ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume 7, Numbers 1–2, 2004, 9–43

BalkaNet: Aims, Methods, Results and Perspectives. A General Overview

D. TUFIŞ[1], D. CRISTEA[2], S. STAMOU[3]

[1]Institute for Artificial Intelligence, Romanian Academy, Bucharest, Romania
[2]Faculty of Computer Science, "Al. I. Cuza" University of Iaşi, Romania
[3]Research Academic Computer Technology Institute, Patras, Greece

E-mail: [1]`tufis@racai.ro`, [2]`dcristea@infoiasi.ro`, [3]`stamou@cti.gr`

Abstract. BalkaNet is an EC funded project (IST-2000-29388) that started in September 2001 and will end in August 2004. It aims at developing [109] aligned wordnets for the following Balkan languages: Bulgarian, Greek, Romanian, Serbian, Turkish and to extend the Czech wordnet previously developed in the EuroWordNet project. BalkaNet project has insofar delivered many useful results in the fields of both Computational Lexicography and Natural Language Processing. However, most of these results have been only partially disseminated in different conferences and journals. This is the first attempt to provide an overall description of the findings, methodologies and results of the project as well as a detailed account on each monolingual wordnet. The paper also presents the freeware multilingual tools designed for the development, maintenance and efficient exploitation of the aligned BalkaNet wordnets. A preliminary approach on BalkaNet's application towards indexing Web documents and Information Retrieval is described, following the consideration that semantic networks are valuable in the context of real world systems and user communities. Last but not least, a rather thorough analyses of wordnet applications over the last years is intended to put in evidence the hottest themes for further developments based on wordnets. The ultimate objective of this contribution is to spread the knowledge and experience that we have acquired, to the benefit of the research and industrial communities. We also hope that our shared experience will be helpful for other wordnet-builders.

1. Introduction

Semantic networks [96] are among the most popular Artificial Intelligence formalisms for knowledge representation that have been widely used in the 70's and 80's to represent structured knowledge. Like other networks, they consist of nodes and links. Nodes represent concepts, i.e., abstract classes whose members are grouped together on the basis of their common features and/or properties, while arcs between these nodes represent relations between concepts and are labelled so as to indicate the relation they represent. In a semantic network, usually, the concepts' labels are mnemonics, informative for the knowledge engineer developer. The semantics of the concepts resides not in the name of the associated labels, but in the concepts' properties and relations to other concepts of the semantic network. In the last 20 years or so there have been a tremendous resurrection of interest in semantic networks formalisms boosted, among others, by CYC, the impressive work of Lenat and his colleagues [54]. The ontological representation of general and domain specific knowledge is now claimed to be a sine-qua-non support to any attempt to intelligently solve the hard problems faced by the modern information technology. A special form of the traditional semantic networks came out from the pioneering work of George Miller and his co-workers [69] at Princeton University. They developed the concept of a lexical semantic network, the nodes of which represented sets of actual words of English sharing (in certain contexts) a common meaning. These sets of words, called synsets (synonymy sets), constitute the building blocks for representing the lexical knowledge reflected in WordNet, the first implementation of lexical semantic networks. As in the semantic networks formalisms, the semantics of the lexical nodes (the synsets) is given by the properties of the nodes (implicitly, by the synonymy relation that holds between the literals of the synset and explicitly, by the gloss attached to the synset and, sometimes, by specific examples of usage) and the relations to the other nodes of the network. These relations are either of a semantic nature, similar to those to be found in the inheritance hierarchies of the semantic networks, and/or of a lexical nature, specific to lexical semantics representation domains. The convergence of the representational principles promoted both by the domain-oriented semantic networks and ontologies, and by WordNet's philosophy in representing general lexical knowledge, is nowadays an apparent trend, motivated not by fashion, but by the significant improvements in performance and by the naturalness of interaction displayed by the systems that have adopted this integration. Several NLP systems based on semantic networks initially (80's) relied on (limited) domain specific semantic lexicons for mapping synonymic words used in the input to the same concept of the underlying semantic net. The IURES system [112], [113] is just one such example.

However, the tremendous technological advancement of the recent years in computers' speed and storage capacity, the unforeseen Web evolution, the widespread of understanding and usage of WordNet, as well as the maturity of the ontology-based technologies, made possible up-scaling the integration of domain knowledge and lexical knowledge at an unprecedented level. This interdependency, which is not always explicit, motivated several researchers' doubts on language independent ontologies [97], [40], [38], etc.

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

The EuroWordNet (EWN) project (LE-2 4003 & LE-4 8328), which started in March 1996 and ended in June 1999, extended the PWN approach with the multilingual dimension adding an Inter-Lingual Index (ILI) to which all the monolingual wordnets for the languages represented in the project were aligned. The ILI was based on the PWN 1.5, the synsets of which played the role of language independent concepts. The ILI was further extended with some language (other than English) specific concepts. Another major extension was the association with each of the so-called Base Concepts of an ontological description subject to be inherited by all more specific concepts in the ILI. For the monolingual wordnets the same structuring as in PWN [69] was preserved and via the ILI (interconnecting the languages) it is possible to go from the words in one language to semantically close words in any other language. The index also gives access to a shared top-ontology of 63 semantic distinctions. This top-ontology provides a common semantic framework for all the languages, while language specific properties are maintained within the individual wordnets. The languages represented in EWN were Dutch, Italian, Spanish, German, French, Czech, Estonian and obviously English. The alignment of the monolingual wordnets on the basis of the interlingual index as well as the shared top-ontology turned the EWN multilingual lexicalized semantic network into a multilingual lexical ontology. A detailed presentation of the principles, methodology and results of the EWN project is given in [126] and on the EWN website (`http://www.illc.uva.nl/EuroWordNet/`).

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

A major contribution to the furthering of the EWN principles [100] is the ongoing European project BalkaNet (IST-2000-29388) which initially aimed at extending the pool of the EWN languages with five South-Eastern European languages from the Balkan area: Bulgarian, Greek, Romanian, Serbian and Turkish. In the consortium have been included the Czech and the French teams that participated in the EWN, to liaise towards a perfect compatibility with previously developed wordnets. Also the coordinator of the EWN project, Dr. Piek Vossen, was solicited and accepted to be a consultant for the BalkaNet project. Besides compatibility with the other aligned wordnets, the BalkaNet project ambitioned to a better quality and to a much wider cross-lingual coverage than in EWN. Therefore, the quality control was much stricter.

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

Before going into further details, let us define three terms relevant for the discussions to follow: "sense", "meaning" and "concept". Although closely related, and sometimes interchangeably used, these notions are slightly different distinguishing the perspective from which the encoded knowledge is considered. The notion of *sense* is strictly referring to a word. The polysemy degree of a word is given by the number of senses the respective word has. A traditional explanatory dictionary provides definitions for each sense of a headword. The notion of *meaning* generalizes the notion of *sense* and it could be regarded as a set-theoretic equivalence relation over the set of senses in a given language. In colloquial speech one says *this word has the same meaning with that word* while a more precise (but less natural) statement would be *the M^{th} sense of this word has the same meaning with the N^{th} sense of that word.* Synonymy, as this equivalence relation is called, is a lexical relation that represents the formal device for clustering the word senses into groups of lexicalized meanings. The meaning is the building block in wordnet-like knowledge representations. In PWN and all its followers the meanings in the respective languages are represented as *synsets* (synonymy sets) and they are implemented as sets of word senses. Each synset is associated with a gloss that covers all word senses in the synonymy set. The meaning is thus a language specific realization of a *conceptualization* which might be very similar to conceptualizations in several other languages. Similar conceptualizations are generalized in a language independent way, by what we call *interlingual concepts* or simply *concepts.* The meanings in two languages that correspond to the same concept are said to be equivalent. One could arguably say that the interlingual concepts cannot entirely reflect the meanings in different languages (be it only for the historical and cultural differences), however, concepts are very useful generalizations that enable communication across speakers of different natural languages. In multilingual semantic networks the interlingual level ensures the cross-lingual navigation from words in one language to words in the other languages. Both EWN and BalkaNet adopted as their interlingual concepts the meanings of PWN. This choice was obviously a matter of technological development and a working compromise: the PWN displayed the greatest lexical coverage and is still unparalleled by any other language. To remedy this Interlingua status of English, both EWN and BalkaNet considered the possibility of adding in the ILI concepts which represent language specific meanings (or meanings specific to a group of languages).

The remainder of the paper is organized as follows: Section 2 presents the motivation and the general objectives of the BalkaNet project; Section 3 provides a thorough presentation of the specifications adopted for developing BalkaNet with emphasis on the methodology that was followed for developing the monolingual wordnets. The resources that were employed towards lexical data acquisition are presented and a description of some of the tools developed for the processing of these resources is given. Following on from this, some quality control issues and highlights the main approaches adopted towards validating the quality of the monolingual wordnets. Section 4 reports on BalkaNet's application in the framework of a Web search engine, and introduces our approach towards conceptual indexing by utilizing BalkaNet's hierarchies. Section 5 provides an up-to-date overview of the worldwide wordnet developments and their applications during the last four-five years demonstrating the huge potential of the multilingual wordnets in the immediate future. Section 6 concludes the paper with some general remarks emerging from our experience and points to future research directions.

2. Background Presentation of BalkaNet

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

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

G1) developing at least 8 000 synsets per new language-specific wordnet, commonly selected so that even with this small size, the wordnets should be useful in real applications;

G2) ensuring maximal interlingual overlap among the BalkaNet wordnets and compatibility with the wordnets developed in the EWN project;

G3) building free software tools for the efficient management and exploitation of the multilingual semantic lexicon;

G4) development of application demonstrators such as Word Sense Disambiguation (WSD), intelligent document indexing, cross-lingual Information Retrieval

(CLIR), etc. However, the project is still under development and, as such, it partially deals insofar with application issues. In particular, a system for WSD has already been developed (see [45] in this volume), intelligent document indexing is underway via the search engine and the same goes for Multilingual IR.

