BalkaNet, IST-2000-29388 1st Annual Report 2002

www.dblab.upatras.gr/balkanet.htm

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 2001-2002 Activities

Within the first year of the project many tasks have been performed, most of which fall within the following two domains: a) processing of lexical resources and corpora for the implementation of the monolingual wordnets and b) development of the technical infrastructure via which wordnet editing as well as wordnet cross-linguistic navigation would be made efficient. More specifically, the following achievements have been reached: a survey has taken place regarding the requirements that should be followed for the implementation of the project. More specifically, the survey was targeted towards both user and developers' requirements. Following this, various individuals and interested parties were contacted in order for the consortium to acquire feedback on the application of the lexical resources and corpora, beyond the

framework of the project. With respect to the validation of the linguistic work, which has been performed within the first year of the project, several tools have been developed and services have been implemented so as to check vocabulary coverage and completeness of the monolingual wordnets. Moreover, in this respect a tool has been developed which performs a validation of synsets's quality[1]. In addition, a wordnet editor, namely VisDic, that forms that basic infrastructure on which monolingual wordnets' development is taking place, has been developed and fully is functional in both Windows and Linux platforms. Finally, a prototype of the WordNet Management System (WMS) that forms the main module via which communication of WordNets and browsing of the resources is being performed, has been developed and is currently under excessive testing and modifications. Regarding WMS various peripheral services have been implemented that allow viewing of and navigation in the linguistic data in rather efficient ways.

Summarizing, following the abovementioned achievements, it is foreseen that within the following year of the project many additional tasks will be performed, the most important of which are summarized below. The WMS is going to be fully functional and available to interested parties, monolingual wordnet development tasks will have been almost finalized and an evaluation of the resources will take place. An extensive market research will take place in order for the so far project's results to be incorporated in various applications. Finally, within the forthcoming year the infrastructure that will form BalkaNet's application, namely, indexing of web documents based on their semantics will have been finalized.

Important work area

User Requirements and Product Profile

Balkanet multilingual semantic network is targeted towards a wide spectrum of potential users and as such can be an extremely valuable resource in a range of applications such as information retrieval, language teaching, lexicographic work etc. within the limited time of the project the envisaged application of Balkanet resource is focused on a preliminary classification of web pages during indexing time. Based on the above, project requirements fall into two major categories: developers' requirements and users' requirements with sufficient overlap between the two. In particular, developers' requirements concern completeness for each monolingual wordnet of the language in question. Completeness is determined not only in terms of vocabulary coverage and overlap but also in terms of synsets' validation and quality reassurance. Moreover, from a technical point of view developers' requirements aim at the implementation of various tools and services, each of which is targeted towards multiple aspects of the project. Particularly, requirements of the technical infrastructure concern efficiency of tools developed for the processing of the lexical resources and corpora, capacious storage mechanisms, effective wordnet editor and finally a well-designed and flexible wordnet management system, which will form the main source for distributing the final project's results to a wider audience. All requirements listed above have been taken into serious consideration by all members of the consortium and form the basis on which Balkanet development is being conducted. On the other hand users' requirements as defined after taking into account

[1] The respective tool has been developed by the RACAI team

market needs fall in the following categories: a) quality of data and b) reusability and openness of the technical infrastructure. Concerning the first point, which is largely shared by the developers' group too, it is desirable that in each monolingual wordnet, lexicalised patterns of the underlying languages are being reflected in a meaningful and thorough way. Based on the assumption that the final multilingual resource will be adopted in a series of tasks we are paying particular attention so as to meet users' needs and hence construct a resource out of which data acquisition would not pose any burden on the final user. Similarly, the technical infrastructure that will be developed within the framework of the project is intended to be as open and easy-to-use, as possible. In this respect WMS is a platform independent wordnet browser and the main wordnet editor, namely VisDic, runs under both windows and Linux platforms.

