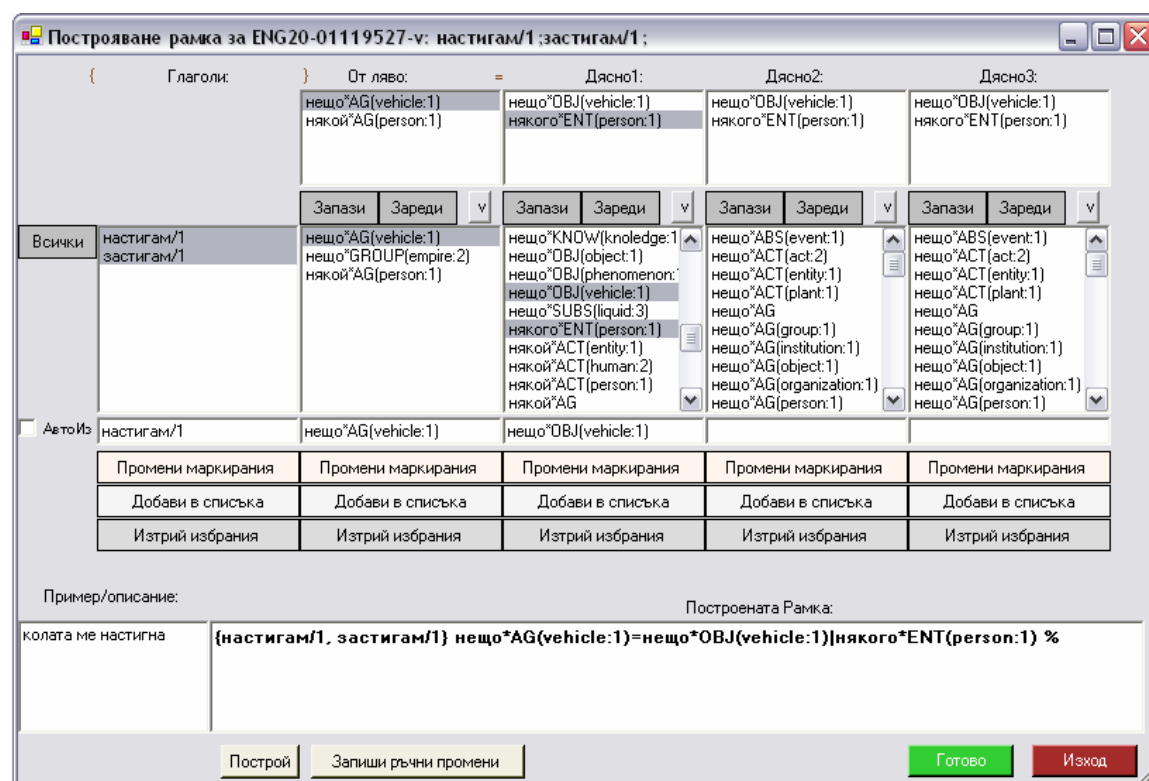


Figure 19. Build subsystem of the *Frame Editor*

Furthermore the *Frame Editor* helps in creation of frames for a verb synset from the *TLWN* in cases when the corresponding synset (via the ILI number) in the *SLWN* doesn't exist or its frames are not developed yet.

The Frame Editor allows construction of verb frames without analogues using:

- all preliminary developed frames for the *TLWN*;
- *TLWN* or *SLWN* frames of the ascending synsets (according to the hypernym relations from the *WN*) of the worked one;
- *TLWN* or *SLWN* frames of the descending synsets (according to the hyponyms relations) of the worked one;
- any argument list from the template library (e.g. created on the basis of a verb classification).

The frames for 1090 Bulgarian verb synsets are created till now. The BVN total number of frames: is 1587 and the number of corresponding unique frames – 846. About 25% of the BG verb valency frames completely coincide with the Czech ones. Some results we have arrived at (see and Table 9. Czech valence frames results) for are shown in the following table

| Verb frame                       | Frequency | Sense characterization               |
|----------------------------------|-----------|--------------------------------------|
| AG (person:1) = PAT (person:1)   | 95        | relations between persons            |
| AG (person:1) = OBJ (object:1)   | 58        | manipulating with objects            |
| AG (person:1) = ACT (act:2)      | 39        | solving tasks, performing activities |
| AG (person:1) = LOC (location:1) | 33        | motion verbs                         |
| AG (person:1) = SUBS (food:1)    | 13        | verbs of eating                      |

**Table 9:** Bulgarian Valence frames results

### Conclusion

A software tool for semantic text analysis, based on BVN, is in an experimental stage. On the other hand the methodology for BVN application is already developed and experimented. For example, a method [Totkov'90] (related to the MVC of the Bulgarian verb) is laid at the foundations of algorithms for semantic recognition of the meaning of unknown words [Totkov, Tanev'99]. The idea of a computer experiment called *Semantic wave* [13] for “extraction” of semantic characteristics of words and phrases from the input text, including “unknown” verb frames is related again to the use of a database similar to BVN.

The linguistic module of the tool is to be expanded with more heuristics and grammar rules for extraction of phrases and word sense disambiguation. It is supposed that in future the BWN will contain the basic models of subcategorization (not only for verbs) and that the experimented software tools for automatic extraction of knowledge for the semantics of Bulgarian words will be fully developed. At this stage a classification of the Bulgarian verbs, suitable for NLP implementation, doesn't exist. On the basis of the collected information (statistics) about concrete verb frames and using the BWN relations (hyperonym, verb groups, etc.), an attempt for an automated verb clustering and classification of the Bulgarian verbs will be made. The creation of a suitable model of the Bulgarian verb system would allow the improvement of the tools for BVN building.

## Using Balkanet as a training environment for students – the Romanian experience<sup>5</sup>

It is worth mentioning that, during a period of time that preceded the Balkanet project, the UAIC Romanian partner has developed activities with students in Computer Science oriented towards the acquisition of a preliminary set of Romanian synsets. This set was afterwards used during the developing process as one of the sources for the Romanian wordnet. The activity in itself, apart from having a certain practical value in the acquisition of the Romanian wordnet, had also a pedagogical merit since it has given the students a very clear idea of wordnet and the technologies addressing the issue of its acquisition.

Then, during a period of 6 year, that includes also the project's time, at least 20 diploma and dissertation thesis have been developed with students in Computer Science and the master studies in Computational Linguistics from the "A.I. Cuza" University of Iasi. These thesis had topics that covered wordnet lexical issues as well as interface developments, software for the exploitation of wordnet and applications using wordnet.

During the years of the project, the Balkanet database was intensively used as a teaching environment for master students in Computational Linguistics. Not only that they have acquired the basic knowledge of wordnet and Balkanet, and the skills to use the Balkanet interfaces, mainly VisDic, but they were taught the basic technology for the acquisition of Romanian synsets through the WNBuilder acquisition tool (Tufis&Barbu, 2004). Moreover, we have used Balkanet as a platform on which to test diverse scientific suppositions, potentially opening paths for valuable research efforts. For instance, in a group exercise pursued during the university year 2003-2004 we investigated with our students the feasibility of using Balkanet for the detection of semantic structures for automatic translation. We have given to 10 groups of students, 4 members each, the George Orwell's "1984" English-Romanian aligned parallel corpus, initially tagged in both languages to part-of-speech, and instructed them to recognize senses of words, to annotate these senses conforming to two aligned lexical thesaurus, the PWN and the ROWN, and to build parallel semantic frames of translation-equivalent verbs. More precisely, their task was:

- to find all verb occurrences in English and to sort them in the descending order of their frequency;
- among the most frequent verbs, each group had to choose 10 English verbs and to select from the parallel corpus all the language-pair segments of occurrences;
- they had to annotate the occurrences of these verbs, in both English and Romanian, to senses (using the ILI codes), according to both PWN and RoWN;
- subcategorisation constituents of verbs had to be annotated: their syntactic role, the head word, and the sense of the head word – using also the ILI codes;
- then, students had to select all occurrences in which a verb was considered to have the same sense and to generalize a semantic frame out of the set of constituents found around it. For a given constituent, say the direct object role, the generalization had to be the lowest concept in the wordnet hierarchy subsuming all senses of head words found on the role of direct object in the selected examples. If no generalization of this kind could be found, due to the fact that, for each part-of-speech, wordnet contains a collection of graphs, not just one, the union of the lowest computed role-concepts was computed;

---

<sup>5</sup> Cf. (Cristea et al, 2004)

- the final goal was to report a collection of English-Romanian frames around verbs that have given rise to parallel translations, which could be considered the kernel of a semantic transfer grammar.