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

S1) the inter-lingual index (based on PWN 1.5) and the inter-lingual relations were defined the same way as in EWN; based on this decision was possible to select a set of common concepts to be implemented in each BalkaNet wordnet thus maximizing and controlling the cross-lingual coverage;

S2) since the available language resources, useful in building the monolingual wordnets, were different for each partner, both in format and coverage, each team had to build their own acquisition, development and validation tools so that to make maximum use of the available data; however, because all the wordnets were supposed to be integrated into a single multilingual environment, a common XML format was agreed (see [37] and [106], in this volume); this format is used by the BalkaNet multilingual viewer and editor VisDic [88]. The most recent and more powerful version of VisDic is presented in [37], this volume.

S3) because of various improvements apparent in recent versions of PWN we decided to update consequently our inter-lingual index, so that the final BalkaNet multilingual database is based on the PWN 2.0. This is not really a departure from the EWN compatibility (based on PWN 1.5) since more than 90% of the mappings among different versions of PWN (1.5, 1.6, 1.7.2, 2.0) are done automatically.

S4) to ensure quality control over the monolingual wordnets the consortium decided a set of validation tests, checking the syntactic and structural correctness (see [106] in this volume);

An initial step in the BalkaNet project was bringing the PWN in the same XML format to be followed by all the monolingual wordnets. The synset IDs were given unique values made up from the string "ENG15–" (to specify the used version) followed by a sequence of digits representing the offset in the original database of the respective synsets, followed by a tag denoting the part of speech of the literals in the encoded synsets. The BalkaNet initial ILI was represented by the codes representing ID values of the synsets in the XML version of the PWN 1.5[1]. The interlingual alignment is made explicit by assigning the ILI in all languages to the synsets that are equivalent to the PWN with the same ID value. The concepts to be implemented by all the monolingual wordnets, described in the next section, were specified in terms of the ILI codes from which every partner could visualize the associated synset in PWN. The commonly agreed set of concepts (BCS: BalkaNet Concept Set) was obligatory for each monolingual wordnet and it contains 8 516 concepts. Besides BCS each partner has the autonomy to extend his/her own wordnet by selecting other ILI codes according to language specific criteria. However, since the monolingual selection cri-

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

teria were similar, more than 8 516 common concepts are found between different pairs of languages.

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

In the vast majority of cases the hierarchical structures in PWN (nouns and verbs) were preserved over the monolingual wordnets following the *principle of hierarchy preservation* [114]. The wordnets hierarchies are inheritance structures (more often than not a synset has only one direct parent) with lower meanings being specialisation of their ancestors.

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

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

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

After each concerted development phase of the monolingual wordnets, the ILI concepts are normalized (see [106], in this volume) for being loaded into VisDic, the BalkaNet multilingual browser and editor. VisDic has powerful editing facilities by means of which an expand model approach (cf. following section) in the development of a wordnet is strongly supported. Due to its browsing facilities, VisDic supports the merge or combined approaches as well, pinpointing alignment problems which can be easily corrected. As a multi-wordnet viewer, it allows for synchronized search: the search is performed in a source wordnet but (via the ILI) the results are (or can upon request) be displayed for all other target wordnets loaded within the editor. VisDic is an open source tool that currently runs under both Linux and Windows platforms.

In addition to the multilingual wordnets editor and viewer, it has been implemented a storage, querying and browsing infrastructure, namely the Wordnet Management System (WMS) [47] which enables the efficient navigation within and across wordnets. WMS is essentially a distributed network of servers, each one hosting a monolingual wordnet. A central server is responsible for establishing a coherent communication among the peripheral servers and it also holds responsibility for handling multilingual search requests. A variety of services have been incorporated into the WMS, which enable the efficient navigation within wordnets' hierarchies and the retrieval of their information contents. The current version of the Wordnet Management System is described in [47], [48].

Besides checking the syntactic and structural correctness of the wordnets as they were developed, a more challenging validation process is prepared for the end of the project, when the core monolingual wordnets will be in a stable state. This is called the *semantic interlingual validation* [45] and it will check, against a multilingual parallel corpus, how the synsets of each monolingual wordnet cover actual use of language and to what degree the established interlingual equivalences among synsets of different wordnets are supported by parallel human translations. This procedure has been already applied to the Romanian-English pair of wordnets, a detailed presentation being provided in [45], this volume. Ensuring an accurate inter-lingual alignment and a large cross-lingual coverage is essential for the performance of the final project's application, which envisages the incorporation of BalkaNet resource in an IR system.

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

3. Design Strategies of the BalkaNet Semantic Network

This section elaborates on the design strategies adopted in order to achieve the main goals of the project described in the previous section.

Following the principles adopted in EWN [124] and PWN [69], 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, 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) [123]. 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 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.

3.1. Selection of the BalkaNet Common Set of Concepts

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

The first BalkaNet Concept Set (BCS1) was identical to the EWN Base Concept set. Base Concepts were selected for reasons convincingly argued in [100] and they represent concepts that are lexicalized in all the languages represented in the BalkaNet resource. In the present versions of the BalkaNet ILI, the number of BCS1 concepts is 1 218[2]. As in EWN, all concepts in BCS1 have attached a Top Ontology description (cf. [125]). For the selection of BCS2 and BCS3 the concepts which were lexicalized in most languages represented in EWN were taken into account. This statistical information was provided by MEMODATA, our French partner. Also, each partner suggested a list of candidate concepts that would be relevant for their languages. The concepts proposed by at least two partners were also considered as candidates for the common set of concepts. An additional selection criterion was that the concepts in BCS1, BCS2 and BCS3 should correspond to dense sub-networks in PWN. We called this selection restriction the *conceptual density* criterion and it can be stated as follows:

a) once a nominal or verbal concept (i.e. an ILI concept that in PWN is realized as a synset of nouns or as a synset of verbs) was selected in the BCS, all its direct and indirect ancestors (i.e all ILI concepts corresponding to the PWN synsets, up to the unique beginners) will be also included in BCS.

b) all the descriptive adjectival concepts (i.e. ILI concepts that in PWN are realized as synsets of descriptive adjectives) included in BCS should represent values of attributes named by nouns already presented in the chosen set of BCS. This relationship is encoded in PWN by the relation *be-in state.*

A detailed description of the BCS selection process is given in (BKN-D.4.2, 2003) but the figures have changed due to migration from PWN 1.7.1 to PWN 2.0. Currently, the set of BCS (1, 2 and 3) consists of 8 516 concepts implemented in all but one of the monolingual wordnets. The exception is the wordnet of the Serbian subcontractor which implemented 5 381 concepts of the BCS (anyway, 3.5 time more synsets than planned in the project). The difference is explained by the fact that the Serbian wordnet development started in a later phase of the project.

Because the conceptual density criterion operates only on nominal, verbal and (descriptive) adjectival synsets, the BCS includes concepts that correspond neither to adverbial synsets nor to relational adjectives synsets. The selection of these categories of synsets was left in the responsibility of each partner. Each monolingual wordnet has been further extended beyond covering the BCS. In general, the wordnets enrichment process followed a top-down approach starting with the synsets that have been already mapped onto the BCS. The monolingual wordnets extension was mainly guided by language-specific criteria in order to make sure that in spite of the ILI guided development of the first aligned synsets, the lexical distributional properties in each language were not overlooked. Each team responsible for their own language

[2] This is slightly different from the BC in EWN: 1 024 (representing 796 nominal concepts and 228 verbal concepts). The difference is due to finer grained synsets of PWN 2.0 (BalkaNet ILI) as compared to PWN 1.5 (EWN ILI).

wordnet made various statistical studies on large corpora or used existing lexical resources to identify frequently occurring general words or word senses that should be included into the respective wordnets. More information on this aspect is provided in the papers on this volume reporting on monolingual wordnets.

During the concerted development of the BalkaNet wordnets, several quality control policies have been adopted and implemented by each wordnet developer. Yet, an overall quality control was performed by one of the partners. The methodology and the evaluation results are largely described in a separate paper of this volume [107].

Some extended monolingual wordnets included synsets that either represent concepts specific to some Eastern European area, or they automatically derive because of their regular morphological patterns and their easy to predict semantics. The analysis of the latter monolingual synsets in a multilingual context seems to open very interesting pathways. A lexical relation found in one language (by means of a derivative analysis), the semantics of which is predictable, might be relevant as a paradigmatic relation in many non-derivative languages. For instance, in Turkish, two derivationally related words might correspond to two morphologically unrelated words in Greek. However, the semantics of the Turkish derivative affix could be very useful in assigning between the two Greek words a semantic relation (a case relation for instance [9]).

