BalkaNet, IST-2000-29388 **2nd Annual Report 2003**

www.dblab.upatras.gr/balkanet.htm

Project's General Description

Balkanet aims at combining effectively Balkan lexicography and modern computation. The most ambitious feature of Balkanet is its attempt to represent semantic relations and organize lexical information from Balkan languages in terms of word meanings. Its main objective is the development of six monolingual WordNets and their combination in a common infrastructure, namely the WordNet Management System (WMS). Moreover, Balkanet aims not only at combining Balkan word forms in an on line dictionary but at a further expansion of EuroWordNet by tracing and exploring the relationships among Romance and Balkan languages. Finally, it aims at promoting the study of the less studied Balkan languages by creating a large-scale linguistic resource, which would be useful in various NLP applications, ranging from information retrieval to machine translation and language teaching.

Summary of 2002-2003 Activities

Within the second year of the project many tasks have been performed, which concern both the implementation of the project in a linguistic and technical level, and the application of the project's outcome in an IR system. Specifically, concerning the development of the monolingual Wordnets the following have been achieved: (i) the full set of BCs subset I and II have been developed and are cross-linked across languages. To enable the development of the monolingual Wordnets several tasks were carried out which fall within the ILI's contents modifications as well as the processing of the lexical resources available to the consortium. At the reporting period subset III of BCs is under development and is expected to be delivered by the end of January 2004.

Besides the actual development of the monolingual networks some steps were commonly adopted so as to reassure their qualitative content. Following several discussions among the consortium members and taking into account the feedback of the PO as well as the project's midterm reviewers, there has been established a firm and well-documented validation plan, which aims at monitoring the quality of the individual Wordnets in terms of completeness and accurate representation of the languages' characteristics. The quality control task has been followed for the last couple of months and as such qualitative results cannot be reported at this point.

Besides the linguistic components of the project several tasks have been performed concenring the technical infrastructure that stores the lexical data as well as the one extensively used for editing the lexical elements and their semantic relations.

In specific these are: a prototype version of the Wordnet Management System (WMS) in which all monolingual Wordnets are stored and cross-linked has been developed and released to project members so as to perform some tests and provide members of the DBLAB team with useful feedback concerning areas that need further improvement or modifications. In addition, to the prototype release, new services are being incorporated within the framework of the WMS, that mainly concern services providing information of statistical nature such as the distance of a synset to the top of a relation. Also, the necessary services for the addition of a predefined domain ontology to the wordnets and the retrieval of this information were designed and are going to be implemented.

For the last year of the project the WMS is expected to be further improved by the incorporation of cross-lingual services and further disseminated beyond the project's consortium.

As a final achievement of the project's second year is Balkanet's pilot incorporation in an IR system (search engine) currently under development. To this respect the core infrastructure of the web search engine is going to be developed, which will support the storage of conceptual information for the indexed documents of the engine, the retrieval of documents based on conceptual information and the appropriate mechanisms for these tasks.

With respect to the Balkanet's incorporation in the engine's mechanisms the following tasks are currently in progress: the selection and implementation of an appropriate algorithm for the indexing and retrieval of the documents, using the conceptual information provided by the wordnets, and the intra-communication between the engine and the WMS. Both are expected to be delivered by mid January 2004.

Following the abovementioned achievements, it is foreseen that within the last year of the project the following tasks will be performed.

- The extended WMS will be made publicly available
- Data and lexical links validation tests will be finalized so as to reassure the quality of the monolingual Wordnets

- All Balkan wordnets will have been fully developed and cross linked into the WMS infrastructure
- The web search engine that will host Balkanet will be fully functional
- The conceptual indexing mechanisms incorporated in the search engine will be implemented and tested by the project's end user, OTEnet.

Important work area

Quality Control and Approaches towards Balkanet's final application

The main objective of the quality control plan was to reassure the delivery of qualitative monolingual Wordnets in order for the latter to be incorporated in a web search engine and perform content-based indexing as opposed to keyword-based indexing so far adopted by traditional IR systems. With respect to the quality control plan the following tasks have been accomplished.

Data Quality Control

For reassuring the delivery of qualitative data within and across Balkan Wordnets, a quality control plan has been established and followed by all consortium members. The main objectives and the methodology of carrying our quality control tasks have been summarized in a detailed report, which has been uploaded on the project's information server and has been made available to the EC and the project's reviewers. The main actions carried out while checking the quality of the monolingual Wordnets fall within two main areas: (1) quality control of the monolingual synsets in terms of coverage and representativess of the involved languages, and (ii) validation of the lexico-semantic relations' quality in terms of completeness, correctness and accuracy in reflecting lexicalized concepts in the respective languages. At the time being the validation task is in progress and is expected to last till the end of the project. The ultimate goal of the consortium is within the time limits of the project to be able to check and where necessary correct as many synsets as possible.

Special Wordnet Features

A major feature of the Balkanet semantic network concerns the incorporation of "conceptual domains" as the top most entries of the ILI nodes. 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. Each element of a conceptual domain is built into a taxonomic structure and each taxonomy links concepts that belong to that particular domain. Besides organizing concepts, a key attribute of Balkanet's ILI is its flexibility in incorporating new concepts and/or languages by allowing the percolation of shared semantic attributes to all concepts represented within a taxonomy.