Data Compilation

To ensure compatibility across monolingual wordnets we have started off by a common set of concepts, which emerged from Subsets 1 and 2 of the Eurowordnet project. These concepts are currently shared by all wordnets and form the core part of each monolingual wordnet comprising currently of approximately ~5,000 synsets. Following to this, we adopted the EWN Inter-Lingual-Index (ILI) via which cross-linking of monolingual wordnets would be made efficient. ILI mainly included Princeton WordNet 1.5 synsets but it has been further improved and updated with WN 1.6 synsets as well. Moreover, each team has processed various monolingual resources, such as explanatory dictionaries, corpora, glossaries etc., in order to extract terms that would form the basis for new synsets to be developed. Selection of terms is based on their importance where the latter is determined by their frequency scores or by the number of the relations they bear with other concepts or word meanings etc.

The abovementioned Base Concepts have been incorporated into the multilingual resource, via the VisDic editor, under the Top Ontology inherited by EWN. For the time being only 1st and 2nd level entities have been incorporated, which terms falling within these two have higher frequency values in the monolingual resources.

Moreover, terms described above have been linked to the ILI records based on their glosses so that navigation within wordnets is facilitated. However, there have been noted several cases where linking could not be performed due to lexicalised differences between languages. Remedial actions have been taken for such cases and the possibility to introduce new language internal relation in this respect is under consideration.

Technology and innovative features

This contribution describes a new tool named VisDic for browsing and editing WordNet databases. It was developed in the Natural Language Processing Laboratory at the Faculty of Informatics, Masaryk University. In fact, it is not designed as a specialized tool for processing WordNet data only, generally, it has been developed as a tool for viewing and editing any lexical database as e.g. multilingual dictionaries, monolingual dictionaries, corpora, etc. From this point of view, WordNet can be also understood as a dictionary with special features.

1. Functions of VisDic

VisDic application and its functions will be described in this part. The performance of the tool is intimately tied up with the data representation mentioned above. A good definition of the dictionary structure can lead to the fast data searching and modifying as the WordNet XML structure from the previous part shows.

VisDic is designed to hold up to ten dictionaries at the same time. The user can work with WordNet databases, multilingual dictionaries, monolingual dictionaries, **corpora** or other type of lexical databases, all in one window. One can edit or browse with even more copies of one dictionary at the same time, but only one of them can be changed, the other ones can be viewed only. This feature enables the user to look up for additional information in the same data source.

Another reason for working with dictionaries simultaneously is the possibility of their cooperation. According to the same value of the entry identification value (key attribute, which was discussed above), the respective entry can be shown in the other dictionary. Then, in the case of monolingual dictionary, say the English one, it is possible to translate the same entry into another monolingual dictionary, e.g. the Czech one, although none of these dictionaries are multilingual. For example, this feature is very suitable for the WordNet multilingual view.

1.1 Main Application Window

In the application, every dictionary has its own window frame. The frames are arranged from left to right. Their positions displaying the active dictionaries can be changed anytime. A window frame contains an edit box for dictionary querying where the user can specify what exactly he wants to find. The corresponding entries will be listed in a list box below. After selecting the desired entry it will be shown in the last part of the frame with regard to the type of view discussed further. For illustration see Figure 3.

