ASSESSMENT AND EVALUATION OF THE PROJECT'S RESULTS

Deliverable D.9.1, WP9, BalkaNet, IST-2000-29388

Databases Laboratory (DBLAB)

Computer Engineering & Informatics Department

Patras University, Greece

GR 26500

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

BalkaNet

Identification number	IST-2000-29388
Type	Report – Document
Title	Assessment of the project's results
Status	Final
Deliverable	D.9.1
WP contributing to the deliverable	WP9
Task	T.9.1
Period Covered	September 2001-Septeber 2004
Date	September 2004
Version	0.5
Status	Public
Number of pages	35
WP / Task Responsible	DBLAB
Other Contributors	RACTI, UAIC, RACAI, SABANCI, FI MU, MEMO, MATF, DCMB, PU, UOA
Authors	{Sofia Stamou, Vassilis Andrikopoulos, Sofia Raikou} **DBLAB** {Cvetana Krstev, Gordana Pavlović-Lazetić, Ivan Obradović, Duško Vitas} **MATF** {Özlem Çetinoglu, Orhan Bilgin} **SABANCI** {Dan Tufis, Radu Ion, Verginica Mititelu, Eduard Barbu} **RACAI** {Dan Cristea, Catalin Mihaila, Corina Forascu } **UAIC** {Karel Pala, Pavel Smrz, Anna Sinopalnikova} **FI MU** {Dominique Dutoit, Patric de Torcy} **MEMO** {Svetla Koeva} **DCMB** {George Totkov} **PU**

EC Project Officer	Erwin Valentini
Project Coordinator	Professor Dimitris Christodoulakis Director of DBLAB Databases Laboratory, Computer Engineering & Informatics Department Patras University GR 26500, Greece Phone: +30 (61) 960 385 Fax: +30 (61) 960 438 E-mail: dxri@cti.gr
Keywords	Assessment of results, quantitative data, quality control policies, evaluation outcome, final status, evaluation metrics, wordnet applications
Actual Distribution	Public

Abstract	In this report the final assessment of BalkaNet project is presented with emphasis on qualitative issues and the main achievements delivered by the project. In particular, the objectives set at the initial and intermediate phases of the project are examined in comparison to the final results delivered by the reporting period. BalkaNet aimed at developing a multilingual semantic network of 8,000 comparable synsets across the Balkan languages. Additionally, it envisaged the implementation of open platform and sophisticated technical infrastructure that would enable editing, browsing and evaluating semantic networks irrespectively of languages or lexical resources. During the implementation of the project the initial objectives as well as the market advances and needs were always considered in order to deliver up to date, useful resources for the lesser studied Balkan languages. Following three years of extensive and meticulous work, Balkanet has now delivered a huge network of lexical databases for all Balkan languages, along with tools, editors, and a Wordnet Management System, meeting fully the objectives set. More specifically, with respect to the lexical data, BalkaNet consortium delivered more than double of the synsets, initially foreseen. Synsets are comparable across languages and enable the navigation from one language to the other via BalkaNet's ILI. Concerning the latter several upgrades have taken place in order for the ILI records to represent the latest advances of the English WordNet. Besides the lexical data delivered, the project has also resulted in a variety of technical components and tools that have helped the development, the evaluation and the interconnections of the monolingual Balkan wordnets. These tools have been built in such as way that they are platform and application independent and can serve several linguistic tasks irrespectively of the languages and/or resources used. Finally, another important outcome of the project concerns the incorporation of the Balkan wordnets within a Web search engine that indexes Web pages in all Balkan languages and English. The search engine has been implemented from scratch with the framework of the project and it is being continuously improved and updated. It started of by indexing Web documents from the BalkanTimes news Web archive for all languages in questions and developed into a Web search engine of the reachable Web. Several linguistic components have been incorporated into the engine, like spellers, normalizers, taggers and of course the Balkan wordnets. These components offer a variety of services to the end users, ranging from query expansion, domain search, classification, thematic directories and so forth.