The experiment was described in a paper presented in the LREC 2004 workshop on Teaching Computational Linguistics (Cristea et al, 2004). Although the results were rather unequal, from spectacularly good to poor, overall the project was successful, since, on one side it has given us a first indication on the feasibility of the problem of frame alignment and, on the other hand, was an extremely interesting working theme for students in Computational Linguistics. They have acquired thus a very clear sense of the advantages of using annotated corpora in NLP, and they learned the technology to obtain and exploit annotated corpora. Moreover, the best rated projects have thrown the seeds for furthers master and Ph.D. level research.

## Current Status of the Balkan Wordnets

### *Status of the Greek Wordnet*

|                  |        |
|------------------|--------|
| Synsets          | 18461  |
| Nouns            | 14426  |
| Verbs            | 3402   |
| Adjectives       | 617    |
| Adverbs          | 16     |
| Literals         | 24366  |
| Literals/ synset | ~ 1,33 |

#### Base Concepts

|     |      |
|-----|------|
| BC1 | 1218 |
| BC2 | 3462 |
| BC3 | 3825 |

|                           |       |
|---------------------------|-------|
| Domain specific synsets   | 238   |
| Law                       | 75    |
| Politics                  | 72    |
| Economy                   | 91    |
| Balkan specific synsets   | 309   |
| Greek specific synsets    | 52    |
| Lexico-semantic relations | 24368 |

|                |       |
|----------------|-------|
| HYPERNYM       | 18324 |
| HOLO_MEMBER    | 1320  |
| HOLO_PART      | 2660  |
| HOLO_SUBSTANCE | 57    |
| HOLO_PORTION   | 162   |
| VERB_GROUP     | 424   |
| BE_IN_STATE    | 143   |
| SUBEVENT       | 132   |
| CAUSES         | 76    |
| ALSO_SEE       | 210   |
| SIMILAR_TO     | 46    |
| DERIVED        | 103   |
| NEAR_ANTONYM   | 689   |
| ANTONYM        | 22    |

### *Status of the Turkish Wordnet*

|                           |        |
|---------------------------|--------|
| Synsets                   | 14,626 |
| Nouns                     | 11,059 |
| Verbs                     | 2,725  |
| Adjectives                | 802    |
| Adverbs                   | 40     |
| Literals                  | 20,310 |
| Literals/ synset          | 1.39   |
| Base Concepts             |        |
| BC1                       | 1,220  |
| BC2                       | 3,479  |
| BC3                       | 3,794  |
| Domain-specific synsets   | 300    |
| Law                       | 100    |
| Politics                  | 100    |
| Economy                   | 100    |
| Balkan-specific synsets   | 103    |
| Turkish-specific synsets  | 204    |
| Lexico-semantic relations |        |
| HYPERNYM                  | 12,197 |
| HOLO PART                 | 1,746  |
| NEAR ANTONYM              | 1,500  |
| HOLO MEMBER               | 1,114  |
| ALSO SEE                  | 973    |
| VERB GROUP                | 924    |
| BE IN STATE               | 608    |
| SIMILAR TO                | 311    |
| HOLO PORTION              | 230    |
| SUBEVENT                  | 131    |
| CAUSES                    | 100    |

### ***Status of the Romanian Wordnet***

|                   |       |
|-------------------|-------|
| Synsets           | 19839 |
| Nouns synsets     | 13345 |
| Verb synsets      | 4808  |
| Adjective synsets | 852   |
| Adverb synsets    | 834   |
| Token literals    | 33690 |

|  |              |
|--|--------------|
| Type literals                            | 19511        |
| The medium length of synsets             | 1.70         |
| The average number of senses per literal | 1.72         |
| Valency frames                           | 477          |
| Balkanet Common Set                      |              |
| BCS1                                     | 1218         |
| BCS2                                     | 3471         |
| BCS3                                     | 3827         |
| <b>Lexico-semantic relations</b>         | <b>25885</b> |
| <b>Hypernym</b>                          | <b>18134</b> |
| <b>holo_part</b>                         | <b>112</b>   |
| <b>holo_part</b>                         | <b>1174</b>  |
| <b>also see</b>                          | <b>422</b>   |
| <b>similar to</b>                        | <b>899</b>   |
| <b>verb_group</b>                        | <b>1050</b>  |
| <b>near_antonym</b>                      | <b>1660</b>  |
| <b>holo_member</b>                       | <b>797</b>   |
| <b>causes</b>                            | <b>129</b>   |
| <b>be_in_state</b>                       | <b>558</b>   |
| <b>subevent</b>                          | <b>195</b>   |
| <b>category_domain</b>                   | <b>755</b>   |
| Domain specific synsets                  | 286          |
| Law                                      | 98           |
| Politics                                 | 89           |
| Economy                                  | 99           |
| Balkan specific synsets                  | 151          |
| Romanian specific synsets                | 545          |

### ***Status of the Serbian Wordnet***

|  |        |
|--|--------|
| Synsets                                  | 8059   |
| Nouns                                    | 5919   |
| Verbs                                    | 1803   |
| Adjectives                               | 324    |
| Adverbs                                  | 13     |
| Literals                                 | 13295  |
| The medium length of synsets             | ~ 1.65 |
| The average number of senses per literal | ~ 1.21 |
| Base Concepts                            |        |
| BC1                                      | 1219   |
| BC2                                      | 3469   |

|                             |                      |
|-----------------------------|----------------------|
| BC3                         | 1369                 |
| Domain specific synsets     | 305                  |
| Law                         | 103                  |
| Politics                    | 101                  |
| Economy                     | 101                  |
| Balkan specific synsets     | 117                  |
| Serbian specific synsets    | 206                  |
| Lexico-semantic relations   | 12787                |
| HYPERNYM                    | 7601                 |
| HOLO_MEMBER                 | 956                  |
| HOLO_PART                   | 423                  |
| HOLO_PORTION                | 39                   |
| VERB_GROUP                  | 154                  |
| BE_IN_STATE                 | 176                  |
| SUBEVENT                    | 68                   |
| CAUSES                      | 54                   |
| ALSO_SEE                    | 116                  |
| SIMILAR_TO                  | 16                   |
| DERIVED                     | 114                  |
| DERIVED-POS                 | 42                   |
| DERIVED-GENDER              | 20                   |
| NEAR_ANTONYM                | 533                  |
| PARTICLE                    | 9                    |
| CATEGORY_DOMAIN             | 242                  |
| Usage (examples)            | 718 (in 630 synsets) |
| Glossies                    | 8035 (99.7%)         |
| Morphosyntactic information | 7870 literals        |
| non-lexicalized concepts    | 36                   |