Moreover, the possibility of incorporating language-specific lexical relations is under investigation and will take place where necessary.

WordNet Management System architecture (WMS)

WMS provides the following services: search for synset data by literal name and/or synset id (ILI number), retrieval of hierarchic semantic data for a given synset concerning its relations, retrieval of the conceptual domain of a given synset. Additionally, a number of services of statistical nature are being implemented concerning the statistics of a wordnet, the position of a synset in the tree formed by one or more type of relations and the distance of two or more synset on this tree and/or across wordnets. WMS will also provide services for the addition of the domain ontology into a monolingual wordnet and the retrieval of information concerning the domain that a given synset belongs to.

WMS is a system that functions as an interconnection and communication link between a user (where user is any data consuming entity) and any of the involved monoligual systems. It also facilitates the need for flexible mechanisms and services for navigation in a multilingual linguistics environment, that are provided by a main technical infrastructure of the multilingual network. However, keeping all the benefits of the Web, such as distributed work environment, concurrent access to the data and multiple views of the data are achieved through the WMS. The benefit behind using the WMS is that project developers will be tightly linked with other wordnets and valuable suggestions for new terminology fields will be facilitated.

A variety of methods for the communication of WMS with the search engine were tested and are being implemented and tested for suitability and effectiveness. Since the actual interconnection of the search engine with the network of servers that consist the WMS system would prove highly ineffective in terms of availability and performance – since it would depend on the status of the network, the option of a stand-alone WMS Server hosting the latest versions of the monolingual wordnets was selected.

User Group, Promotion and Awareness

The major project's objective is strengthening the ties between the academic and information technology communities in European countries. To this respect 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 other NLP applications. Several of them have been admitted access to the project's intermediate results on the grounds that they will be 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, and keeping in mind the final incorporation of the project's results in an Information Retrieval system, the consortium has contacted several 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 are

currently performed by the consortium members 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.

With respect to the dissemination of the scientific project's results these have been so far presented in various national and international conferences and colloquia and aim at stimulating the community's interest towards the new approaches addressed in the Balkanet project. In light of the above members of the consortium have demonstrated a rather active participation in the second Global Wordnet Conference (GWC) organized by Global Wordnet Association (GWA) on January 2004 in Czech republic.

Besides participating in a Conference mainly devoted to semantic networks issues, several other publications to NLP conferences have taken place. For a detailed account of the scientific contribution the project has made so far please refer to the last table of the present report. In addition, the Balkanet consortium has approached the steering board of the Global Wordnet Association expressing their interest in becoming actively involved to the Association's activities. By joining GWA each contractor will hold responsibility 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.

A final step towards dissemination of the so far project's results concerns the compilation of a Balkanet exploitation plan that would go beyond the framework of the project. Such a plan is expected to form the backbone of a potential follow up proposal and in this direction several European NLP companies have been contacted. At the present time useful feedback is being collected on their behalf so as to come up with a firm and coherent proposal as to which the project's future directions would be.

Future Work / Exploitation Prospects

Within the remaining time of the project the following tasks will be completed. All monolingual Balkan Wordnets will have been finalized and cross-checked in terms of semantic overlap and qualitative data. Furthermore, the Wordnet Management System will be fully functional and publicly available to any interested party. The WMS will offer not only lexical resources storage capabilities but also a variety of services that would help end users exploiting the lexical information hosted within WMS.

Following on from this the search engine that will incorporate conceptual information encoded within the Balkanet multilingual resource will be developed. The search engine will offer a variety of mechanisms emerging from the project's results which are expected to be rather useful for the IR community. The most innovative mechanisms developed concern the implementation of a conceptual indexing infrastructure. The latter will fully exploit Balkanet's structure so as to trace conceptual information of the indexed documents and enable content-based indexing. Another feature of the search engine that will delivered at the end of the project will be the availability of query expansion modules. Query expansion approach is targeted towards assisting end users in formulating their search requests in such a way and by using terms that are aslo used by the search engine while indexing documents. The query expansion module will form an individual component that could be employed by other IR systems. The main contribution of Balkanet's semantic network towards

query expansion will focus on the usage of the lexical internal relations used to cross-link the concepts represented within each monolingual network.

Besides the project's final application towards IR another major outcome of the project concerns the availability of a large-scale parallel Balkan corpus (i.e., 1984 corpus), which will be in part semantically annotated with information encoded within the monolingual Balkan synsets. Corpus annotation is a task that will take place within the framework of the quality control activities but reported results on the annotation process will be made available shortly before the end of the project. This is essentially due to the nature of the semantic annotation task which requires a lot of manual work. However, to speed up the process several scenarios are investigated so as to develop the mechanisms and the technical infrastructure that would support annotators while tagging corpus terms.

Finally, another expected benefit of the project that will come within its final year concerns the availability of a detailed evaluation report in which both quantitative (statistical) and qualitative data will be addressed. Such a detailed report would be a useful guide for researchers and industrial parties who are either interested in developing their own semantic networks or who aim at exploiting and incorporating semantic networks within their applications and NLP components.