	By the reporting period all project's objectives have been successfully met and the project has delivered both quantitative and qualitative results. The most important outcome of the project besides the actual deliveries is considered to be the close and mutual collaboration that has been established across the Balkan countries perhaps for the very first time. BalkaNet consortium managed to work smoothly and in an harmonized way and an excellent networking community of research and academic players has been established. It is our hope that this collaboration is further continued and that BalkaNet's initiative will find Europe's support in the future.
Status of abstract	Complete
Send on	September 2004

TABLE OF CONTENTS

INTRODUCTION

The final assessment of BalkaNet project is presented with emphasis on qualitative issues and the main achievements delivered by the project. In particular, the objectives set at the initial and intermediate phases of the project are examined in comparison to the final results delivered by the reporting period.

BalkaNet aimed at developing a multilingual semantic network of 8,000 comparable synsets across the Balkan languages. Additionally, it envisaged the implementation of open platform and sophisticated technical infrastructure that would enable editing, browsing and evaluating semantic networks irrespectively of languages or lexical resources.

During the implementation of the project the initial objectives as well as the market advances and needs were always considered in order to deliver up to date, useful resources for the lesser studied Balkan languages.

Following three years of extensive and meticulous work, Balkanet has now delivered a huge network of lexical databases for all Balkan languages, along with tools, editors, and a Wordnet Management System, meeting fully the objectives set.

More specifically, with respect to the lexical data, BalkaNet consortium delivered more than double of the synsets, initially foreseen. Synsets are comparable across languages and enable the navigation from one language to the other via BalkaNet's ILI. Concerning the latter several upgrades have taken place in order for the ILI records to represent the latest advances of the English WordNet.

Besides compatibility with the English WordNet, the BalkaNet consortium delivered as set of synsets that are specific to the Balkan languages in question, i.e., for which there are no lexicalized English counterparts. These synsets are the so-called Balkan-specific synsets and represent concepts that are common across the Balkan area, such as family relations, foods, garments, religious concepts and so forth. These concepts besides being informative on the common lexical elements across the languages in question are also a strong indicator of the cultural commonalities in the Balkan area and the influences each country has taken from other neighbouring regions.

Additionally, a set of domain-specific synsets pertaining to the thematic categories, of Law, Politics and Economy have been delivered for all Balkan languages. These specialized synsets were generally missing for those languages and they are really useful in a variety of NLP applications. To enable the development of domain-specific synsets an ontology on top of the BalkaNet Base Concepts has been built mainly originating from a combination of the SUMO ontology and the WordNet Domains 1.0.

Besides the lexical data delivered, the project has also resulted in a variety of technical components and tools that have helped the development, the evaluation and the interconnections of the monolingual Balkan wordnets. These tools have been built in such as way that they are platform and application independent and can serve several linguistic tasks irrespectively of the languages and/or resources used.

Additionally, the 1984 corpus earlier available for some of the Balkan languages has been enriched with its Turkish and Greek versions and has been semantically annotated with the framework of BalkaNet. Corpus annotation was carried out as a part of the cross-lingual qualitative evaluation of the project's results. With respect to this task several tools have been developed to facilitate annotation, the most useful of which is a WSDTool developed by RACAI team which is being continuously improved ever since. Finally, another important outcome of the project concerns the incorporation of the Balkan wordnets within a Web search engine that indexes Web pages in all Balkan languages and English. The search engine has been implemented from scratch with the framework of the project and it is being continuously improved and updated. It started of by indexing Web documents from the BalkanTimes news Web archive for all languages in questions and developed into a Web

search engine of the reachable Web. Several linguistic components have been incorporated into the engine, like spellers, normalizers, taggers and of course the Balkan wordnets. These components offer a variety of services to the end users, ranging from query expansion, domain search, classification, thematic directories and so forth.

The initial goal for launching the search engine was to organize the documents at the index level on the basis of their thematic subjects, a task frequently called within the consortium as conceptual classification of the Web pages. However, several additional features have been incorporated in the engine like the ones mentioned above, resulting this way into a search engine specifically tailored for the Balkan languages which can however easily be modified and extensile to other user groups and languages.

By the reporting period all project's objectives have been successfully met and the project has delivered both quantitative and qualitative results. The most important outcome of the project besides the actual deliveries is considered to be the close and mutual collaboration that has been established across the Balkan countries perhaps for the very first time. BalkaNet consortium managed to work smoothly and in a harmonized way and an excellent networking community of research and academic players has been established. It is our hope that this collaboration is further continued and that BalkaNet's initiative will find Europe's support in the future.

FINAL ASSESSMENT OF THE SEMANTIC NETWORKS

BalkaNet's objectives

Following the principles adopted in EWN and the English WordNet, 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). 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 conceptual 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.