### ***Status of the Czech Wordnet***

|                   |       |
|-------------------|-------|
| Synsets           | 28456 |
| Nouns synsets     | 21009 |
| Verb synsets      | 5155  |
| Adjective synsets | 2128  |
| Adverb synsets    | 164   |
| Literals          | 43918 |



|                           |       |
|---------------------------|-------|
| Literals / synset         | ~1.54 |
| Valency frames            | 1344  |
| Balkanet Common Set       |       |
| BCS1                      | 1218  |
| BCS2                      | 3471  |
| BCS3                      | 3827  |
| Lexico-semantic relations | 25683 |
| hypernym                  | 24312 |
| holo_part                 | 357   |
| holo_member               | 1781  |
| also see                  | 769   |
| similar to                | 1138  |
| verb_group                | 936   |
| near_antonym              | 1798  |
| holo_member               | 1089  |
| causes                    | 119   |
| be in state               | 602   |
| Subevent                  | 225   |
| Category domain           | 1136  |
| Domain specific synsets   | 304   |
| Law                       | 103   |
| Politics                  | 101   |
| Economy                   | 100   |
| Balkan specific synsets   | 257   |
| Czech specific synsets    | 257   |

### ***Status of the Bulgarian Wordnet***

|  |        |
|--|--------|
| Synsets                                  | 21441  |
| Nouns                                    | 14174  |
| Verbs                                    | 4169   |
| Adjectives                               | 3089   |
| Adverbs                                  | 9      |
| Literals                                 | 44956  |
| The medium length of synsets             | ~2.09  |
| The average number of senses per literal | ~ 1.36 |
| Base Concepts                            |        |
| BC1                                      | 1218   |
| BC2                                      | 3471   |
| BC3                                      | 3827   |
| Domain specific synsets                  | 2065   |
| Law                                      | 1007   |
| Politics                                 | 365    |
| Economy                                  | 693    |

|                             |                       |
|-----------------------------|-----------------------|
| Balkan specific synsets     | 220                   |
| Bulgarian specific synsets  | 116                   |
| Lexico-semantic relations   | 28599                 |
| HYPERNYM                    | 18370                 |
| HOLO_MEMBER                 | 921                   |
| HOLO_PART                   | 1388                  |
| HOLO_PORTION                | 114                   |
| VERB_GROUP                  | 882                   |
| BE_IN_STATE                 | 622                   |
| SUBEVENT                    | 182                   |
| CAUSES                      | 108                   |
| ALSO_SEE                    | 1186                  |
| SIMILAR_TO                  | 1594                  |
| DERIVED                     | 1166                  |
| NEAR_ANTONYM                | 2010                  |
| PARTICLE                    | 56                    |
| BG_DERIVATIVE               | 7920                  |
| Extralinguistic relations   | 61                    |
| USAGE_DOMAIN                | 29                    |
| REGION_DOMAIN               | 32                    |
| Usage (examples)            | 9920 (in 630 synsets) |
| Glosses                     | 21441(100 %)          |
| Morphosyntactic information | 41 099                |
| Valency frames              | 1 165                 |
| Non-lexicalized concepts    | 226                   |

## BalkaNet's Applications

### *Objectives and Current Status*

A critical element while building BalkaNet was not only to develop a rich structured sense inventory for the languages in question, but also to develop a scalable resource that would be utilized by various NLP applications and user communities. To that end we decided to incorporate BalkaNet in an IR system, in an attempt to provide end users with meaningful search results. BalkaNet's incorporation in an IR system is a continuous task that the consortium wishes to constantly improve. Within the scope of BalkaNet project the following tasks have been accomplished:

A web search engine has been launched. The engine indexes English documents as well as documents in all Balkan language represented within BalkaNet. Several components have been implemented and incorporated in the engine, ranging from query expansion modules, to domain search capabilities and organization of the indexed documents into topical directories. The main intuition for employing BalkaNet's shared ontology towards IR is that the ontology could be used as a deep *conceptual map* of the data sources stored by Web search engines, allowing thus information seekers to navigate within the Web's conceptual space. The conceptual ontology can also help search retrieval algorithms deal with the word mismatch problem by making connections between terms used in a search request and semantically related terms that might be found in the indexed documents. In this respect, a core infrastructure that employs BalkaNet ontology as a guide towards a more meaningful organization of the data sources that are indexed by Web search engines was developed. The conceptual indexing approach combines knowledge representation techniques and classical approaches for indexing words, so as to perform content-based IR as opposed to exact keyword matching.

Conceptual indexing is a process that, given a set of document's keywords, tries to map these keywords onto the available conceptual taxonomies and, based on that knowledge, to decide the conceptual domain under which the given document would be indexed. In this direction BalkaNet was employed as the conceptual knowledge resource that would be utilized by a Web search engine in order to organize indexed documents. To reassure that BalkaNet ontology would be effectively employed by the search engine, an additional layer of semantic information was incorporated into BalkaNet's Inter-Lingual-Index. This layer concerns conceptual domains knowledge and was appended to the nodes of the ILI's hierarchies. The nodes of the ILI's taxonomies are linked to conceptual domains and, through the transitivity of the taxonomic ILI links, the domains knowledge are transferred to all ILI nodes belonging to the respective taxonomy. Conceptual domains are treated as conceptual ontologies and serve to the transfer of the respective semantic attributes within monolingual wordnets and across the ILI network. BalkaNet's conceptual domains emerged from the thematic areas of approximately a 410,000 Web document collection that we have indexed in a local Web search engine. In particular, a search engine that indexes multilingual documents from the Balkan Times Web site (<http://www.balkantimes.com>) has been developed. Web documents hosted by the respective website follow a preliminary classification into major thematic categories, such as politics, law, economy, religion, etc. Out of those categories three were selected, namely Law, Economy and Politics that formed the conceptual domains into which BalkaNet's taxonomies would be structured. Having defined the conceptual clusters into which Web documents would be organized, the SUMO ontology was employed of which all ILI concepts falling into any of the pre-defined conceptual domains were extracted. All ILI hierarchies that belong to the SUMO ontology domains are marked-up with explicit domain information, which is automatically transferred to the corresponding monolingual wordnet taxonomies through inter-ILI equivalence links.

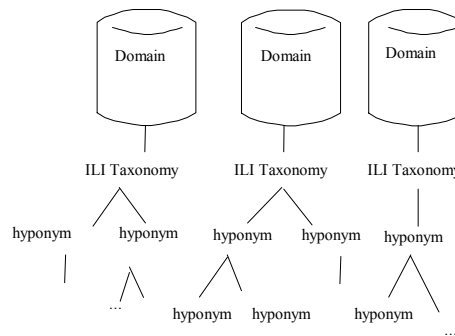
Following web documents morphosyntactic processing and keywords' extraction, conceptual indexing takes place via an internal mapping between documents' high-weighted terms and

ILI nodes and by calculating the distance of the conceptual nodes within the taxonomy. Conceptual distance reflects semantic similarities between terms and tackles sense ambiguity issues in case a term is distributed over several ILI nodes. Based on the distance of concepts in wordnets' graphs sophisticated modules that have been incorporated in the engine calculate the thematic category of a Web document and index it into the respective index. In case that multiple indexed match a document's subject then the document is indexed into all matching domains.

### **Introducing Conceptual Domains in BalkaNet ILI**

The main rationale for structuring the ILI is that a language independent conceptual taxonomy employed as the backbone of a conceptual indexing infrastructure would result in a semantically meaningful organization of the indexed data. In order to utilize the conceptual taxonomy to efficiently locate where in the taxonomy a concept belongs to, it is necessary to first organize the concepts of the taxonomy in such a way so that every concept has explicit pointers to its most specific concepts (hyponyms) and from its most general concepts (hypernyms).