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

3.2. BILI: the Structured BalkaNet Interlingual Index

As mentioned before, initially the BalkaNet project adopted EWN's ILI, which is defined as an unstructured collection of concepts [89] represented by records of the form ($\langle$ ILI-index $\rangle$ $\langle$ ontological description$\rangle$ $\langle$gloss$\rangle$ $\{\langle$domain$\rangle\}$) and the development of the monolingual wordnets started on that basis. However, ILI's unstructured nature introduced some difficulties while trying to define the best translation equivalents for the ILI's concepts. Quite often, the ILI gloss was not sufficiently informative to make a right lexicographic choice and the lack of any direct access to the hierarchical contexts of the target concept was frustrating. Most importantly, though, an unstructured ILI would hamper the project's final application in an IR system. Therefore, with the release of PWN 1.7.1 it was decided to replace the bag of ILI records with the PWN 1.7 itself, exported from its original database format into the BalkaNet XML format. The IR system which is the project's final application will make intensive use of the structures defined over ILI. This way we have imposed the principle of hierarchy preservations [114] over the whole network of wordnets in BalkaNet. Relying on a structured ILI, when following the expand model, improved the quality of the synsets translation. Where the merge model was followed, using a structured ILI improved the quality of the synsets interlingual mapping. At a later phase in the project, our ILI was further updated by switching to the latest PWN version released, namely version 2.0. Issues related to ILI updates were automatized by the Czech partner to a large extent, but some manual work was necessary. Specifically all monolingual concepts that were linked to PWN 1.7.1 ILI's nodes by that time were automatically mapped against their equivalents in the PWN 2.0 ILI nodes. 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. 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) [79]. SUMO is an ontology that was created by merging publicly available ontological content [80] into a single structure and provides definitions for general-purpose terms. SUMO was used as the base ILI ontology for several reasons. First and foremost, it has already been mapped to PWN's synsets that form the main repository of BalkaNet's ILI. Secondly, it combines resources from many fields, and, most importantly, it is freely available and extensible. 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.

3.3. Lexical Data Acquisition

In the EWN project there were defined two distinct development models, namely the expand and the merge model [122].

The expand model essentially is a translation-driven wordnet development approach, in which the literals in each PWN synset are being translated as faithful as possible. The relations of the translated synset are to a large extent automatically imported and the original gloss is translated into the target language. During this process, some new literals could be inserted in the target synset or some literals in the source synset hard to translate are ignored. In such an approach a high quality bilingual machine readable dictionary (MRD) can speed up dramatically the development in competition with a bilingual human lexicographer. Such an approach was strongly supported by the VisDic multilingual editor and browser.

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

Both models exhibit advantages and disadvantages. To benefit from the advantages offered by both models it was decided to develop BalkaNet by using a combination of both models. Depending on the lexical resources the partners had at their disposal, the individual wordnets development approaches came closer to one or the other development models. However none of the partners adopted a pure expand or merge model. This way it is reassured that the monolingual wordnets are richly encoded and comparable across languages (guaranteed by the expand model), while at the same time language-specific properties are reflected into the monolingual wordnets (guaranteed by the merge model).

The papers in this volume describing each monolingual wordnet provide more in-depth details on the language resources used in the respective wordnets development and, where necessary, the tools[3] that were developed for the purpose of this endeavour.

3.4. Quality Control Policies

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 connection to the other wordnets within BalkaNet, on the other hand. Besides the validation tests developed by each partner for their own wordnets, centralized and cross-lingual 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 relation loops in the wordnets, etc. A web implementation of these tests and several others has been also implemented (by the Bulgarian team, reported in [46], this volume) 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

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

partners resulted in a quality control methodology, which was implemented in various versions. One of the papers included into this volume [106] addresses this very methodology and its implementation.

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 parallel texts should also be retrieved (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” [22] 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 1 810 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 (PWN 2.0). One of the Romanian partners developed a Word Sense Disambiguation platform [118] called WSDtool incorporating a highly accurate word aligner in parallel corpora [115]. This system, as well as the results and conclusions of the cross-lingual validation between PWN and Romanian wordnet, are described in a separate paper [45] included in this volume. Similar evaluations are currently conducted for all the other BalkaNet wordnets.

4. Utilizing BalkaNet’s Conceptual Taxonomies to Index Web Documents

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 still underway and it is expected to be over by August 2004. In this section we describe how we envisage BalkaNet's use in IR and we present some early achievements that have been accomplished so far in this direction.

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 *paraphrase problem* [129], 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 [101], so as to perform content-based IR [4] as opposed to exact keyword matching.

4.1. Conceptual Clustering

To enable documents' conceptual organization, the ILI's conceptual domains are treated as clusters under which Web documents would be classified. Conceptual clustering takes place via an internal mapping between documents' terms and ILI's concepts and by calculating their semantic similarity [110]. Semantic similarity is captured by means of the information content of the concepts in the hierarchical net, following the approach reported in [99]. The main objective behind clustering Web documents on the basis of their conceptual relatedness is to provide a meaningful organization of the indexed data. Semantically grouped Web documents are expected to improve the performance of the engine's indexing modules.

Conceptual clustering 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 via the *belongs_to* semantic relation. 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[4] were selected, namely Law, Economy and Politics that would form the conceptual domains under which BalkaNet's taxonomies would be structured. Having defined the conceptual clusters under which Web documents would be organized, the SUMO ontology [79] was employed, out of which all ILI concepts falling into any of the pre-defined conceptual domains were extracted. SUMO's incorporation within the ILI, as well as its main contribution in structuring BalkaNet's Interlingual, are described in section 3.2 of this article. 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.

ILI's conceptual domains are going to be utilized by the engine's clustering modules that will attempt the organization of the Web documents at the index level. More specifically, the engine's storage space that is currently under development, comprises three distinct indices. Each index corresponds to one of the pre-specified BalkaNet domains (i.e., law, politics and economy) and stores Web documents coming from the BalkanTimes web site that semantically belong to each domain. This way all documents dealing with the theme of economy are grouped together under the "*economy*" index, documents of the politics theme are stored under the "*politics*" domain, and so forth. The clustering module of the engine is responsible for deciding to which of the conceptual domains a given document should be assigned. Such decision is based on the concepts mentioned by the document that match a specific wordnet taxonomy. Once the appropriate cluster is decided, the engine's indexing modules are responsible for directing the given Web document under the decided domain. Besides the three conceptual indices, a supplementary plain index is maintained. This index stores Web documents that could not be classified under any of the above domains. The conceptual clustering module is still in an early development stage. Alternatively, experiments could be conducted with the agglomerative word sense clustering algorithm used, as a back-off mechanism, in the WSDtool word-sense-disambiguation system [45]. Although developed for another purpose, the word sense clustering algorithm can be easily adapted for document clustering.

4.2. Towards Conceptual Indexing; Challenges to Conceptual Retrieval

The first step to be taken before the actual indexing of the Web documents, concerns their classification on the basis of their topical relations and semantic similarities. More specifically, each Web document crawled by the search engine passes through the clustering algorithm in order to decide on the domain under which the given document would be classified and stored. Once this decision is made the engine's indexing algorithm is responsible for actually indexing the document under the respective domain's index and for retrieving it upon users' requests.

[4] Selection was based on quantitative criteria, i.e., the number of Web documents corresponding to each thematic category.

In order to enable the clustering of Web documents, their morphological preprocessing is required, so as to extract a set of lexicalized concepts out of them. Morphological processing involves: (i) document tokenization, (ii) POS-tagging, and (iii) lemmatization.

Henceforth, a frequency occurrence formula, namely the normalized *tf*idf* weighting scheme [102], is employed against all documents' content-terms[5]. Terms with high frequency weights are those that lexicalize the most representative concepts of a given document and are the ones on which clustering will be based. Conceptual clustering 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. More specifically, in cases where the high-weighted terms extracted from the Web documents are monosemous and each one belongs to a unique ILI node, then the given document is clustered under the domain of the taxonomy that contains the most ILI matching nodes. On the other hand, where document's terms match several ILI nodes, then some disambiguation is needed before deciding the domain under which the respective document will be classified. Disambiguation is attempted by estimating the semantic similarity between the different ILI matching nodes. Precisely, we assume that ILI nodes being closer in the hierarchy display more common semantic properties in comparison to nodes that are spread over distant parts of the taxonomy. Based on this assumption, we attempt to select for each document term the ILI node that best reflects the term's semantics. This selection takes into account the number of the ILI matching nodes that belong to each conceptual domain, their distance within the ILI graph and their density. The closer two ILI nodes are within a deep and dense part of the ILI hierarchy, the closer these concepts are in terms of semantic relatedness. The discriminative power of the information content measure (or any other measure) of semantic similarity, frequently, is not enough to guarantee a sharp distinction among the classification domains. To account for such border line classifications we allow for a document to be clustered under multiple conceptual domains. This way a document about economic legislation would be indexed under both *Economy* and *Law* conceptual domains. To ensure the soundness of our approach we are also investigating other proposals towards calculating semantic similarity, such as the conceptual density approach introduced in [1].

An alternative solution will be to adopt a vector-space strategy as in [116] where all senses are taken into account, but with different weights established by a previous training phase on already classified documents. A semantically ambiguous word could provide different weighted hints for more classification domains but the final decision would come from all the words considered in building the document' vector space. An interesting result demonstrated in [116] is that not all the words of a document are necessary to be taken into account but only a random sample.

Using the ILI for the document conceptual classification, indexing and retrieval, in the context of ILI-based aligned wordnets, makes irrelevant the language of the

[5] As content-terms we consider those having received a N (noun), V (verb), ADJ (adjective) or ADV (adverb) POS-tag.

document to be classified or retrieved. When it happens that the respective document is available in two or more languages (for which ILI-based aligned wordnets exist) their conceptual classification and indexing become significantly easier as argued by [45] in this volume.

Irrespective of the implementation solution, a training set of data of already classified domains should be used. Fortunately, the documents at `http://www.balkantimes.com` are already thematically classified and, as such, provide an invaluable multilingual source for training and evaluating of the conceptual classification and indexing engines.

We mentioned that for clustering, indexing and retrieval purposes, the BILI has been conceptually enriched so that it can be regarded (when needed) as a *conceptual taxonomy*. The objective of the conceptual taxonomy is to support the engine's indexing modules with information on the documents' semantics so as to index the documents into conceptual domains, each conceptual domain corresponding to a separate index.