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 Interlingua 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 BalkanTimes website which will are 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 1.6 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 1.6 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.

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 is 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 by (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, they were selected from all the English nouns and verbs occurring in the corpus, only those that belong 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 211 had at least two senses. These words occurred altogether 1810 times not always translated in every language present in the parallel corpus. All the partners received the list of the 211 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.

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 WSDtool methodology has been described at length elsewhere (i.e. see the recently published article "Dan Tufis, Radu Ion, Nancy 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):

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

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 wordnetl;
- **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 poing remark is the same as above.

Considering one language, if we subtract the last three rows of the table from the total number of annotated occurrences, and consider the sum of the first two rows (WSDOK1 and 2) as providing useful information for the word sense disambiguation process (being this way a piece of information that indicates how good is the individual source wordnet with respect to the list of the target words that was used), then the percentage of the occurrences of the target words that were successfully disambiguated due to the individual wordnet can be found in the corresponding column of the following table:

	GR	***BG***	***SR***	***RO***
TOTAL Occs.	824	937	1171	1159
WSD Accuracy	657 (**79.73%**)	810 (**86.44%**)	1090 (**93.08%**)	1155 (**99.65%**)

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)

Quantitative evaluation across Balkan wordnets – Overall Statistics

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	19680
Nouns synsets	13345
Verb synsets	4649
Adjective synsets	852
Adverb synsets	834
Token literals	33514
Type literals	19473
The medium length of synsets	1.70
The average number of senses per literal	1.72
Valency frames	151
Balkanet Common Set	
BCS1	1218
BCS2	3471
BCS3	3827

Lexico-semantic relations	25683
Hypernym	17982
holo_portion	112
holo_part	1174
also_see	416
similar_to	899
verb_group	1022
near_antonym	1658
holo_member	797
causes	126
be_in_state	558
subevent	189
category_domain	750

Domain specific synsets	286
Law	98
Politics	89
Economy	99
Balkan specific synsets	?
Romanian specific synsets	545

Status of the Serbian Wordnet

Synsets	7780
Nouns	5646
Verbs	1803
Adjectives	324
Adverbs	13
Literals	12736
Literals/ synset	~ 1.64
Base Concepts	
BC1	1219
BC2	3469
BC3	1352

Domain specific synsets	305
Law	103
Politics	101
Economy	101
Balkan specific synsets	117
Serbian specific synsets	198

Lexico-semantic relations	9886
HYPERNYM	7326
HOLO_MEMBER	735
HOLO_PART	423
HOLO_PORTION	37

VERB_GROUP	154
BE_IN_STATE	176
SUBEVENT	68
CAUSES	54
ALSO_SEE	116
SIMILAR_TO	16
DERIVED	177
DERIVED-POS	42
DERIVED-GENDER	20
NEAR_ANTONYM	533
PARTICLE	9

Usage (examples)	718 (in 630 synsets)
Morphosyntactic information	7699 literals

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_portion	357
holo_part	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

	Nouns	Verbs	Adjectives	Adverbs	Total
Synsets	14178	4169	3088	9	21444
Literals	24823	15010	5099	12	44944
Literals/synsets	1.75	3.6	1.65	1.33	2.1
Graphic words	20 219	8934	3844	12	33009
Literals/words	1.23	1.68	1.33	1	1.36
ILR	22960	10090	4851	10	37911
ILR per synset	1.61	2.42	1.57	1.1	1.77
Definitions	14178	4169	3088	9	21444
Frames	1165				
Frames/Verbs	~0,279				

Table 1. Statistical data characterizing BulNet

WN	N nodes	Tops N	V nodes	Tops V
Eng2.0	79 689	9	13 508	554
BulNet	14 178	9	4 169	397