In addition, we introduce the notion of *conceptual domains*, which are treated as conceptual ontologies and which serve to the transfer of the respective semantic attributes within monolingual Wordnets and across the ILI network. The BalkaNet ILI is organized as a set of conceptual taxonomies for certain conceptual domains, which are inherited from the SUMO ontology (<http://ontology.teknowledge.com/>). SUMO is an upper ontology that contains concepts general enough to address a broad range of domain areas. Concepts specific to particular domains are included within ILI's taxonomies, whereas SUMO provides a structure upon which ontologies need to be constructed for particular domains. The architecture of the conceptual taxonomies linked to the SUMO ontology domains is illustrated in Figure 16. We chose SUMO as a base ILI ontology for three reasons. First and foremost, it was already mapped to Princeton WordNet's synsets, which are contained in the Balkanet ILI. Secondly, it combines resources from many fields, and, most importantly, it is freely available and extensible.



**Figure 16:** Balkanet ILI classified taxonomies

Each element of a conceptual domain is built into a taxonomic structure and each taxonomy links concepts that belong to that particular domain. All ILI hierarchies that belong to the SUMO ontology domains are marked-up with explicit domain information, which is automatically transferred to the equivalent monolingual Wordnet taxonomies through inter-ILI equivalence links. This way, conceptual domains are assigned automatically to monolingual Wordnet synsets.

### **Conceptual Indexing Using Domain Taxonomies**

To demonstrate the potential that conceptual taxonomies have in Web indexing, we employ the BalkaNet shared ontology as a baseline for a more meaningful organization of the data records that are to be indexed by Web search engines. The main component of our conceptual indexing approach is a conceptual classification formula, which clusters the contents of the engine's index on the basis of their topical relations and semantic similarity. To perform con-

ceptual clustering, we treat ILI's conceptual domains as topics under which Web documents are classified. Conceptual clustering takes place via an internal mapping between documents' representative terms and ILI's concepts, and by calculating their semantic similarity. Based on corresponding index terms, each document is assigned to a specific domain(s).

The first step towards classification concerns the morphological pre-processing of documents in order to extract a core set of lexicalized concepts, represented in each document. To address multilingual conceptual indexing, the clustering module employs the language denoting tags accompanying each document as a guide towards morphological processing and towards the use of the information encoded within the respective monolingual wordnet. Morphological processing involves document tokenization, part-of-speech tagging and lemmatization. Henceforth, term weighting schemes (for example the normalized  $tf*idf$  formula (Salton and Buckley, 1988)) are employed against all documents' content terms<sup>6</sup>. Terms with high frequency weights are those that lexicalize the most representative concepts of a given document, and are the ones on which indexing and clustering are based. These terms are then located in the corresponding monolingual wordnet and their ILI's conceptual equivalents are retrieved simply by following the semantic links. Document clustering then takes place by traversing the conceptual taxonomies of the retrieved ILI nodes. The closer the matching nodes are to a topmost node (the shortest path), the more likely that a given document belongs to that cluster. However, relying exclusively on the idea of the shortest path for measuring conceptual distance is not sufficient per se for ensuring the successful conceptual clustering of documents. This is essentially the case where a document's terms are mapped against several ILI concepts, each of which belongs to a different taxonomy and whose distance from each taxonomy's root node are equal (or comparable). To account for such conflicting cases we allow for a document to be clustered under multiple conceptual domains.<sup>7</sup> For calculating conceptual distances we follow Resnik's (1995) approach that captures semantic similarity by means of the information content of the concepts in a hierarchical network. Conceptual distance is not only used to reflect semantic similarities between terms, but also to tackle sense ambiguities issues in cases a term is distributed over several ILI nodes.

## Challenges

Developing a language-independent, consistent and comprehensive conceptual ontology that can be used for semantic indexing is not an easy task. The major difficulty we encountered while structuring our sense inventory concerned inter-lingual alignment issues. In particular, we were challenged to incorporate (into the ILI) language-specific concepts that are common across the Balkan languages, but for which there were no lexicalized English counterparts. We tackled such cases by allowing complex ILI relations, an approach that reassures that ILI remains a language neutral conceptual knowledge base. Inter-ILI links also guarantee a level of consistency across wordnet mappings. Moreover, the adoption of the SUMO ontology domains helped us structure the ILI taxonomy in a meaningful way and gave us the flexibility to enrich the ILI with new concepts without imposing any need for structural changes. This flexibility is due to the percolation of the shared semantic attributes to all the concepts represented in each ILI taxonomy.

Further, the BalkaNet shared ontology can serve as a baseline for multilingual conceptual indexing. We have presented an approach that clusters documents according to the conceptual domains to which their representative terms belong. Documents can be classified under multiple domains, while the problem of ambiguous terms is addressed on the grounds of conceptual distances within the taxonomy. So far in our experiments we have used Resnik's approach to calculate semantic similarities, but we are also considering other approaches, like the conceptual density approach (Agirre and Rigau, 1996).

---

<sup>6</sup> As content terms we consider nouns, verbs, adjectives and adverbs.

<sup>7</sup> This way a document about *tuition fees* would be clustered under both *education* and *economy* domains.

The proposed approach for clustering documents based on the classified ILI taxonomy exhibits several advantages. One benefit for clustering ILI's taxonomies under the SUMO domains is that each taxonomy can be viewed as a domain-specific Wordnet and, as such, it can be employed by applications that require specialized knowledge sources. Another advantage of our structured ILI is that it can be extended with other languages and/or concepts without requiring any modifications. Moreover, the conceptual indexing infrastructure we have designed maintains distinct multilingual indices for each conceptual domain, a feature that makes the engine's repository manageable upon updates and has a strong potential in supporting specialized cross-lingual Web searches. In addition to indexing, the suggested classified and structured sense inventory enables the efficient maintenance of the ILI's hierarchies, and contributes in dealing with the proliferation of ILI's concepts among individual wordnets.

We believe that the BalkaNet shared ontology can be further used to improve IR performance by using conceptual indexing, as conceptual taxonomies have a strong potential in helping information seekers satisfy their needs. We argue that a core component of a conceptual retrieval system is a conceptual indexing module that groups indexed documents under conceptual domains on the basis of their semantics, and organizes them on the basis of their conceptual closeness. The objective of the conceptual taxonomy is, therefore, to feed the engine's indexing modules with information on the documents' semantics so as to index them under conceptual domains. Thus, the main idea for employing BalkaNet's shared ontology towards IR is that the ontology could be used as a deep conceptual map of the data sources stored by a Web search engine, allowing users to navigate within the Web's conceptual graph. In that respect, the conceptual ontology can help retrieval algorithms make connections between terms used in a search request and semantically related terms that might be found in the relevant indexed documents.

## **Steps for Web Documents Pre-processing**

### **Pre-process web pages**

HTML parsing, markup removal, tokenization, lemmatization and stop words elimination.

### **Lexical chains**

Segment the document into paragraphs, identified by the HTML source tags. If a paragraph tag (*p*) is not found in the HTML source, use shingling.

*Shingling*: Group 50 adjacent words of a page to form a shingle. Treat each shingle as a paragraph.

For every paragraph, generate lexical chains and compute their scores. If a paragraph produces multiple lexical chains, keep the chain of the highest score as the most representative chain.

Merge chains of all paragraphs and eliminate duplicate lexical elements. Take the remaining elements together and form a single chain, which is the chain of the whole document. Finally compute a new score for the new chain. This is the final chain upon which we will rely for finding the topic category of Web pages.

### **Find the topic category of a web page**

Elements of the lexical chains are mapped against the ontology's nodes and if a mapping is found, the ontology's hierarchies of the matching nodes are traversed up to the top level nodes (i.e., the topic categories concepts), following the IS-A links.

If all elements of a page's lexical chain map to the same ontology domain, then index the page into this domain.