However, developing a language-independent, consistent and comprehensive conceptual taxonomy is not an easy task. The integration of the SUMO ontology domains in Balkanet helped us to structure the ILI taxonomy in a meaningful way and gave us the flexibility to enrich ILI with new concepts without imposing any need for structural changes. A benefit for clustering ILI's own taxonomies under SUMO domains is that each taxonomy can be viewed as a specialized-semantic network and, as such, can be employed by applications that require domain-specific knowledge sources. Another advantage of our structured ILI, besides its universality, is that it can be extended with other languages and/or concepts and it can be reusable by other applications without requiring any modifications.

However, a significant amount of work remains to be accomplished prior the proposed indexing module is fully functional. In this respect we are currently testing the performance of our conceptual indexing framework against a small set of Web documents collected from the Southeast European Times Web site, which contains news articles in all Balkan languages.

In the future, we plan to develop an efficient searching mechanism that would utilize conceptual taxonomies while processing submitted queries in order to retrieve relevant documents. An important component of the searching mechanism is a query expansion module (responsible for expanding queries with semantically related terms encoded within BalkaNet), that is currently under development. Query expansion is performed within and across wordnets' synsets and aims at providing the users of the engine with alternative wordings for their submitted queries, in any of the languages represented within the BalkaNet repository. We also envisage incorporating in our retrieval system advanced searching modes that would enable the user to specify the conceptual domain out of which s/he wishes to retrieve information. In sum, we argue that conceptual taxonomies have a strong potential towards content-based IR and they can significantly contribute in helping information seekers satisfy their needs. Many challenging issues still need to be addressed before we end up with an operational conceptual information retrieval system.

5. Wordnet Applications

Within the community that contributed to the BalkaNet project there is a strong determination to continue the collective work that resulted in the construction of this considerable aligned multilingual semantic network. Future work will envisage quantitative developments of individual wordnets as well as applications in the area of IR and in other directions. We considered therefore important to overview the state-of-the-art of the worldwide wordnet developments and their applications at least for a period covering the years of the project. Such an analysis should configure the hottest fields of interest for the boosting of future applications.

Therefore in the following subsections we will review some of the work that reported applications of wordnets in domains that include: Word Sense Disambiguation (WSD), Information Retrieval (IR), Question Answering (QA), Information Extraction (IE), summarization, Anaphora Resolution (AR) as well as applications induced by aligned wordnets and efforts oriented towards enriching wordnets.

The investigation resumes work on wordnet developments and applications reported at the most significant Computational Linguistics, Language Technology and Artificial Intelligence conferences and their conjoined workshops (ACL, NAACL, EACL, COLING, LREC, GWC, AAAI), as well as in some of the most representative journals in these fields, along the period 2000–2004.

5.1. Applications in Word Sense Disambiguation and Information Retrieval

Disambiguation of senses, in itself or integrated in an IR task, remains the most abundant wordnet application [75]. Reciprocally, the already rich experience of the first three SENSEVAL[6] editions reveals the growing interest of using wordnet for WSD tasks.

The most often expressed discontent in using PWN for WSD is its too fine-grained sense distinctions [87], [67], [26], [90]. WSD can be improved by a proper grouping of word senses that aims at graining coarser sense distinctions. Most intricate, human and machine errors in distinguishing senses are not reflected in the hypernym relations [49]. Different criteria other than looking in the hierarchy are not easy to find: syntactic criteria are based on differences in subcategorization frames, as in Levin's classes [55], while semantic criteria make use of differences in semantic classes of arguments (abstract versus concrete, animacy versus inanimacy, different instrument types, etc.), differences in entailments (whether an argument refers to a created entity or a resulting state), differences in the type of event (abstract, concrete, mental, emotional, etc.), etc.

In some approaches PWN or other wordnets are used for WSD in combination with other linguistic resources. So, Molina et al. [74] employ a statistical model (Hidden

[6] SENSEVAL is a series of competitions aimed at advancing the state-of-the-art in WSD. The first SENSEVAL took place in the summer of 1998, for English, French and Italian, culminating in a workshop held at Herstmonceux Castle, Sussex, England. The SENSEVAL-2 workshop was held in conjunction with ACL/EACL 2001, in Toulouse, France. The third SENSEVAL workshop was held in Barcelona in July 2004 in conjunction with ACL-2004.

Markov) in WSD. SemCor[7] is used to train and evaluate the model to distinguish WordNet 1.6 senses. Mihalcea and Moldovan [65] also use SemCor, but their method is symbolic and exploits the PWN structure (synsets and hypernymy relations). The information obtained from the WSD module is then used in an indexing process, where the word stem and its semantic tag (synset ID) are marked. The retrieval is based on this kind of word/semantic indexing applied to both query and documents. Kwong [51] reports a WSD method based on an integration of 90% of noun senses of Roget's Thesaurus[8] and PWN using a structurally based sense-mapping algorithm. The references [117], [118] (see also [45] in this volume) are using aligned wordnets and translation equivalents to achieve simultaneous word sense disambiguation in both languages of a parallel corpus.

5.2. Applications in Question Answering

The recent interest in QA research is due to the need for more and more sophisticated paradigms in IR. QA tasks, especially as defined by the TREC[9] series of competitions, generally refer to encyclopedic or factual questions that require concise answers. QA is thus an area of IR but which is more intimately linked to NLP techniques [30].

In many approaches of QA that make use of wordnets, this resource is used to unify the semantic form of the question with the semantic form of the answers from the retrieved paragraphs. In [34], [85], [86] questions and answers are parsed to a dependency structure. Questions are classified in a collection of answer types that are unified towards similar representations of the answer. When unification fails, feedback loops that produce alternations in both question and answers are used. The answer types are developed to the level of an ontology in [86]. Although very different from the PWN hyperonym hierarchy, this ontological resource can be searched to identify specific semantic category representing the answer type. PWN is then used to generate keyword alternations (morphological – to match forms of irregular verbs, lexical – to match forms within a synset, and semantic – to match semantically related words in the question or the answer context). The authors show that the overall precision became more than double when PWN use was integrated into the system. In [71], [36] PWN is used to parse answers from paragraphs that had been identified to contain the question keywords. Together with gazetteers, WN helps in configuring heuristic rules for the recognition of names, persons, organizations, locations, dates, currencies and products. In [57], [58], [60], PWN and the Italian wordnet are used to improve the recall of a Web-based IR system by expanding key-words found in the question with

[7] SemCor is a corpus built from the Brown corpus for American English of the years 1960s which was annotated for word senses (nouns, verbs, adjectives and adverbs) by the Princeton team and used for the development of PWN.

[8] Roget's Thesaurus was invented by Peter Mark Roget (English, 1779–1869), first published in 1852. Words are grouped by meaning and semantic distance and was originally intended as a memory-aid for writers – to find the most appropriate word.

[9] Text Retrieval Conference (TREC) is a series of workshops co-sponsored by the National Institute of Standards and Technology (NIST) and the U.S. Department of Defense, designed to advance the state-of-the-art in information retrieval. The series was started in 1992 as part of the TIPSTER Text program, and will held in 2004 its 13th edition.

synonyms of the domain-related (a variant of sense-related) synsets. Words in the text are tagged with domain labels instead of sense labels, using domain-annotated nouns in both PWN and MultiWordNet (the Italian EuroWN aligned wordnet).

Mann [61] introduces a statistic criterion, called Mutual Information, for inducing correlations between semantic tags and question classes. This criterion is used to generalize over PWN classes, with application in improving a short answer extractor.

Zajac [130] proposes an ontology-based semantic framework for QA in which both questions and answers are parsed into underspecified semantic expressions and answer retrieval is configured as a subsumption and unification process between these expressions. PWN, like any other language ontology, can be used as a source of ontological rules. Moldovan and Novischi [73] describe a methodology to derive lexical chains from Extended WordNet, which are then applied to QA from free texts. Extended WordNet is described below in subsection 5.7. Topical relations can be expressed as lexical chains.

5.3. Applications in Information Extraction, Authoring and Summarization

Chai [14] presents an approach to generate rules for filling data templates as those used in IE, based on PWN. The creation of a semantic space that models a domain, used in an IE setting, is described by Harabagiu and Maiorano [32]. Their method exploits the richness of lexico-semantic information of PWN and the collocational data extracted from corpora for acquisition of linguistic patterns for IE. The method is general enough to allow porting to open-domain IE.

The NAMIC project [6], [7] proposes an approach for multilingual authoring (creating hypertext links among a collection of documents) based on knowledge-intensive IE. At the basis of the hyperlinking is the detection and interpretation of 'events' on the basis of a World Model, which is a collection of subcategorization patterns for verbs within a certain domain. These patterns have references in the EWN hierarchy. The adoption of the Base Concepts ontological structure of EWN augments the semantic classes from 7 (the number of Named Entity classes in MUC – Message Understanding Conference) to 1 024 (the number of EWN concepts).

Harabagiu and Maiorano [33] describe a method of obtaining multi-document summaries with a system called GISTexter. Their method extends single-document summarisation techniques to dealing with multiple documents by involving the notion of topic representation (sets of inter-related concepts, implemented as frames having slots and fillers). Extraction of information for existing frames follows the classical IE slot filling paradigms. Ad-hoc templates and their corresponding extraction rules for new topics are inferred by mining PWN for topical relations, which are revealed as paths between PWN synsets on the bases of gloss and hypernym relations.

5.4. Applications in Anaphora Resolution and Event Tracking

Meyer and Dale [64] use corpora and the PWN hierarchy to derive axioms about associative anaphoric relations (an associative anaphora is a semantic relation existing between referents of definite expressions which are not coreferring, but are somehow related, as for instance in *bus... the driver*, or *okapi... giraffe's head*). Associative constructions of the form ⟨*NP of NP*⟩ and ⟨*genitive NP*⟩ are looked for in a corpus, then the found instances are generalized using the PWN hypernymic relations, and finally their possible antecedents are filtered by looking for their occurrence in contexts (two previous sentences) of the corpus.