Table 2. Number of tops per Bulgarian nouns and verbs

ILR	POS/POS	EW2.0	BulNet
ALSO SEE	A/A V/V	3 240	1 186

BE IN STATE	A/N	1 296	622
BG DERIVATIVE	N/V	36 630	7 920
CATEGORY DOMAIN	N/N V/N A/N B/N	6 166	1 333
CATEGORY MEMBER	N/N V/N A/N B/N	6 166	1 333
CAUSES	V/V	439	108
DERIVED	A/N	6 809	1 165
HOLO MEMBER	N/N	12 205	921
HOLO PART	N/N	8 636	1 386
HOLO PORTION	N/N	787	114
HYPERONYM	N/N V/V	94 844	18 373
HYPONYM	N/N V/V	94 844	18 373
IS CAUSED BY	V/V	439	108
IS DERIVED FROM	N/A	6 809	1 165
IS STATE OF	N/A	1 296	622
IS SUBEVENT OF	V/V	409	182
MERO MEMBER	N/N	12 205	921
MERO PART	N/N	8 636	1 386
MERO PORTION	N/N	787	114
NEAR ANTONYM	N/N A/A V/V	7 642	2 010
PARTICIPLE	A/V	401	56
REGION DOMAIN	N/N V/N A/N B/N	1 280	32
REGION MEMBER	N/N V/N A/N B/N	1 280	32
SIMILAR TO	A/A V/V	22 196	1 592
SUBEVENT	V/V	409	182
VERB GROUP	V/V	1 748	882
USAGE DOMAIN	N/N V/N A/N B/N	983	29
USAGE MEMBER	N/N V/N A/N B/N	983	29
ID		115 424	21 444
L NOTE		0	1 520
LITERAL		203 147	44 944
POS		115 424	21 444
SENSE		203 147	44 944
S NOTE		0	125
USAGE		48 231	8 816
BC1			1218
BC2			3471
BC3			3827

Table 3. Distribution of the BulNet relations

		Nouns	Verbs	Adjectives	Adverbs	Total
Domain Law	Synsets	859	120	26	2	1007
	Literals	1292	358	45	3	1698
Domain Politics	Synsets	308	47	8	2	365
	Literals	522	94	13	2	631
Domain Economy	Synsets	544	146	1	2	693
	Literals	868	365	1	3	1237
Bulgarian specific	Synsets	254	6	80	0	340
	Literals	285	15	91	0	391

Table 4. Domain specific and Bulgarian specific synsets

BALKANET'S CONTRIBUTION TO EC SOCIAL OBJECTIVES AND TECHNOLOGICAL PROSPECTS

BalkaNet delivered a multilingual lexical database for six Balkan languages, forming this way a linguistic resource that meets the aggregated needs of Europe's academic and industrial researchers in the field of language engineering. Its contribution to the community and social objectives focuses on the strengthening of ties with the academic and information technology in European and Balkan countries. The collaboration of all consortium members was deemed as highly beneficial in terms of both experience and knowledge exchanged as well as in terms of bringing together scientists across the Balkans towards pursuing a common goal.

The multilingual BalkaNet semantic network besides being a valuable lexical semantic resource, it captures valuable information about the expressiveness of the Balkan languages and provides at the same time the precise mapping of these languages' vocabulary, expanding this way Europe's linguistic horizons. In addition, BalkaNet is the first unified resources for all Balkan languages, enabling as such retrieval of information about lexicalized patterns across the languages, conceptual density of areas or the vocabulary and distribution of semantic distinctions or relations over different areas of vocabulary.

BalkaNet's impact so far has motivated Language Technology companies and institutions to study these lesser supported languages and on benefiting form the project's results through their incorporation into many NLP products and applications, ranging from word sense disambiguation software, to lexicographic data extraction tools, etc. A significant contribution of BalkaNet is the open source nature of all technical infrastructures that has delivered so far. Both VisDic (the main wordnet editor) and WMS (the core wordnet browser/viewer) have long ago been distributed free of charge via the Internet to anyone interested and the consortium provides continuous support to all their users. It is believed by all consortium members that BalkaNet offered the incentives to the European industry to start integrating Balkan lexical resources and terminology into their NLP application systems. Moreover, the project's contribution to the community's social objectives has paved the way for suture collaborations across members of the Eastern European area and the EC community. BalkaNet addresses the multilingualism and cultural diversity in the Balkan region and proves that above all differentiations there exist many deep commonalities between those countries that should be further explored and supported. Business competitiveness is expected to be increased and definitely the BalkaNet consortium has been requested to participate into much newly formed collaboration of other institutions. We believe that BalkaNet has inspired the initiation of several other similar projects that aim at developing wordnets for other Eastern European languages, such as Russian, Hungarian, Croatian, Latvian, etc.