If elements of a page's lexical chain map to several domains in the ontology, compute a *relatedness score* of every document to each of the ontology's matching domains.

$$\text{Relatedness score: } RScore(i, k) = \frac{\text{Score}(C_i) \cdot \# \text{ of } C_i \text{ elements of } D_k \text{ matched}}{|\# \text{ of } C_i \text{ elements}|}$$

Score ( $C_i$ ) is the score of the page's lexical chain

# of  $C_i$  elements is the number of elements in the lexical chain  $C_i$  that matched the hierarchy of a domain  $D_k$ .

Finally index the page in the domain to which it has the highest relatedness score above a threshold  $T$ .

$$IScore(i, k) = \max RScore(i, k) \text{ where } 1 \leq i \leq T$$

$$T = 0.5$$

For pages with relatedness score below  $T$  index them into all their matching categories.

## **Compiling and Processing a corpus of the BalkanTimes Web Archive**

With respect to the final project application, a significant amount of work has been devoted to the collection of a training data on which our conceptual indexing formula would be based. This training data comprises essentially a small multilingual corpus of 410K Web documents for all languages in question collected over a period of four weeks from the Southeast European Times web archive (<http://www.balkantimes.com>). The above resource provides news articles for all languages participating in the project besides Czech, which are already classified into thematic areas. Three of those themes have been selected for testing the project's results, these are: law, politics and economy and thus all texts classified into one of the themes of interest were extracted and stored in individual text files. For enriching our sample data with Czech documents, members of the FI MU team indicated a set of URLs that host Czech news articles which were also downloaded. This sample corpus was then processed separately by each partner (each one working on the monolingual part of the corpus pertaining to his language) in the way described below.

### Corpus Morphological Pre-processing

For each of the concerned languages found on the BalkanTimes site (Bulgarian, Greek, Romanian, Serbian, Turkish and English for reference) 3300 documents from the Economy, Justice and Politics category of the site were extracted from the search engine database. For this purpose a simple script was used that located for each language the respective documents using the news archive link pages and exported the content of the pages from their search engine cache copies. For the second stage of the extraction, the uniform structure of the documents was utilized in order for the text of each article to be isolated from the rest of the page. The text then was extracted from the documents, stripping them also of any HTML tag found in the article text and saved into a file named with the URL of the document. Since the BalkanTimes doesn't contain Czech documents, three sites in the Czech domain were located that contained archives of stories belonging to equivalent categories with three from the BalkanTimes:

- [www.epravo.cz](http://www.epravo.cz) for Justice
- [ihned.cz](http://ihned.cz) for Economy
- [www.blisty.cz](http://www.blisty.cz) for Politics

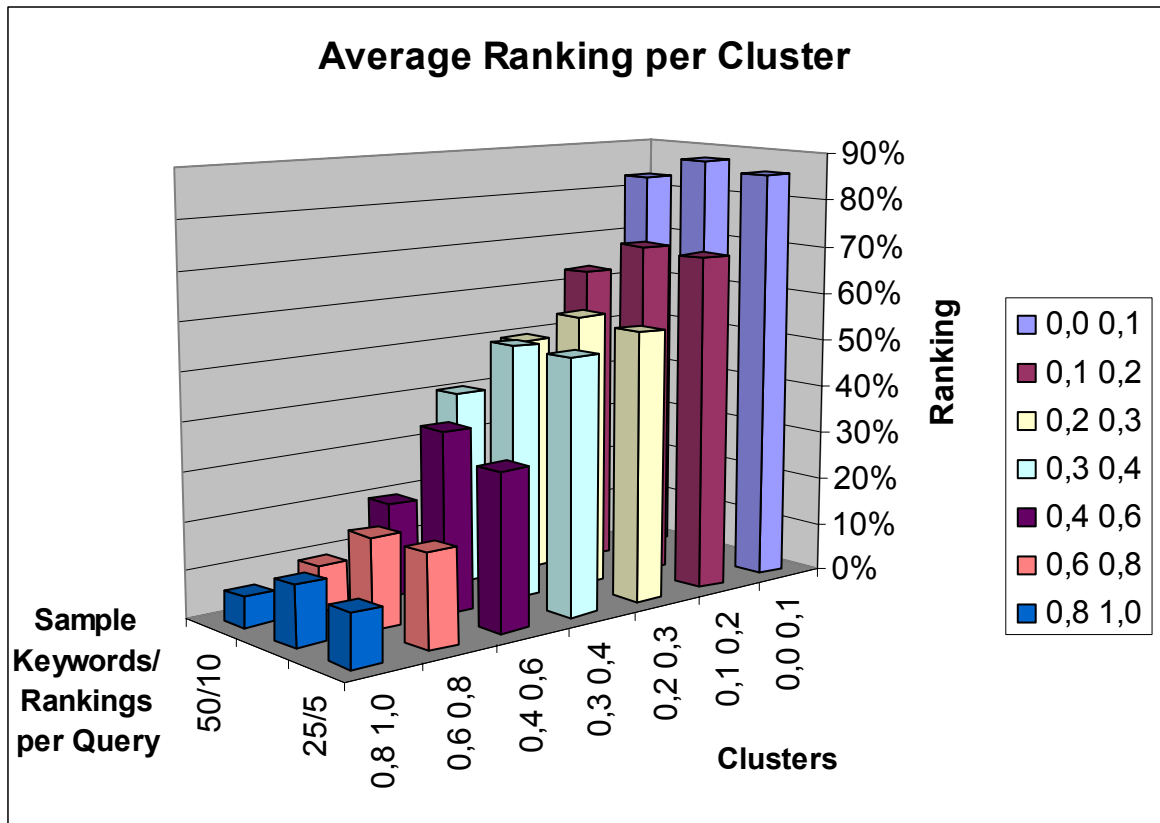
The same procedure as before was followed for the extraction of around 11000 Czech documents from the search engine database.

Following the extraction, each partner automatically tagged the documents for POS using their language-specific taggers. In order to keep the outcome homogeneous, a simple format was agreed that used the subset of the Penn-Treebank tag-set that refers to nouns, verbs, ad-

jectives and adverbs to denote the tags and ignores the rest of the set. Due to the inflectional system of some of the languages (e.g. Greek) a normalization procedure was also applied to the terms of the documents by each partner. The outcome of both the procedures was incorporated into simple tab-delimited files that respected the naming of the original documents.