Cristea et al. [17] and Cristea and Postolache [19] use a PWN-based semantic distance to compute matching scores of noun-to-noun coreference resolution. In all approaches results are expected to ameliorate with the use of a WSD component.

5.5. Applications Induced by Aligned Wordnets

Wordnets can be aligned in a mono-lingual and a cross-lingual way. Examples of alignment of wordnet with other lexical resources of the same language for English is with Roget's Thesaurus, as reported in [51], [63] and [77], where the idea is to disambiguate word senses in Roget's according to PWN. These authors prove that a correspondence between Roget's and PWN senses can be objectified and show that these two resources can be used together in a WSD process.

A cross-lingual (English, Spanish and Catalan) IR approach which has some similarities with the one described in section 4 of this paper is the one adopted in the ITEM search engine [128]. The processing chain in ITEM includes morphological analysis and POS-tagging followed by uni-language WSD. Different from our approach is that indexing and retrieval employ only EWN ILI record codes, therefore no domain tags are used.

Inversely, a multi-lingual aligned corpus, on which word alignment is also performed, can be used to import word sense tags from a corpus in one language to a corpus in another language (see approaches in [42], [43], [44], [21]).

Bilgin et al. [9] describe an approach of exporting paradigmatic semantic relations, in an aligned network of wordnets, from the wordnet of a language that displays these relations in its morphology.

Negri and Magnini [78] use the aligned English-Italian wordnets in MultiWordNet to extend to Italian a previously developed Named Entity Recognition (NER) system for English.

5.6. Other Applications of Wordnets

Fong [31] describes how wordnets can be employed to solve instances of the Frame Problem (in logic the Frame Problem tries to answer the question whether a statement that holds true of a certain given state continues to hold after some action is stated as having been performed). One interesting thing in this research is the way in which a semantic opposition is explained with respect to the network of synonymy/hypernymy/antonymy relations. Within these relations, the shortest paths are computed between the pair of confronted words and, if an antonymy relation is found within a given limit, the semantic opposition is signalled.

Magnini et al. [59] describe a method of NER that extensively uses the rich collection of instances of PWN. An additional advantage of a wordnet-based approach to NER is the growing existence of multi-language networks of wordnets, each contributing with their specific language collection of named entities.

Budanitsky and Hirst [13] apply and then compare five measures of semantic distance in PWN to a task aiming to correct spelling errors as malapropisms (unintentional misuse of words by confusion with others that sound similar).

A number of approaches use wordnets for lexicographic purposes. In [91] the hierarchical structure of PWN is exploited to create a working definition of systematic polysemy and extract polysemic patterns. By comparing different varieties of thesauri (among which Roget, PWN and EWN) Kilgariff and Yallop [49] discover that pairs of lexicographically close meanings are often found in different parts of the hierarchy. They explain this very interesting finding by the fact that close senses are related by polysemous relation which cut across the hierarchy.

5.7. Enriching of WordNet and Integration with Other Linguistic Thesauri

The richer the electronic linguistic resources, the more appealing their integration with other resources is. The opening to integration of any wordnet, and especially of aligned wordnets, is contributed by their lexical, ontological and semantic dimensions. Many approaches tend to exploit the common features of wordnets with other resources in an attempt to obtain a mapping that would allow for the acquisition of new dimensions.

In [3] is described an attempt to enrich PWN concepts with topic signatures. The research is motivated by the critiques that WN lacks relations between related concepts (for instance, there is no relation linking *farm* to *chicken*, or *fork* to *dinner*). A topic signature of a concept in such an enriched wordnet is a list of words topically related to the concept. The topic signatures are automatically collected from the Web and similarities between topic signatures could be used to cluster topically related concepts. However their usefulness in WSD was not proved to be relevant.

The key idea of Extended WordNet, as presented for the first time by Harabagiu et al. [35], was to exploit the part of the world knowledge contained in the PWN glosses, which are now used primarily by humans, to help the development of deep-semantics automatic inference mechanisms. Lexical tokens (single words or multi-word expressions) occurring within the glosses are firstly POS-tagged, then semantically tagged according to senses from WordNet, and finally glosses are transformed into logical forms. The increased connectivity between synsets thus obtained behaves like a Core Knowledge Base that can be used for various tasks, which include Question Answering, Information Retrieval, Information Extraction, Summarization, Natural Language Generation, and other knowledge intensive applications. In [68] a methodology is proposed to automatically POS-tag and sense-tag the words of the glosses by combining diverse taggers (methods) with different degrees of confidence. The interesting fact is that out of the eight sense tagging methods proposed, only one is dependent on the existence of a semantically tagged corpus (SemCor, in their case). All the others exploit only the relations existing within PWN and, therefore, are completely portable to wordnets of other languages. A method for the transformation of the PWN glosses into first-order logical forms is described in [72].

Vossen [127] describes a methodology of automatic extension and trimming of a multilingual wordnet for cross-lingual information retrieval on technical domains. The system builds a hierarchy using terms extracted from documents. This hierarchy is combined with the wordnet hierarchy and then trimmed with a disambiguation method.

A different trend is towards impoverishing WN of its too fine-grained senses, a feature which is often criticised. A coarse grained PWN is claimed for a large range of applications, of which WSD is only one. Mihalcea and Moldovan [67] present a methodology for reducing the polysemy of words in PWN. They propose a set of semantic and probabilistic rules that are used to either merge synsets very similar in meaning or delete synsets that are very rarely used. In other approaches [26], [90] wordnet structures like generalisations, cousins, sisters, twins, and auto-hyponymy are used to cluster senses, in the context of IR applications.

Collocations (lexical sequences frequently co-occurring in normal language use) are not usually treated as lexical entries. Guerreiro [28] proposes the inclusion in wordnets of collocational information associated with component words, as semantic restrictions of word association.

GermaNet (the German component of EWN), as described in [52], supplements the classical wordnet configuration with subcategorization frames for 7 000 verbs. But determining verb-argument structures is an ongoing work with an important impact on applications intended to elucidate the deep understanding mechanisms of human language. Recent research attempted to automatically detect verb-argument structures and acquire the syntactic alternation behaviour of verbs directly form large corpora and to compare them against human judgement of verb classes, as that of Levin [55]. Gildea [25] describes an approach for probabilistic evaluation of verb-argument structures, which applies also to alternations. McCarthy [62] identifies verbs participating in specific alternations and uses PWN to classify role fillers into semantic categories. Schulte im Walde [104] uses an expectation maximization algorithm to automatically cluster verbs, with the attempt to derive Levin classes from untagged data. As in McCarthy's approach, nouns are classified using PWN. In a more recent paper, Schulte im Walde [105] presents clustering experiments on German verbs (based on GermaNet), which explore the relevance of three levels of features: syntactic frame types, prepositional phrase information and selectional restrictions. Green et al. [27] and Korhonen [50] describe similar approaches to semi-automatic classification of verbs to Levin classes via the semantic network of PWN. Brockmann and Lapata [11] investigate five methods for automatic acquisition of selectional preferences, as GermaNet concepts, for German verbs that take into account direct objects and prepositional complements.

Automatic aligning between monolingual ontologies is an interesting path towards acquiring a wordnet ontology or for the completion of wordnets with other resources. Ngai et al. [83] report a procedure of aligning the Mandarin Chinese HowNet with PWN (they speak of an average of 8 to 1 correspondence between HowNet synsets and PWN synsets). The alignment used a greedy maximum forward match algorithm that was run on context vectors extracted from an English-Mandarin bilingual corpus. Chang et al. [15] describe a method to assign domain tags to PWN entries by exploiting the explicit domain information contained in the Far East Dictionary and the contextual information in PWN. Further, the similarity between common lexical taxonomy and the semantic hierarchy of PWN is examined in order to enlarge the domain labelling within the PWN hierarchy in a systematic way. On a similar trend, Buitelaar and Săcăleanu [12] assign domain tags to GermaNet synsets on the basis of the relevance between the synsets' constituent terms that co-occur within representative domain corpora.

A proposal to extend wordnets with *phrasets* (sets of free combinations of words which are recurrently used to express a concept), acquired from dictionaries and corpora in the framework of MultiWordNet (the Italian component of EWN [92]) is proposed in [8].

Palmer et al. [87] and Fellbaum et al. [29] augment the Penn Tree Bank annotation with PWN sense tags.

5.8. Critiques to the Wordnet Concept in Linguistics

There exist voices that deny the concept of wordnet as adequate for recording the linguistic knowledge of a language. The main critiques contest: *a*) wordnet's ability to describe the continuum nature of senses, therefore the very nature of the linguistic concept or meaning; *b*) sometimes its too refined sense distinction[10]; *c*) the adequacy of synsets to describe the synonymy relation ("no words are perfect synonyms" they say, and the evidence is that one cannot replace one word with another in its synset without loosing or adding some fine details); *d*) its still scarce set of semantic relations; and *e*) the lack of words' argument structures.

To give an example related to topic *a*, the contesters say that senses are an invention of our minds because, in reality, meaning are more like continuums than like discrete entities. These linguists claim that concepts are variable and may be better described in terms of features or prototypes[11]. For instance, how close to a prototypical vegetable, such as a pea, does another vegetable come before it really is a vegetable in people's minds?[12] These approaches are all valid and reasonable, and there is some experimental evidence for them. However, they are not necessarily in conflict with PWN. PWN simply takes the words of a language and arranges them according to psychologically valid principles. PWN takes as input what the language reflects about conceptualization. If people think that broccoli is a "better" vegetable than a tomato, that may well be true, but the language does not reflect it, and there is little we can do with it other than observe it. We do not have a word for "80% vegetable" or "bad representative of the concept of vegetable." We cannot foresee a psycho-linguistically motivated and affordable approach which would build a seman-

[10] See [18] in this volume for a discussion on the variability of senses and the difficulty of recognizing the border between senses.