Recently there has been a communication between the BalkaNet consortium and another group of people interested in launching the Central European Wordnet project. This group asked for BalkaNet consortium's support and advice into their initiative, a request that will be fully met in as much as possible. Europe's and the world's interest in BalkaNet increases constantly as the project develops and improves. Many conferences have devoted special sessions on lesser studies languages and many specialized workshops have been organized. Participants have expressed their profound interest in BalkaNet's activities and their wish that this collaboration is further continued beyond the framework of the project.

Europe's linguistic community is richer now that has a large resource of six more lexical databases and a set of tools and software for the processing of these resources. BalkaNet consortium will continue this collaboration in the future in an attempt to keep BalkaNet's objectives and goals alive. Currently alternative means of financial support are being sought that would support BalkaNet's activities and objectives.

On a more practical and technological ground, BalkaNet's impact has also been significant. The tools and applications delivered within the framework of the project have helped the consortium as well as the entire language technology community to facilitate and advance the

processing of the lexical data. Moreover, a multilingual corpus, namely 1984, has been enriched with two more languages (Greek and Turkish) and several explanatory dictionaries earlier available in printed format only have now been digitized and can be processed electronically. Besides the annotated 1984 multilingual corpus that has been delivered by the project, another multilingual corpus was compiled within the framework of the project. This is a corpus of online journalistic texts extracted from the BalkanTimes web archive. Documents classified into thematic areas for all languages involved in the project have been downloaded and morphosyntactically processed, which corresponds to HTML parsing, tokenization, lemmatization and POS tagging.

The VisDic editor is already being used by other groups that are currently developing their Wordnets and has been valuable in their tasks. Moreover, the Wordnet Management System (WMS) is freely available and accessible online or it can be installed locally. Via WMS all Balkan wordnets and the ILI records can be accessed and a variety of information retrieval services is being offered. These range from retrieving synset(s) for one or more languages, to retrieving the entire hierarchy of a concept.

This technical infrastructure delivered by the project has set the ground for market opportunities in the Balkan area as well as across Europe. The lexical resources developed are useful for language engineers and translators who so far due to the lack of linguistic data in electronic form were falling behind Europe's advances. Especially, when it come to Europe's enlargement, the impact and contribution of BalkaNet is significant. Natural language is the vehicle via which people communicate and express their thoughts. Via BalkaNet; language barriers are dropped to a large extend for those countries facilitating this way communication, collaboration and peaceful living. The EC will benefit at large from the availability of BalkaNet resource and it is our aspiration that BalkaNet is perceived as something larger and much more significant than a linguistic resources. Balkanet is a bright beacon of collaboration among 13 different groups across the Balkan region. BalkaNet's success lies dominantly to this collaboration.

Motivation and technological impact of the Wordnet Management System

- *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 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 retrieval 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.

Motivation and technological impact of the VisDic Wordnet Editor

Introduction

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 faciliate 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 hyperonyms). 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
      conjuntive CDATA #IMPLIED
      disjuntive CDATA #IMPLIED
      reversed CDATA #IMPLIED
      negative CDATA #IMPLIED
      factive CDATA #IMPLIED
      non_factive CDATA #IMPLIED>

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

Figure 2. 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 3. 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, hyperonym, 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 orginal 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 5. Comparison of Polaris and XML record

Conclusions

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 wornNet. 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).

References

1. Artale A., Magnini B., Strapparava C. WordNet for Italian and Its use for Lexical Discrimination. In Proceeedings of the 5th Congress of the Italian Association for Artificial Intelligence, Springer Verlag.
2. Miller, G.A. WordNet: An Online Lexical Database. International Journal of Lexicography 3(4) (special issue), 1990.
3. 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).
4. 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.
5. Vossen, P. EuroWordNet: a multilingual database with lexical semantic networks. Kluver Academic Publishers, 1998.

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.

- Organization of a specialized BalkaNet session in conjunction to the 2nd International Global Wordnet Conference (GWC) that was help in Brno, Czech Republic (January, 2004). 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 3rd 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, 27th 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,* 6th Intex Workshop 2003, GWC International Conference 2004, LREC 2004, IEEE International Conference on Advanced Learning Technologies (ICALT 2004), 7th Intex Worksop 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 Automatisation 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 Scool 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. 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.

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.

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 indexed 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 area 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 handing single term queries since after all those are the most frequent types of queries issued in IR systems especially by inexperienced end users.

BalkaNet 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.