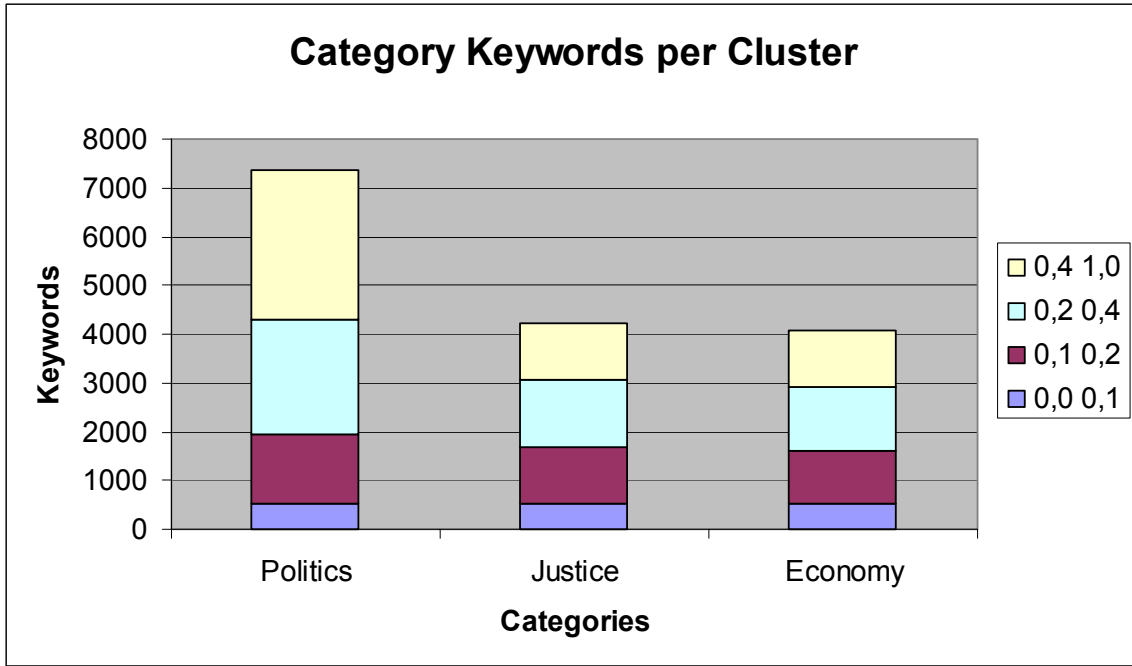
**Keyword Extraction**

Following the extraction and processing of the documents for all the languages, a procedure was decided for the extraction of keywords from the English documents. The reasoning behind this decision was twofold: first it complemented the domain-specific sense selection process and second it facilitated the conceptual indexing by narrowing down the variables (i.e. the terms) that had to be processed. For this purpose, and following the bibliography on this matter it was optioned to calculate weights for each term in each document and estimate threshold values for the keyword selection. The weight was calculated using a normalized variance of the TFIDF metric and the results were clustered into value groups. In order to locate the most effective cluster group, a variable number of randomly selected from the group terms were extracted and used as simple keywords in queries to the respective indexed documents of the search engine. For each cluster group the retrieval efficiency of the average search engine rankings for the respective keywords was calculated:

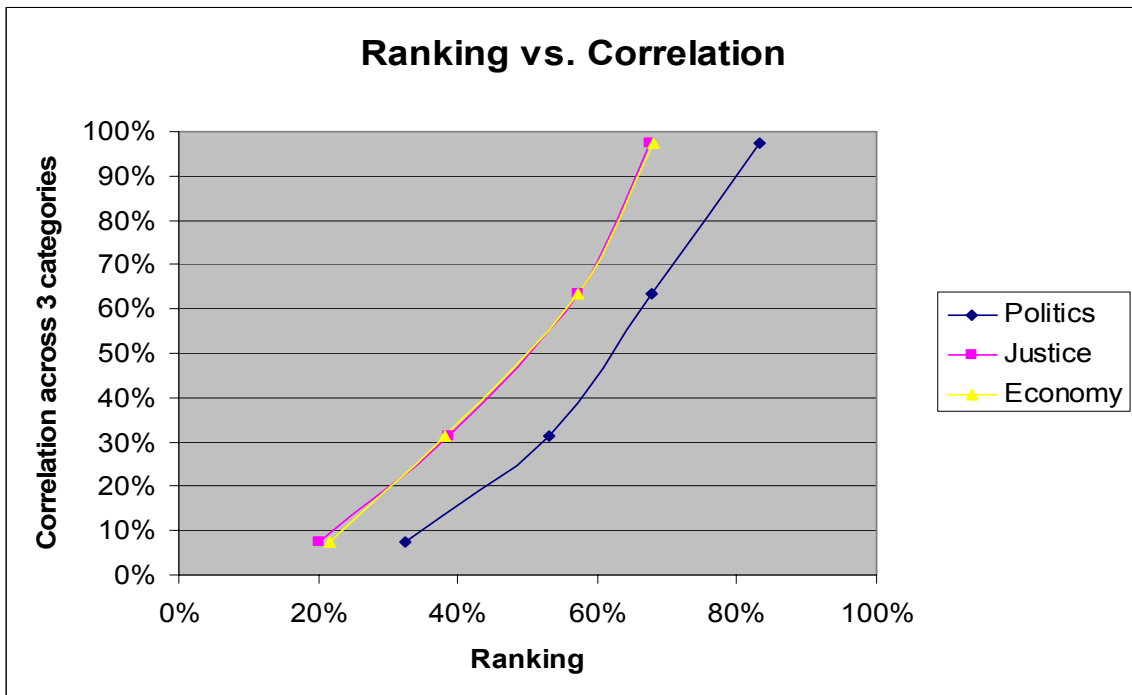


As shown in the diagram, the most efficient group was for weight values from 0.0 to 0.1 with the ranking degrading rapidly for the next groups. The fact that some of the clusters produced results that were very close was exploited by grouping them further and creating bigger clusters for weight values of 0.2 – 0.4 and 0.4 – 1.0. At the next stage, a list of (unique) keywords was extracted for each category and cluster group:





In order to check the suitability of the keywords as category discriminators, the correlation amongst the extracted lists for each category for each cluster group was examined. The results are shown in the figure:



The curve in the Ranking vs. Correlation shows the existence of a trade-off between Ranking and Correlation and additionally that the more unique keywords (after the removal of multiple occurrences) appear for the middle spectrum of the weight values. Given these two parameters, it was decided that the optimal solution was to choose the cluster group 0.2 – 0.4 which produced after the removal of the duplicate among lists 67 terms for Justice, 65 terms for Economy and 594 terms for Politics.

## ***Impact***

Having briefly outlined the work accomplished towards the project's application it is worth mentioning the core objective and the expected impact of the application. BalkaNet aimed at delivering a useful multilingual semantic network, whose usefulness would be deemed besides the availability of the lexical resources. That was the main reason why the consortium decided to incorporate BalkaNet network into a Web search engine and tests the network's contribution in delivering qualitative search result. Of course within the limited time frame of BalkaNet and given that the lexical networks should be developed from scratch, the project's application can by no means be seen as a complete and functional tool that is readily applicable. BalkaNet's aim was to perform a feasibility study on the network's application in a search engine and to this respect it has been successful. BalkaNet demonstrated that semantic networks for the languages in question can altogether be imported within the searching modules of an IR system. Moreover, the searching mechanisms and services built verify that multilingual IR for the Balkan language has now a significant starting point that could and should be explored by Internet Service Provider in the area.

The success of the project's application can be summarized in the following points: a multilingual Web search engine for six languages was launched and it currently indexes a large number of Web documents with weekly updates. For the first time, a multilingual wordnet is utilized by the indexing modules of an IR system in order to organize Web documents thematically, a query expansion module has been implemented that performs both monolingual and multilingual query expansion and which proves that the problem of multilingualism on the Web can be substantially alleviated for the lesser studies languages.

## ***Testing Specifications***

The consortium has defined a set of tests that could help Internet Service Providers, who will incorporate BalkaNet's results into their systems, evaluate the contribution of the semantic network. Members of the consortium will actively participate in the performance of the tests and will provide detailed feedback on the project's specifications and implementation approaches.

The tests should 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 various levels in the hierarchy and by making use of different kind of lexical information (ambiguous, polysemous terms etc.). Furthermore, it needs to be investigated the extend to which 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 should be based as a measurement of the additional functionality and quality of the monolingual wordnets. In addition, the queries shall 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 for evaluating the system's performance are summarized below:

- ❑ Precision scores obtained by the engine and relevance scores provided by end users and evaluators (i.e., relevance feedback)
- ❑ 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 need to be applied:

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

Some of these tests are underway and currently performed using the search engine provided by OTENET. However, in order to compare the acquired results with the performance of other systems we also need to test the performance of other systems that support documents and/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 handling single term queries since after all those are the most frequent types of queries issued in IR systems especially by inexperienced end users.

BalkaNet however, will be constantly improved so that its contribution to NLP tasks and applications is enhanced. It is our hope that BalkaNet is only the beginning for the development of IR players across the Balkan region.