[11] For alternative approaches see Conceptual Dependency frames of the Yale school [103], Latent Semantic Indexing of Landauer and Littman [53], The Generative Lexicon [94], [95], Microkosmos [81], [82], Sensus [41], Acquilex [16], FrameNet [56], Lexical Conceptual Structures [23] and The Integral Dictionary [24].

[12] The "vegetable" argument was given, in a discussion with one of the authors, by Christiane Fellbaum (included here with her kind acceptance).

tic net with many different degrees of strength between, e.g. hypo- and hypernyms. And, anyway, how could we measure, psychologically safe, these strengths for all the words in a language. And, supposing a solution to this exists, what would be the evidence that all speakers of that language perceive the same relation with the same degree of strength?

The critiques encumbered with "*the too fine sense distinction of PWN*" argument are answered in a very simple way: one could keep the very fine-grained sense distinction of PWN where there is a need for it, and one can use a coarser grained set of senses in all other cases (see [67] for a method to throw away many synsets or to merge the synsets of a wordnet).

To all the other critiques the response is that the wordnets, as the concept is understood today, represent updatable resources (as the richness of the approaches that try to combine PWN with other resources, presented in the previous sections can prove). It is clear that PWN and all other language specific wordnets will tremendously gain from the integration with different linguistic resources, and methods that convincingly prove the feasibility of such attempts are numerous, even if the automatic alignment is challenged.

Despite all these critiques, the PWN has reached today an extraordinary popularity. First, there is experimental evidence for the psychological reality of the wordnet concept, the way people organize meanings and store knowledge about them. There is strong evidence in favor of mental representations, in favor of the fact that they really exist in people's minds, even separate from lexicons, which only name these meanings. And the wordnets records only those mental representations that have names in one language. Then, at least for English, we know of no similarly large implementation for a lexicon and displaying such a rich lexico-semantic structure. Finally, there is a growing interest nowadays towards building similar resources for other languages and aligning them on inter-lingual concepts.

6. Conclusions

The present-day success of the PWN and the proliferation of instantiations of the wordnet concept in many languages should be attributed not only to its rich organisation but also to the fact that PWN is a comprehensive knowledge source. In a domain where the difficulty to acquire exhaustive and reliable electronic linguistic thesauri is notorious, its accuracy and completeness makes it appealing for so many approaches. This characteristic is of a tremendous importance and should be maintained in the community of aligned wordnets. Without satisfactory coverage and correctness, the individual language wordnets will not exhibit the same popularity and the effort to develop them up to the present level will be under-exploited. From this point of view, the BalkaNet project boosted the development of five new wordnets and the significant extension of a sixth one. All the six wordnets have a significant cross-lingual coverage via PWN. By adopting the EWN methodology, the BalkaNet wordnets are extending the pool of aligned wordnets to an unprecedented semantic network for 15 European languages. Although larger than specified in the project's technical annex, the BalkaNet wordnets are still prototypes, much smaller than PWN. Much effort and funds should be invested in order to bring them at the PWN level of lexical coverage. Yet, due to the harmonized criteria in ILI concepts, these wordnets are immediately usable in various applications.

Balkanet is important also from a different point of view. It has developed tools that are freely available and has refined the methodology for building wordnets to a level that makes much easier and more accurate the integration of new language wordnets within the envisaged global network. Any new wordnet linked to the current network adds value to each of the existing wordnets. To give just an example, name entities in individual wordnets, representing language specific contribution, add value to all aligned wordnets for many NLP and AI applications. It is also important to notice that, although originally PWN was designed as an electronic resource intended for manual consultation, the interest shifted more and more towards automatic accessing. In this respect EuroWordNet and Balkanet provide software interfaces for plugging-in this important resource onto diverse applications.

The upshot is that although the wordnet model is certainly not a perfect way to represent human lexical knowledge, we know of no better one. Moreover its wide acceptance and the proliferation of wordnets for so many languages, to which Balkanet adds five more languages, makes it a tremendously important international resource, the greatest ever build.

Acknowledgement. The work reported in this paper is conducted in the framework of the BalkaNet project (IST-2000-29388) supported by the European Commission. We would like to thank all members of the BalkaNet consortium for their significant contribution in the objectives of the project, namely: Databases Laboratory of Patras University, Greece; Research Academic Computer Technology Institute, Greece; Sabanci University, Turkey; Masaryk University, Czech Republic; Athens University, Greece; MEMODATA, France; Plovdiv University, Bulgaria; "Alexandru Ioan Cuza" University of Iaşi, Romania; Research Institute for Artificial Intelligence of the Romanian Academy, Romania; Bulgarian Academy of Sciences, Bulgaria; Belgrade University, Serbia and OTEnet Internet Service Provider, Greece.

Many thanks go also to Dr. Christiane Fellbaum, the coordinator of the PWN project, and to Dr. Piek Vossen, the former coordinator of the EWN project for their valuable support and suggestions throughout the implementation of BalkaNet. Finally, we would like to thank the Project's Officer, Mr. Erwin Valentini and BalkaNet's mid-term reviewers, Dr. Frederique Segond and Dr. Petek Bojan, for their support.

References

[1] AGIRRE, E., RIGAU, G., *Word Sense Disambiguation Using Conceptual Density*, in *Proceedings of the COLING Conference*, Copenhagen, Denmark, 1996.

[2] AGIRRE, E., ANSA, O., HOVY, E., MARTINEZ, D., *Enriching Very Large Ontologies Using the WWW*, in *Proceedings of the ECAI Workshop on Ontology Learning*, Berlin, Germany, 2000.

[3] AGIRRE, E., ANSA, O., MARTINEZ, D., HOVY, E., *Enriching WordNet Concepts with Topic Signatures*, in *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, Pennsylvania, 2001.

[4] AMBROZIAK, J., WOODS, W., *Natural Language Technology in Precision Content Retrieval*, in *International Conference on Natural Language Processing and Industrial Applications*, Moncton, New Brunswick, Canada, 1998.

[5] BALKOVA, V., SUKHONOGOV, A., YABLONSKY, S., *Russian Wordnet: from UML to Internet/Intranet Implementation*, in *Proceedings of the 2^{nd} International Global Wordnet Conference*, 31–38, Brno, Czech Republic, 2004.

[6] BASILI, R., PAZIENZA, M.T., VINDIGNI, M., *Corpus-driven learning of event recognition rules*, in *Proceedings of Machine Learning for Information Extraction workshop, held jointly with the ECAI 2000*, Berlin, Germany, 2000.

[7] BASILI, R., PAZIENZA, M.T., ZANZOTTO, F., CATIZONE, R., SETZER, A., WEBB, N., WILKS, Y., PADRO, L., RIGAU, G., *Multilingual Authoring: the NAMIC approach*, in *Proceedings of the ACL-EACL 2001*, Toulouse, France, 2001.

[8] BENTIVOGLI, L., PIANTA, E., *Beyond Lexical Units: Enriching Wordnets with Phrasets*, in *Proceedings of EACL-2003*, Budapest, Hungary, 2003.

[9] BILGIN, O., ÇETINOĞLU, Ö., OFLAZER, K., *Morphosemantic Relations In and Across Wordnets*, in *Proceedings of the Global Wordnet Conference*, Brno, Czech Republic, 2004.

[10] BKN-D.4.2, Tracing the common base concepts, Balkanet Deliverable D.4.1, WP 4, February, 2003.

[11] BROCKMANN, C., LAPATA, M., *Evaluating and Combining Approaches to Selectional Preference Acquisition*, in *Proceedings of EACL-2003*, Budapest, Hungary, 2003.

[12] BUITELLAR, P., SĂCĂLEANU, B., *Ranking and Selecting Synsets by Domain Relevance*, in *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, Pennsylvania, 2001.

[13] BUDANITSKY, A., HIRST, G., *Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures*, in *Proceedings of the Workshop on WordNet and Other Lexical Resources*, Second meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, Pennsylvania, 2001.

[14] CHAI, J.Y., *Evaluation of a Generic Lexical Semantic Resource in Information Extraction*, in *Proceedings of LREC-2000*, Athens, Greece, 2000.

[15] CHANG, E., HUANG, C.-R., KER, S.-J., YANG, C.-H., *Induction of Classification from Lexicon Expansion: Assigning Domain Tags to WordNet Entries*, in *Proceedings of COLING-2002*, Taipei, Taiwan, 2002.

[16] COPESTAKE, A., SANFILIPPO, A., *Multilingual lexical representation*, in *Proceedings of the AAAI Spring Symposium: Building Lexicons for Machine Translation*, Stanford, California, 1993.

[17] CRISTEA, D., DIMA, G.E., POSTOLACHE, O.D., MITKOV, R., *Handling complex anaphora resolution cases*, in *Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquium*, Lisbon, Portugal, 2002.

[18] CRISTEA, D., MIHĂILĂ, C., FORĂSCU, C., TRANDABĂŢ, D., HUSARCIUC, M., HAJA, G., POSTOLACHE, O., *Mapping Princeton WordNet synsets onto Romanian wordnet synsets*, in this volume, 2004.

[19] CRISTEA, D., POSTOLACHE, O.D., *How to deal with wicked anaphora*, to appear in António Branco, Tony McEnery and Ruslan Mitkov (edts): *Anaphora Processing: Linguistic, Cognitive and Computational Modelling*, Benjamin Publishing Books, 2004.