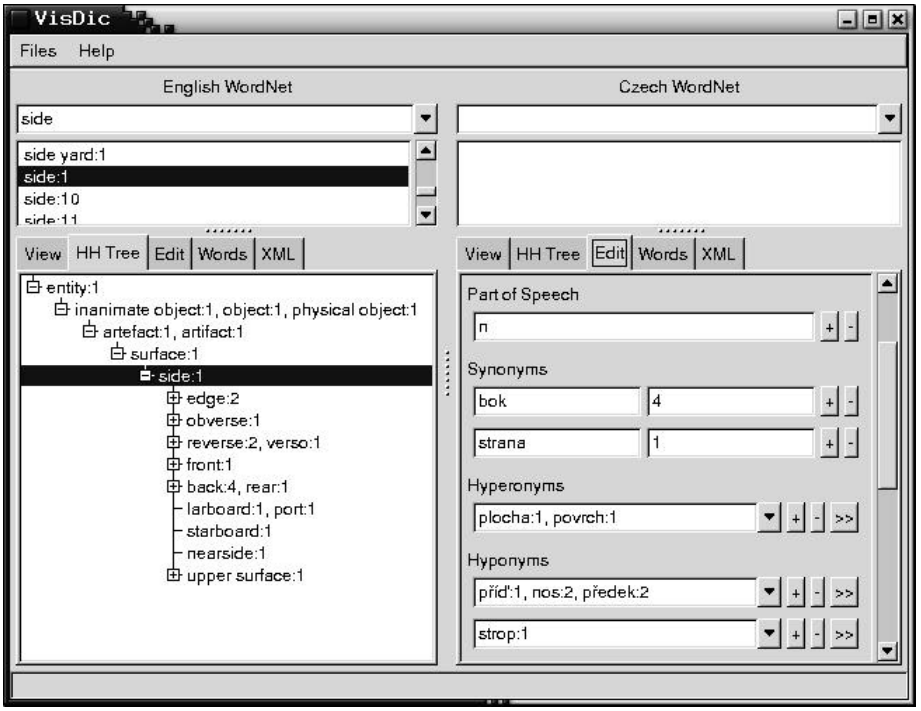


Figure 3. VisDic layout

1.2 The Types of Entry View

There are five basic types of entry view offered by VisDic.

1. **XML view** displays the raw XML format in which an entry is represented.
2. **Text view** shows the information contained in the individual part of entry, such as its head, senses, definitions, etc., in the user-specified format. It is possible to specify how the items will be displayed – in which font and which color - all in the dictionary configuration file (CFG).
3. **Tree view** shows the current entry as it is linked to the other entries. This information is then displayed in a tree structure. In the CFG file one can define which attribute will be understood as a parent and which attributes will indicate the children. Especially in WordNet databases, the typical tree structures can be formed by the hyperonym and hyponym ILR[2], or by some kind of the meronym and holonym ILR. Every tree node can be expanded or collapsed. During expansion, the number of children is counted and displayed. Furthermore, a node can be also fully expanded. It means that every node belonging to the subtree indicated by a given node will be expanded. This function is helpful for counting all the entries of the subtree.
4. **Editing view** enables to change information stored in the entry attributes by means of special edit boxes. The structure of the edit view can be defined in the CFG.
5. **Word view** is not specific for a current entry. In it one can view all the words belonging to a special attribute. This attribute can be also specified in CFG. The list is alphabetically sorted and can come in handy in the case when the user wants to browse the dictionary systematically. It is possible to drag a specified word to the query edit box and look up the corresponding entries.

2. Special WordNet Features

The VisDic tool displays specific functions as well. They help to maintain the WordNet databases. Although all features presented here can be also applied to other dictionaries, they were primarily intended for manipulation within WordNet databases. That is the reason why they are mentioned separately.

2.1 Functions for Finding Topmost Entries

The special function of the tree view is to find all the entries which do not have any parent. In a WordNet specification of hypero-hyponymical tree (H/H tree) it can find all synsets, which do not have any hyperonyms. In an ideal case these topmost synsets are exactly the ones corresponding to the top ontology synsets. However during WordNet editing some relations can be broken. Especially deleting the hyponyms belonging to a node can make the synset free, i.e. not having any parent, and thus they will become also the topmost ones. This function can be useful to prevent this kind of inconsistency.

[2] In fact, these two tags do not form a tree in all cases, because there is no restriction to have only one hyperonym for a synset, but for most of the synsets this condition is fulfilled.

2.2 Tree View Functions

VisDic has a special function in the tree view that is able to convert the whole structure of the subtree given by a specific node to the other WordNet. All the synsets that do not exist in the target database are copied to it. Then it is possible to translate their literals while their relations are kept as they are defined.

VisDic offers also a possibility to assign the key value to the current synset from the synset of another WordNet database. Then these two synsets become equivalent. This feature is helpful mainly when WordNet is built using Merge model. Developers of the new WordNet can build selected synsets, join them with ILR and then link them simply to another WordNet.

2.3 Importing and Exporting files

VisDic 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 that is not readable, but allows the fast searching and editing entries.