## DISSEMINATING BALKANET

### ***Conferences, Workshops, Special Sessions***

Several actions have been taken from all members of the consortium towards the dissemination of the project's results. Participations in National and International Conferences and Workshops are excellent opportunities for stimulating the interest of the scientific community and end users.

The main awareness activities that have been performed are summarized below:

- Publication of a double special journal issue on BalkaNet. The issue was published by the Romanian Academy Journal Publishing House, in *Journal of Science and Technology*. There are 13 papers in approx. ~250 pages, a term glossary and a preface written by Dr. Christiane Fellbaum and Dr. Piek Vossen.
- The project member Faculty of Informatics at Masaryk University (FI MU) organized the 2<sup>nd</sup> International Global Conference (GWC) in Brno, Czech Republic (January, 2004). At the conference several presentations of the project and its results took place and were disseminated to a wider audience. Moreover, within the framework of the conference a special session was also organized especially devoted to BalkaNet project. In the session project participants presented to a large audience the main achievements accomplished by the project and demonstrated the technical infrastructure implemented within BalkaNet. This session contributed to BalkaNet's promotion due to the fact that a large group of specialized researchers and individuals attended the conference.
- Organization of a BalkaNet workshop in conjunction to the 3<sup>rd</sup> International LREC Conference, Las Palmas, May 2002. The workshop was entitled: "Wordnet Structures, Standardization and Applications (WSA) for Lesser-studied Languages" and aimed at bringing together researchers that have recently started developing their own Wordnets (e.g. Balkans, Scandinavians etc.), in order to exchange ideas on approaches for linguistic structures and architectures of semantic networks and demonstrate their preliminary results to a wider audience.

Furthermore, several presentations of the project have taken place in National and International Conferences, such as LREC 2002, GWC International Conference 2002, 9th International Conference on Computational Linguistics COLING 2002, International Conference *Romanian Language and Globalisation* 2002, International Conference on Information Communication Technologies in Education 2002, 27<sup>th</sup> International Conference ICT&P 2002, International Conference on Text, Speech and Dialogue 2003, *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts; Data Driven Machine Translation and Beyond* 2003, International Conference ICT&P'2003, *Balkan Conference in Informatics* 2003, 6<sup>th</sup> Intex Workshop 2003, GWC International Conference 2004, NLUCS 2004, COLING 2004, ACL 2004, LREC 2004, IEEE International Conference on Advanced Learning Technologies (ICALT 2004), 7<sup>th</sup> Intex Workshop 2004, DAARC 2004, Control and Information 2002 Conference, DIALOGUE 2003 Workshop, Computer Treatment of Slavonic Languages Workshop, 2003, ICT&P 2003 and 2004 Conferences, Conference on Automatisations and Informatics, 2004, etc. Finally, BalkaNet has been disseminated in several events and meetings across Europe, as for example in the Europrix Summer School, Salzburg, Austria (Sept. 2002), European Summer School on Logic, Language and Information, ESSLLI 2004 (August 2004, Nancy, France).

The results of the BalkaNet project were disseminated also at national level. To mention only the last year's activities, the Iasi branch of the Romanian Academy has hosted two invited

talks given by BalkaNet members. At the National Conference on Computer Human Interaction (Bucharest, 22-23 September, 2004) Dan Tufiş gave the invited talk: “BalkaNet: a Multilingual Lexical Ontology”<sup>8</sup>. A presentation of Balkanet (objectives and realisations) was done also at IRST-Trento by Dan Cristea (member of UAIC) on August 2004.

### ***Joining Global Wordnet Association***

Following a communication between the BalkaNet consortium and the steering board of the Global Wordnet Association, the consortium became actively involved to the Association’s activities, by joining GWA. Each contractor is responsible for providing guidance and advice to other wordnet developers as well as to monitor the feedback of the entire research and industrial community concerning the functionality and usefulness of the project’s results.

### ***User Groups /Promotion and awareness***

One of BalkaNet’s objectives is strengthening the ties between the academic and information technology communities in European countries. BalkaNet’s user group falls within a wide spectrum of institutions and individuals. In particular, academic as well as industrial parties have contacted members of the consortium in order not only to acquire more information on the project, but also to express their interest in further exploiting the project’s results in various NLP applications. Several of them have been admitted access to the project’s intermediate results on the grounds that they are exploited only for research purposes. Moreover, various well-known linguistic communities have expressed their interest in the project’s results and as such several publications and presentations of the project’s outcomes have taken place.

Additionally, due to the incorporation of BalkaNet’s results into a Web search engine, the consortium is continuously in contact with Internet Service Providers in order for the latter to embody BalkaNet’s content and technical infrastructure into their systems’ components. To this respect the contribution of the project’s end user, namely OTEnet, is valuable and has already expressed their intention in incorporating BalkaNet’s results into their commercial Web search engine. Moreover, concerning the dissemination of the project’s results some attempts have been performed by the consortium so as to develop flexible and modular components that would be adopted in a number of applications, ranging from IR query expansion to the development of services for the semantic web.

---

<sup>8</sup> D. Tufiş: BalkaNet: ontologie lexicală multilingvă. In Şt. Trăuşan, C. Probeanu (eds): Interacţiune Om-Calculator, Printech Pubs., Bucharest, 22-23 September, 2004, pp.9-22.

## References

- Agirre E. & Rigau G. (1996). Word Sense Disambiguation Using Conceptual Density. In *Proceedings of the COLING Conference*, Copenhagen, Denmark
- Artale A., Magnini B., Strapparava C. WordNet for Italian and Its use for Lexical Discrimination. In *Proceedings of the 5th Congress of the Italian Association for Artificial Intelligence*, Springer Verlag.
- Baker, F. Collin, J. Ruppenhofer, FrameNet's Frames vs. Levin's Verb Classes. In J. Larson and M. Paster (Eds.), In *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society* 2002, 27-38.
- Bilgin O., Cetinoglu O., Oflazer K. (2004). Morphosemantic Relations In and Across Wordnets: a Study Based on Turkish. In *Proceedings of the 2<sup>nd</sup> Global WordNet Conference (GWC)*, Brno, Czech Republic.
- Cristea, D., Teodorescu, H.-N. and Tufi, D. (2004): Student Projects in Language and Speech Technology. In *Proceedings of the LREC 2004 Workshop on CL Learning*, Lisbon, Portugal.
- Dorr J. B., M. B. Olsen, Multilingual Generation: The Role of Telicity in Lexical Choice and Syntactic Realization, *Machine Translation*, 11:1-3, 1996, 37-74.
- Fillmore C. J., Baker C. F., Frame Semantics for Text Understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL*, Pittsburgh, June, 2001.
- Hirst G. and St-Onge D. (1998) Lexical chains as representations of context for the detection and correction of malapropisms, in Fellbaum C. (ed.) *WordNet: An Electronic Lexical Database*, The MIT Press, Cambridge, MA
- Horák, A., Smrž, P. (2004) VisDic –Wordnet Browsing and Editing Tool. In *Romanian Journal of Information Science and Technology*, Vol. 7(1-2) (2004)
- Ion, R., Tufiş, D. (2004): Multilingual Word Sense Disambiguation Using Aligned Wordnets. In *Romanian Journal of Information Science and Technology*, Vol. 7(1-2) 1-35 (2004)
- Kilgariff A. and Yallop C. (2000) What's in a thesaurus?, in *Proceedings of LREC-2000*, Athens, Greece
- Kingsbury P., M. Palmer, M. Marcus, Adding Semantic Annotation to the Penn TreeBank, In *Proceedings of the Human Language Technology Conference*, San Diego, California. 2002.
- Kipper K., H. T.Dang, M. Palmer, Class-Based Construction of a Verb Lexicon. *AAAI-2000 Seventeenth National Conference on Artificial Intelligence*, Austin, TX, July 30 - August 3, 2000.
- Koeva S. (2004) Arguments – semantic relations and semantic realizations. in *Argument structure. Problems of simple and clause sentences*. Sofia: Sema RS, 2004, 19-34.
- Koeva S. (2004) Theoretical model for a formal representation of syntactic frames, *Scripta and e-Scripta*, Vol.2 , Sofia, 2004, 9-26.
- Koutsoubos I.D., Andrikopoulos V., Christodoulakis D. (2004). Wordnet Exploitation through a Distributed Network of Servers. In *Proceedings of the 2<sup>nd</sup> International Global Wordnet Conference*, Brno, Czech Republic.
- Lenat D.B., (1995). CYC: A Large-Scale Investment in Knowledge Infrastructure. In *Communications of the ACM*, vol. 38, no.11.
- Levin, B. (1993) *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago
- Miller G. (1990). Five Papers on WordNet. *Special Issue in International Journal of Lexicography*, vol.3, no.4.
- Miller G.A., Beckwith R., Fellbaum C., Gross D., Miller K.J. (1990). Introduction to WordNet: An On-Line Lexical Database. In *International Journal of Lexicography*, vol. 3, no. 4 (winter 1990), pp. 235-244.
- Miller, G.A. WordNet: An Online Lexical Database. *International Journal of Lexicography* 3(4) (special issue), 1990.