As a general outcome of the project which however should not be underestimated is the fact that all Balkan countries will have available perhaps for the first time a unified, common and richly encoded lexical resource that will open up wider possibilities for exploring the involved languages' patterns and come up with more competent and large scare real life applications.

Further Information

All documents, reports and public data can be downloaded from the BalkaNet information server: http://is.dblab.upatras.gr and the BalkaNet web site: http://www.dblab.upatras.gr/balkanet.htm

Deliverables:

D.0	"Quarterly and Semestrial Management Reports – Cost Statements"	DBLAB Project Coordinator in cooperation with the consortium
D.3.2	"The Wordnet Management System along with its peripheral tools"	UOA
D.4.1	"The local Base Concepts"	UAIC
D.4.2	"The common Base Concepts"	UAIC
D.5.1	"The language internal relations for nouns and verbs"	SABANCI
D.5.2	"First version of the individual Wordnets"	SABANCI

Published Balkanet scientific results

"Automated Improving and Forming Synsets on Conventional (non computer based) Synonym Dictionaries", IT Conference Sofia, November 2002, G. Totkov, P. Ivanova
"Bulgarian WordNet – Problem and Prospects", International Conference Electronic

Description and Edition of Slavic Sources, September 2002, Svetla Koeva
"Logic for WordNet", Annual journal of Sofia University, 2002, Tinko Tinchev, Stoyan Mihov, Svetla Koeva, Angel Genov
"Turkish WordNet" – CeBIT Eurasia Information Technologies Fair, 3-8 September 2002, Kemal Oflazer, Ozlem Cetinoglu
"Building a Romanian WordNet; problems, solutions and prospects, Seminar Sprachwissenschaft Abt. Computer linguistik, Universitatea Eberhard Karls, Tubingen, November 2002, Dan Tufis
"Word Sense Clustering Based on Translation Equivalence in Parallel Texts: A Case Study in Romanian" Proceedings of the International Conference on Speech and Dialogue, Bucharest 8-10 April 2003, Dan Tufis, Radu Ion
"TREQ-AL: A word alignment system with limited language resources" in Proceedings of the HLT-NAACL 2003 Workshop: Building and Using Parallel Texts; Data Driven Machine Translation and Beyond, Edmonton, May 31, 2003, pp. 36-39 Dan Tufis, Ana-Maria Barbu, Radu Ion
"Extracting multilingual lexicons from parallel corpora", *Computers and the Humanities*, vol. 37, 2003, 33 pages Dan Tufis, Ana-Maria Barbu, Radu Ion
"Automated Improving and Forming WordNet Synsets on Conventional (non computer based) Synonym and Bilingual Dictionaries, in A. Narin'iyani (ed.), Computational Linguistics and its Applications", Proceedings ?f the International Workshop DIALOGUE'2003, Protvino, June 11-16 2003, Totkov G., P. Ivanova, Iv. Riskov,
"On Bulgarian Text-to-Speech System", Proceeding of the International Conference ICT&P'2003, Varna, 23-26 June T G. Totkov, V. Angelova
"Towards Building Bulgarian WordNet: Language Resources and Tools", Proceeding of the International Conference ICT&P'2003, Varna, 23-26 June G. Totkov
"Application of Intex in Refinement and Validation of Serbian Wordnet", 6th Intex Workshop, 28-30th May 2003, Sofia. Obradovic, R. Stankovic, C. Krstev, G. Pavlovic-Lažetic, "Corpora Issues in Validation of Serbian Wordnet", Text, Speech, Dialogue, Proceedings of the 6th International Conference, September 2003, Ceské Budejovice, Czech Republic, V. Mataušek, P. Mautner (eds.), Springer 2003, pp. 132-137. C Krstev, G. Pavlovic-Lažetic, I. Obradovic, D. Vitas
"Regular Expressions Builder and Parser for Unicode Systems", 1st Balkan Conference in Informatics (BCI'2003), Thessalonica, Greece, November 21-23, 2003. G. Totkov, D. Blagoev, R. Dokov (to be presented)
"Bipartite Finite State Transducers as Morphology Analyser, Synthesizer, Lemmatizer and Unknown-Word Guesser", 1st Balkan Conference in Informatics (BCI'2003), Thessalonica, Greece, November 21-23, 2003, G. Totkov, R. Doneva (to be presented)
"Cross Lingual Validation of Multilingual WordNets", Global Wordnet Association Conference, Brno, Czech, January 2004. Dan Tufis, Radu Ion, Eduard Barbu, Verginica Barbu, (to be presented)
"Wordnet exploitation through a distributed network of servers" Global Wordnet Association Conference, Brno, Czech, January 2004. Christodoulakis D., Koutsoubos I-D, Andrikopoulos V.
"Morphosemantic Relations in and across Wordnets: A Study Based on Turkish", Global Wordnet Association Conference, Brno, Czech Republic, January 2004. Orhan Bilgin, Ozlem Cetinoglu, Kemal Oflazer (to be presented)