Another VisDic function can export all WordNet or the specified subtree of any kind to a XML file. This file can be also modified by another tool or application and can be later imported back to VisDic again. This behavior is very similar to the Polaris importing and exporting mechanism. However, every synset is stored at one line in the text file. This feature is suitable for getting statistics of the WordNet. Especially in Linux systems many statistics can be obtained just using simple scripts or even one command (such as grep, etc.).

3. Conclusions

VisDic is able to work with XML format, which can be regarded as a standard now and which is readable by many other applications. However, it can be seen that there are still functions, mainly related to the multi-user processing that should be added. VisDic tool is going to be integrated in WordNet Management System providing the necessary services for Balkanet project (see [4]). For this purpose a client/server version of VisDic tools is being developed presently by FI MU team.

References:

[1] Introduction to Wordnet: An On-line Lexical Database, Miller G.A., Beckwith R., Fellbaum Ch., Gross D., Miller K, Princeton University, 1993

[2] Vossen, P., Final Report on EuroWordNet, CD Rom, Amsterdam University, August 1999

[3] Polaris User's Guide, Louw M., Lernout & Hauspie - Antwerp, Belgium, 1998, p. 59-82

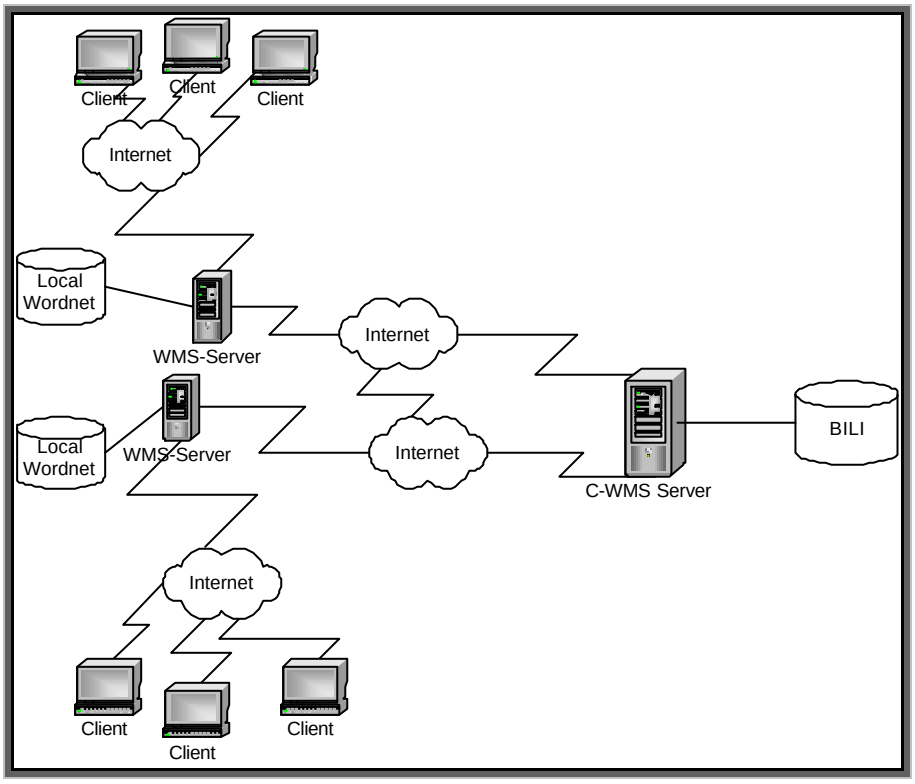
[4] Balkanet Project, No. IST-2000-29388, led by D. Christodoulakis, University of Patras, DBLAB.

WordNet Management System architecture (WMS)

WMS will follow a mixed approach regarding the architecture of the system: a client-server model for the clients (browsers, applications and Web Services users) and a distributed server model for the servers themselves. The basic model of WMS is comprised of two types of servers:

- the WMS Server, which is the basic type of WMS server and which is responsible for serving the requests of the clients. It hosts one or more versions of one or more monolingual wordnets and redirects requests for unknown data to the respective WMS Servers. The server possesses also the knowledge of the existence and capability of service of the other servers and is cooperating with the Central Server in order to get this information. In the basic model of the WMS Architecture, each partner has one such server.
- the C-WMS Server, which is the main key point in the WMS system. It has a dual role, providing the WMS Servers with information about the status of the network but also serving the requests for BILI links' data and domain information.