- Niles I. & Pease A. (2001). Towards a Standard Upper Ontology. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS), Ogunquit, Maine, pp. 2-9
- Pala K. and P. Smrz, Building Czech Wordnet, Romanian Journal of Information Science and Technology, Dan Tufis (ed.), Special Issue on BalkaNet, Romanian Academy, vol. 7, no. 2-3, 2004
- Pavelek Th. & Pala K. (2002). VisDic: A new Tool for WordNet Editing. In *Proceedings of the 1<sup>st</sup> International Global Wordnet Conference*, Mysore, India
- Pavelek, T., How to Convert Wordnets into XML representation, www pages, NLP Lab. FI MU, Brno 2001 (<http://nlp.fi.muni.cz/projekty/mt/visdic/ewn2visdic.html>).
- Popova M., Short Valency Dictionary of the Bulgarian Verbs, Sofia, BAS, 1987.
- Princeton WordNet 2.0 <ftp.cogsci.princeton.edu>
- Quillian R. (1968). Semantic Memory. In M. Minsky (Ed.) *Semantic Information Processing*, MIT Press, Cambridge, M.A., pp.216-170
- Quine, Willard Van Orman (1960): *Word and Object*. The MIT Press, Cambridge, MA
- Resnik P., (1995). Disambiguating Noun Groupings with Respect to WordNet Senses. In Proceedings of the 3<sup>rd</sup> Workshop on Very Large Corpora, MIT, pp. 54-68
- Rodriguez, H., Climent, S., Vossen, P., Bloksma, L., Peters, W., Alonge, A., Bertagna, F., Roventini, A.(1998). The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. In Piek Vossen (ed.) *EuroWordNet: A Multilingual database with lexical semantic networks*, Computers and Humanities, Vol. 32, Nos 2-3.
- Roventini A., Alonge A., Bertagna F., Magnini B. and Calzolari N. ItalWordNet: a large semantic database for Italian. Proceedings of LREC-2000, *Second International Conference on Language Resources and Evaluation*, pp. 783-790, Athens, Greece, 2000.
- Salton G. & Buckley C. (1988). Term Weighting Approaches in Automatic Text Retrieval In *Information Processing and Management*, 24(5): 513-523
- Stamou S., Oflazer K., Pala K., Christodoulakis D., Cristea D., Tufis D., Koeva S., Totkov G., Dutoit D., Grigoriadou M. (2002b). BALKANET: A Multilingual Semantic network for the Balkan Languages. In *Proceedings of the 1<sup>st</sup> International Global Wordnet Conference*, Mysore, India.
- Stamou S., Nenadic G., Christodoulakis D. (2004) Exploring BalkaNet Shared Ontology towards Multilingual Conceptual Indexing. To appear in *Proceedings of the 4<sup>th</sup> Language Resources and Evaluation Conference (LREC)*, Lisbon, Portugal.
- SUMO <http://ontology.teknowledge.com/#FOIS>
- Totkov G., Ch. Tanev, Computerised extraction of Word Semantic through Connected Text Analysis, in A. Narin'iyani (eds.), *Computational Linguistics and its Applications*, Proc. of the International Workshop DIALOGUE'1999, 360-365.
- Totkov G., Conceptual and Computer Modelling of Language Structures and Processes (in case of Bulgarian), Doct. Thesis, Plovdiv, 2003.
- Totkov G., Robust Analysis of Bulgarian Texts and Development of Linguistic Processor, Math. And Math. Education, 19<sup>th</sup> Conf. of Union of Bulgarian Mathematicians, Sunny Beach, April 6-9, 1990, 295-302.
- Tufiş D., Cristea D., (1985a). IURES: A Human Engineering Approach to Natural Language Question Answering. In W. Bibel, B.Petkoff (eds.) *Artificial Intelligence: Systems, Applications, Methodology*, North Holland.
- Tufiş D., Cristea D.(1985b). A Pragmatic Implementation of the Dialogue Principles. In *Automatische Sprache Verarbeitung*, J.Kunze (ed.), Zentral Institut fur Linguistics Berlin, Germany.
- Tufiş D. & Cristea D. (2002). Methodological in Building the Romanian Wordnet and Consistency Checks in BalkaNet. In *Proceedings of the LREC Special Workshop on wordnets*, pp. 35-41, Las Palmas, Spain.
- Tufiş, D., Cristea, D., Stamou, S. (2004): BalkaNet: Aims, Methods, Results and Perspectives, a General Overview. In Romanian Journal of Information Science and Technology, Vol. 7(1-2) 1-35 (2004)

Tufiş, D., Ion, R., Ide, D. (2004): Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets. In proceedings of the 20th International Conference on Computational Linguistics, COLING, Geneva, 2004 pp. 1312-1318.

Mihalcea, R., Chklovski, T., Tufiş, D., Năstase, V., Tătar, D., Hristea, F. An Evaluation Exercise for Romanian Word Sense Disambiguation, In the Proceedings of the SENSEVAL-3 Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, ACL 2004 workshop, Barcelona, Spain, July 25-26, 2004.

Tufiş, D., Ion, R. Interlingual wordnets validation and word-sense disambiguation. In Proceedings of the Natural Language Understanding and Cognitive Systems Symposium, Porto, 2004, pp. 97-105

VisDic <http://nlp.fi.muni.cz/projekty/visdic/>

Vossen, P. EuroWordNet: a multilingual database with lexical semantic networks. Kluwer Academic Publishers, 1998.

Vossen P. (1996). Right or Wrong: Combining Lexical resources in the EuroWordNet Project. In M. Gellerstam, J. Jarborg, S. Malmgren, K. Noren, L., Rogstrom, C.R. Pappmehl (Eds.) *Proceedings of the Euralex Workshop*, pp. 715-128, Göteborg, Sweden.

Vossen P., Bloksma L., Rodriguez H., Climent S., Calzolari N., Roventini A., Bertagna F., Alonge A., Peters W. (1997a). The EuroWordNet Base Concepts and Top Ontology, LE-4003, Deliverable D017, D034, D036, University of Amsterdam.

Vossen P., Diez-Orzas P., Peters W. (1997b) The Multilingual Design of EuroWordNet. In P. Vossen, Calzolari N., Adriaens G., Sanfilippo A., Wilks Y. (Eds.) *Proceedings of the ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid, Spain.

Vossen P. (Ed.) (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic publishers, Dordrecht.

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.