[20] DAUDÉ, J., PADRÓ, L., RIGAU, G., *Mapping WordNets Using Structural Information*, in *Proceedings of ACL-2000*, Hong Kong, China, 2000.

[21] DIAB, M., *An Unsupervised Method for Multilingual Word Sense Tagging Using Parallel Corpora: A Preliminary Investigation*, in *Proceedings of the SIGLEX Workshop on Word Senses and Multilinguality*, in conjunction with ACL-2000, Hong Kong, China, 2000.

[22] DIMITROVA, L., ERJAVEC, T., IDE, N., KAALEP, H.J., PETKEVIC, V., TUFIŞ, D., *Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages*, COLING, Montreal, Canada, 1998.

[23] DORR, B.J., *Large-scale dictionary construction for foreign language tutoring and interlingual machine translation*, *Machine Translation*, **12**, 1997.

[24] DUTOIT, D., *A Text → Meaning → Text Dictionary and Process*, in *Proceedings of LREC-2000*, Athens, Greece, 2000.

[25] GILDEA, D., *Probabilistic Models of Verb-Argument Structure*, in *Proceedings of COLING-2002*, Taipei, Taiwan, 2002.

[26] GONZALO, J., CHUGUR, I., VERDEJO, F., *Sense Clusters for Information Retrieval: Evidence from SemCor and the EuroWordNet InterLingual Index*, in *Proceedings of the SIGLEX Workshop on Word Senses and Multilinguality*, in conjunction with ACL-2000, Hong Kong, China, 2000.

[27] GREEN, R., PEARL, L., DORR, B.J., RESNIK, P., *Mapping Lexical Entries in a Verbs Database to WordNet Senses*, in *Proceeding of ACL-EACL-2001*, Toulouse, France, 2001.

[28] GUERREIRO, P., *Improving Lexical Databases with Collocational Information: data from Portuguese*, in *Proceedings of LREC-2000*, Athens, Greece, 2000.

[29] FELLBAUM, C., PALMER, M., DANG, H.T., DELFS, L., WOLF, S., *Manual and Automatic Semantic Annotation with WordNet*, invited paper, in *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, Pennsylvania, 2001.

[30] FERRET, O., GRAU, B., HURAULT-PLANTET, M., ILLOUZ, G., JACQUEMIN, C., *Terminological variants for document selection and question/answer matching*, in *Proceedings of the ACL-EACL Workshop on Open-Domain Question Answering*, Toulouse, France, 2001.

[31] FONG, S., *On Mending a Torn Dress: The Frame Problem and WordNet*, in *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, Pennsylvania, 2001.

[32] HARABAGIU, S.M., MAIORANO, S.J., *Acquisition of Linguistic Patterns for Knowledge-Based Information Extraction*, in *Proceedings of LREC-2000*, Athens, Greece, 2000.

[33] HARABAGIU, S.M., MAIORANO, S.J., *Multi-document summarization with GISTexter*, in *Proceedings of LREC-2002*, Las Palmas, Spain, 2002.

[34] HARABAGIU, S., MOLDOVAN, D., PAŞCA, M., MIHALCEA, R., SURDEANU M., BUNESCU, R., GÎNJU, R., RUS, V., MORĂRESCU, P., *The Role of Lexico-Semantic Feedback in Open-Domain Textual Question-Answering*, in *Proceedings of ACL-EACL*, Toulouse, France, 2001.

[35] HARABAGIU, S., MILLER, G., MOLDOVAN, D., *WordNet 2 – a morphologically and semantically enhanced resource*, in *Proceedings of SIGLEX-99*, University of Maryland, 1999.

[36] HARABAGIU, S., PAŞCA, M.A., MAIORANO, S.J., *Experiments with Open-Domain Textual Question Answering*, in *Proceedings of COLING-2000*, Taipei, Taiwan, 2000.

[37] HORÁK, A., SMRŽ, P., *New Features of Wordnet Editor VisDic*, in this volume, 2004.

[38] HIRST, G., *Ontology and the Lexicon*, 2003.

[39] HIRST, G., ST-ONGE, D., *Lexical chains as representations of context for the detection and correction of malapropisms*, in Fellbaum C. (ed.), *WordNet: An Electronic Lexical Database*, The MIT Press, Cambridge, MA, 1998.

[40] HOVY, E., NIRENBURG, S., *Approximating an Interlingua in a Principled Way*, in *Proceedings of the DARPA Speech and Natural Language Workshop*, Arden House, New York, 1992.

[41] HOVY, E., *Ontologies for Machine Translation*, presented at *The Japan-US MT Workshop*, Washington, D.C., 1993.

[42] IDE, N., *Cross-lingual sense determination. Can it work?*, in *Computers and Humanities*, **34**, 1999.

[43] IDE, N., ERJAVEC, T., TUFIŞ, D., *Automatic Sense Tagging Using Parallel Corpora*, in *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, Tokyo, Japan, 2001.

[44] IDE, N., ERJAVEC, T., TUFIŞ, D., *Sense Discrimination with Parallel Corpora*, in *Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, Pennsylvania, 2002.

[45] ION, R., TUFIŞ, D., *Multilingual Word Sense Disambiguation Using Aligned Wordnets*, in this volume, 2004.

[46] KOEVA, S., MIHOV, S., TINCHEV, T., *Bulgarian Wordnet – Structure and Validation*, in this volume, 2004.

[47] KOUTSOUBOS, I.D., ANDRIKOPOULOS, V., CHRISTODOULAKIS, D., *Wordnet Exploitation through a Distributed Network of Servers*, in *Proceedings of the 2nd International Global Wordnet Conference*, Brno, Czech Republic, 2004.

[48] KOUTSOUBOS, I.D., ANDRIKOPOULOS, V., TZAGARAKIS, M., CHRISTODOULAKIS, D., WMS: *Towards a distributed network of Semantic Networks*, in this volume, 2004.

[49] KILGARIFF, A., YALLOP, C., *What's in a thesaurus?*, in *Proceedings of LREC-2000*, Athens, Greece, 2000.

[50] KORHONEN, A., *Assigning Verbs to Semantic Classes via Wordnet*, in *Proceedings of COLING-2002 SemaNet workshop on Building and Using Semantic Networks*, Taipei, Taiwan, 2002.

[51] KWONG, O.I., *Word Sense Disambiguation with an Integrated Lexical Resource*, in *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, Pennsylvania, 2001.

[52] KUNZE, C., *Extension and use of GermaNet, a lexical-semantic database*, in *Proceedings of LREC-2000*, Athens, Greece, 2000.

[53] LANDAUER, T., LITTMAN, M.L., *Fully automatic cross-language document retrieval using latent semantic indexing*, in *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, Waterloo Ontario, 1990.

[54] LENAT, D.B., *CYC: A Large-Scale Investment in Knowledge Infrastructure*, Communications of the ACM, **38**, no.11, 1995.

[55] LEVIN, B., *English Verb Classes and Alternations: A Preliminary Investigation*, University of Chicago Press, Chicago, 1993.

[56] LOWE, J.B., BAKER, C.F., FILLMORE, C.J., *A frame-semantic approach to semantic annotation*, in *Proceedings of SIGLEX Workshop ANLP'97*, Washington D.C., 1997.

[57] MAGNINI, B., PREVETE, R., *Exploiting Lexical Expansions and Boolean Compositions for Web Querying*, in *Proceedings of the ACL-2000 Workshop on Recent Advances in NLP and IR*, Hong Kong, China, 2000.

[58] MAGNINI, B., CAVAGLIÀ, G., *Integrating Subject Field Codes into WordNet*, in *Proceedings of LREC-2000*, Athens, Greece, 2000.

[59] MAGNINI, B., NEGRI, M., PREVETE, R., TANEV, H., *A Wordnet-based Approach to Named-Entites Recognition*, in *Proceedings of COLING-2002 SemaNet workshop on Building and Using Semantic Networks*, Taipei, Taiwan, 2002.

[60] MAGNINI, B., STRAPPARAVA, C., *Experiments in Word Domain Disambiguation for Parallel Texts*, in *Proceedings of the SIGLEX Workshop on Word Senses and Multilinguality*, in conjunction with ACL-2000, Hong Kong, China, 2000.

[61] MANN, G.S., *A Statistical Method for Short Answer Extraction*, in *Proceedings of the ACL-EACL Workshop on Open-Domain Question Answering*, Toulouse, France, 2001.

[62] McCARTHY, D., *Using Semantic Preferences to Identify Verbal Participation in Role Switching Alternations*, in *Proceedings of NAACL*, Seattle, WA, 2000.

[63] McHALE, M.A., *A Comparison of WordNet and Roget's Taxonomy for Measuring Semantic Similarity*, in *COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, 1998.

[64] MEYER, J., DALE, R., *Using the Wordnet Hierarchy for Associative Anaphora Resolution*, in *Proceedings of COLING-2002 SemaNet workshop on Building and Using Semantic Networks*, Taipei, Taiwan, 2002.

[65] MIHALCEA, R., MOLDOVAN, D.I., *Semantic indexing using WordNet Senses*, in *Proceedings of the Workshop on Recent Advances on NLP and IR, ACL-2000*, Hong Kong, China, 2000.

[66] MIHALCEA, R., MOLDOVAN, D.I., *Automatic Generation of a Coarse Grained WordNet*, in *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, Pennsylvania, 2001.

[67] MIHALCEA, R., MOLDOVAN, D.I., *EZ.Wordnet: principles for automatic generation of a coarse grained WordNet*, in *Proceedings of FLAIRS 2001*, Key West, Florida, 2001.