The basic model is represented below:



WMS basic services

The number of services, which have been implemented and provided so far, via the WMS Server are described below:

- *Synsetrequest*: query the appropriate WordNet file for synsets with the ILI or LITERAL value(s) of the message. There will be possible multiple selection, but only with the same type of key searching. Then the system returns all the available information for the synset(s).
- *Domainrequest*: returns the domain(s) for one or multiple ILI values. Currently returns the GLOSS of the linked ILI record instead of the domains, since the domain labels have not been applied so far.
- *Hyperonymrequest*: returns the path from the synset to the top of the tree it belongs (base concept), given its ILI value. Currently works for small size wordnets.
- *Hyponymrequest*: returns the subtree with the synset as its root, given its ILI value. Currently works for small size wordnets.
- *Full tree request*: returns the tree where the synset belongs to, given its ILI value. Currently works for small size wordnets (very size-sensitive).

The implementation of the above architecture has been confined to the limits of a prototype system. The system is being built gradually and assembled together as its development proceeds, maintaining its previous functionality and increasing it with the added parts. The main objective of this procedure is to maintain the 'openness' of the entire system. Considering this, WMS is developed as an open platform, which can easily interconnect with a variety of applications such as Web browsers, WordNet clients (such as VisDic) and other WordNet applications.

WMS Innovative features:

- **Data Storage Independency** Currently the system uses exclusively XML files produced by the VisDic tool, but the use of different data storage systems (such as a database) is also possible. Taking such an approach requires the definition of a standard WordNet XML Schema.

- **WEB Services** A Web service is an application that exposes a programmatic interface using standard, Internet-friendly protocols. Web services are designed in order to be used by other programs or applications rather than by humans.

- **Data Management** The main operations of the WMS will be: 1) linking of individual WordNets with the ILI and the Domain Ontology, 2) being able to ensure a high level of compatibility and navigational capabilities between the linked WordNets, 3) providing advanced ILI administration features and a flexible and unified way of data availability between WordNet applications.

Once the WMS is completed and the first milestone of the project will be achieved, i.e., end of the second year, the WMS will be made available to any interested parties.

Access to the WMS will be possible either via the WWW through the site of the project, or via WordNet clients i.e. applications specifically developed with the capability of being integrated into the WMS.

The purpose of the WMS is to provide all the information manipulated, and therefore it is not restricted to WordNet browsing. Any application requiring data concerning

WordNets can be connected to the WMS by using a standard communication protocol. For instance, within the Balkanet project, WMS is going to be applied for conceptual indexing purposes. The conceptual indexing application can be integrated into the WMS and/or interact with it, following the standard communication protocol.

User Group, Promotion and Awareness

Since one of the project's objectives aims at strengthening ties with the academic and information technology communities in European countries, BalkaNet's user groups fall within a wide spectrum of institutions and individuals. More specifically, academic as well as industrial parties have contacted members of the consortium in order not only to acquire more information about the project but also expressed their willingness to join the consortium. Several of them have been admitted access to the project's results on the grounds that this will hold solely for academic and research purposes. Moreover, various well-known linguistic communities have expressed their profound interest in the project's results and as such several publications and presentations of the project have taken place so far.

Following on from this and having in mind the project's final application members of the consortium have contacted various internet service providers in order for the latter to make extensive usage of the project's contribution in Information retrieval tasks. Once such party of the Greek Internet Service provider, namely OTENET that is actively participating in the project and has undertaken the task of incorporating the project's results in a commercial web search engine with the significant objective of improving retrieval performance for the languages in question.

Moreover, members of the consortium have been asked to attend various national and international conferences and meetings in order to disseminate the early project's results to a wider audience. In this respect a workshop on BalkaNet has been organized in conjunction with the 3rd International LREC conference, which took place in Las Palmas, Spain on 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.

The main topics of the workshop were:

- To what extend the architecture of semantic networks rely on language types?
- How structures of semantic networks affect performance of IR applications?
- Semantic relations of the less studied languages and how these are represented?
- Structure as language independent module.
- Are the assumptions of WordNet applicable to other language types?
- Standardization of WordNet representations with respect to metalanguages (XML etc.)
- Consistency checking, comparison and evaluation of WordNet modules

Furthermore, several presentations of the project have taken place in National and International Conferences, such as LREC 2002, 1st WGA 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 and many others, which are listed at the end of this document.

Future Work / Exploitation Prospects

Within the remaining of the project the following tasks will be completed. The WordNet Management System will have been fully developed and distributed to any interested party. Moreover, extensive evaluation of the quality and coverage of the synsets contained within each monolingual WordNet will take place. Emphasis will be given during this task so as to reassure maximal vocabulary overlap whereas at the same time ensuring quality of the links and synsets to be developed. In addition, the language internal relations of each monolingual WordNet are going to be further extended so as to capture any semantic relations that possibly exist. Finally, the project's results are going to be extensively used towards semantic classification of web documents while being indexed. In this respect the contribution of the project's users' groups, (e.g. Otenet Internet Service provider etc.) will be rather useful.

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.1.1	"Project Presentation" (web site, PPT presentation)	DBLAB
D.1.2	"Dissemination and Use Plan"	DBLAB
D.2.1	"Requirement analysis and specification of the methodology"	CTI
D.3.1	"Tools for the construction of the individual WordNets"	UOA

Published BalkaNet papers

"Balkanet-Design and Development of a Multilingual Semantic Network for the Balkan Languages", "Contemporary Information Technologies", October 2001
"Balkanet-A Multilingual Semantic Network for the Balkan Languages", 1st International Conference "Small languages at the Balkans and in Europe", November 2001
"Viewing semantic Networks as Hypermedia", LREC Conference, May 2002 D. Avramidis, G. Kourousias, M. Tzagarakis, S. Stamou, M. Kyriakopoulou
"Requirements for Domain – Specific WordNets", LREC Conference, May 2002 D. Christodoulakis, D.I. Koutsoumpos
"Glosses in WordNet 1.5 and their Standardization/Consistency", LREC Conference, May 2002 Karel Pala, Pavel Smrz
"WordNet Standardization from a practical point of view", LREC Conference, May 2002 Karel Pala
"Methodological issues in building the Romanian WordNet and consistency checks in Balkanet", LREC Conference, May 2002 Dan Tufis, Dan Cristea
"Balkanet – A Multilingual Thesaurus for the Balkan Languages", International Conference Romanian Language and Culture and the Globalisation, May 2002 Dan Tufis
"Lexical token alignment: experiments, results and applications", LREC Conference, May 2002 Dan Tufis, Ana-Maria Barbu
"A cheap and fast way to build useful translation lexicons" 19th International Conference on Computational Linguistics, COLING 2002 Dan Tufis
"Implementing the Greek WordNet: Computational Tools for Information Retrieval from electronic lexica and corpora", International Conference on Information Communication Technologies in Education, July 2002 M. Grigoriadou, E. Galiotou, E. Papakitsos, A. Charcharidou, E. Selimis
"A Stochastic POS Tagger", 27th International Conference ICT&P, June 2002

H. Krushkov, G. Tachev
"Bulgarian WordNet – Problem and Prospects", *International Conference Electronic Description and Edition of Slavic Sources,* September 2002 Svetla Koeva
"AR-Engine- a framework for unrestricted co-reference resolution", LREC Conference, May 2002 Dan Cristea, Oana-Diana Postolache, Gabriela-Eugenia Dima, Catalina Barbu
"Romanian linguistic resources and computer technologies applied to the Romanian language", *The identity of the Romanian language and literature in the globalisation perspective,* Institute of Romanian Philology of the Romanian Academy, August 2002 Dan Cristea, Dan Tufis
"Balkanet: Multilingual thesaurus for Balkan languages", International Conference Romanian Language and Globalisation,
"Logic for WordNet", "Annual Journal of Sofia University", 2002 Tinko Tinchev, Stoyan Mihov, Svetla Koeva, Angel Genov