[68] MIHALCEA, R., MOLDOVAN, D.I., *Extended WordNet: Progress Report*, in *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, Pennsylvania, 2001.

[69] MILLER, G., *Five Papers on WordNet*, Special Issue in International Journal of Lexicography, **3**, no. 4, 1990.

[70] MILLER, G.A., BECKWIDTH, R., FELLBAUM, C., GROSS, D., MILLER, K.J., *Introduction to WordNet: An On-Line Lexical Database*, International Journal of Lexicography, **3**, no. 4 (winter 1990), 235–244, 1990.

[71] MOLDOVAN, D., HARABAGIU, S., T*he Structure and Performance of an Open-Domain Question Answering System*, in *Proceedings of ACL 2000*, Hong Kong, China, 2000.

[72] MOLDOVAN, D.I., RUS, V., *Logic Form Transformation of WordNet and its Applicability to Question Answering*, in *Proceedings of ACL-EACL 2001*, Toulouse, France, 2001.

[73] MOLDOVAN, D., NOVISCHI, A., *Lexical Chains for Question Answering*, in *Proceedings of COLING 2002*, Taipei, Taiwan, 2002.

[74] MOLINA, A., PLA, F., SEGARRA, E., MORENO, L., *Word Sense Disambiguation using Statistical Models and WordNet*, in *Proceedings of LREC-2002*, Las Palmas, Spain, 2002.

[75] MORATO, J., MARZAL, M. Á., LLORÉNS, J., MOREIRO, J., *WordNet Applications*, in *Proceedings of GWC-2004*, Brno, Czech Republic, 2004.

[76] NARAYAN, D., CHAKRABARTY, D., PANDE, P., BHATTACHARYA, P., *An Experience in Building the Indo-Wordnet – a Wordnet for Hindi*, in *Proceedings of the 1st International Global Wordnet Conference*, Mysore, India, 2002.

[77] NĂSTASE, V., SZPAKOWICZ, S., *Word Sense Disambiguation in Roget's Thesaurus Using WordNet*, in *Proceedings of NAACL-2001*, Pittsburgh, PA, 2001.

[78] NEGRI, M., MAGNINI, B., *Using WordNet Predicates for Multilingual Named Entity Recognition*, in *Proceedings of the GWC-2004*, Brno, Czech Republic, 2004.

[79] NILES, I., PEASE, A., *Towards a Standard Upper Ontology*, in Chris Welty and Barry Smith (Eds.), *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems* (FOIS), Ogunquit, Maine, October, 2001.

[80] NILES, I., PEASE, A., *Origins of the IEEE Standard Upper Ontology*, Working Notes of the IJCAI Workshop on the IEEE Standard Upper Ontology, Seattle, 37–42, 2001.

[81] NIRENBURG, S., CARBONELL, J., TOMITA, M., GOODMAN, K., *Machine translation: a knowledge-based approach*, Morgan Kaufmann, San Mateo, California, 1992.

[82] NIRENBURG, S., BEALE, S., MAHESH, K., ONYSHKEVYCH, B., RASKIN, V., VIEGAS, E., WILKS, Y., ZAJAC, R., *Lexicons in the MicroKosmos project*, in Lynne Cahill and Roger Evans (edts.), *Proc. AISB Workshop on Multilinguality in the Lexicon*, Brighton, England, April 1996.

[83] NGAI, G., CARPUAT, M., FUNG, P., *Identifying Concepts Across Languages: A First Step towards a Corpus-based Approach to Automatic Ontology Alignment*, in *Proceedings of COLING 2002*, Taipei, Taiwan, 2002.

[84] PALMER, M., DANG, H.T., ROSENZWEIG, J., *Semantic Tagging for the Penn Treebank*, in *Proceedings of LREC-2000*, Athens, Greece, 2000.

[85] PAŞCA, M., HARABAGIU, S.M., *Answer Mining from On-Line Documents*, in *Proceedings of the ACL-EACL Workshop on Open-Domain Question Answering*, Toulouse, France, 2001.

[86] PAŞCA, M., HARABAGIU, S., *The Informative Role of WordNet in Open-Domain Question Answering*, in *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, Pennsylvania, 2001.

[87] PALMER, M., DANG, H., FELLBAUM, C., *Making Fine-grained and Coarse-grained sense distinctions, both manually and automatically*, in *Journal of Natural Language Engineering*, 2004 (to appear).

[88] PAVELEK, T., PALA, K., *VisDic: A new Tool for WordNet Editing*, in *Proceedings of the 1st International Global Wordnet Conference*, Mysore, India, 2002.

[89] PETERS, W., VOSSEN, P., DIEZ-ORZAS, P., ADRIAENS, G., *Cross-Linguistic Alignment of wordnets with an Inter-Lingual-Index*, in N. Ide, D., Greenstein, P. Vossen (Eds.), *Special Issue on EuroWordNet*, Computers and the Hummanities, **32**, nos. 2–3, 221–251, 1998.

[90] PETERS, W., PETERS, I., *Automatic sense clustering in EuroWordNet*, in *Proceedings of LREC-2000*, Athens, Greece, 2000.

[91] PETERS, W., PETERS, I., *Lexicalised Systematic Polysemy in WordNet*, in *Proceedings of LREC-2000*, Athens, Greece, 2000.

[92] PIANTA, E., BENTIVOGLI, L., GIRARDI, C., *MultiWordNet: developing an aligned multilingual database*, in *Proceedings of the First International Conference on Global WordNet*, Mysore, India, 2002.

[93] * * *, *Princeton WordNet 2.0*, `ftp.cogsci.princeton.edu`

[94] PUSTEJOVSKY, J., *The Generative Lexicon*, Computational Linguistics, **17**(4), 1991.

[95] PUSTEJOVSKY, J., *The Generative Lexicon*, MIT Press, Cambridge, MA, 1995.

[96] QUILLIAN, R., *Semantic Memory*, in M. Minsky (Ed.), *Semantic Information Processing*, MIT Press, Cambridge, M.A., 216–270, 1968.

[97] QUINE, WILLARD VAN ORMAN, *Word and Object*, The MIT Press, Cambridge, MA, 1960.

[98] RADA, R., MILI, H., BICKNELL, E., BLETTNER, M., *Development and Application of a Metric on Semantic Nets*, IEEE Transactions on Systems, Man and Cybernetics, **19**(1), 17–30, 1989.

[99] RESNIK, P., *Disambiguating Noun Groupings with Respect to WordNet Senses*, in *Proceedings of the 3rd Workshop on Very Large Corpora*, MIT, 1995.

[100] RODRIGUEZ, H., CLIMENT, S., VOSSEN, P., BLOKSMA, L., PETERS, W., ALONGE, A., BERTAGNA, F., ROVENTINI, A., *The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology*, in Piek Vossen (ed.), *EuroWordNet: A Multilingual database with lexical semantic networks*, Computers and Humanities, **32**, nos. 2–3, 1998.

[101] SALTON, G., McGILL, M.J., *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

[102] SALTON, G., BUCKLEY, C., *Term Weighting Approaches in Automatic Text Retrieval*, Information Processing and Management, **24**(5), 513–523, 1988.

[103] SCHANK, R., *Conceptual dependency: A theory of natural language understanding*, Cognitive Psychology, **3**(4), 1972.

[104] SCHULTE IM WALDE, S., *Clustering verbs semantically according to their alternation behaviour*, in *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, Saarbrücken, Germany, 2000.

[105] SCHULTE IM WALDE, S., *Experiments on the Choice of Features for Learning Verb Classes*, in *Proceedings of EACL-2003*, Budapest, Hungary, 2003.

[106] SMRŽ, P., *Quality Control for Wordnet Development*, in *Proceedings of the 2nd International Wordnet Conference*, Brno, Czech Republic, 2004.

[107] SMRŽ, P., *Quality Control and Checking for Wordnet Development*, in this volume, 2004.

[108] STAMOU, S., NTOULAS, A., HOPPENBROUWERS, J., SAIZ-NOEDA, M., CHRISTODOULAKIS, D., *EUROTERM: Extending the EuroWordNet with Domain-Specific Terminology Using an Expand Model Approach*, in *Proceedings of the 1st International Global Wordnet Conference*, Mysore, India, 2002.

[109] STAMOU, S., OFLAZER, K., PALA, K., CHRISTODOULAKIS, D., CRISTEA, D., TUFIŞ, D., KOEVA, S., TOTKOV, G., DUTOIT, D., GRIGORIADOU, M., *BALKANET: A Multilingual Semantic network for the Balkan Languages*, in *Proceedings of the 1st International Global Wordnet Conference*, Mysore, India, 2002.

[110] STAMOU, S., NENADIC, G., CHRISTODOULAKIS, D., *Exploring BalkaNet Shared Ontology towards Multilingual Conceptual Indexing*, to appear in *Proceedings of the 4th Language Resources and Evaluation Conference* (LREC), Lisbon, Portugal, 2004.

[111] * * *, *SUMO*, `http://ontology.teknowledge.com/#FOIS`

[112] TUFIŞ, D., CRISTEA, D., *IURES: A Human Engineering Approach to Natural Language Question Answering*, in W. Bibel, B. Petkoff (eds.), *Artificial Intelligence: Systems, Applications, Methodology*, North Holland, 1985.

[113] TUFIŞ, D., CRISTEA, D., *A Pragmatic Implementation of the Dialogue Principles*, in *Automatische Sprache Verarbeitung*, J. Kunze (ed.), Zentral Institut fur Linguistics Berlin, Germany, 1985.

[114] TUFIŞ, D., CRISTEA, D., *Methodological in Building the Romanian Wordnet and Consistency Checks in BalkaNet*, in *Proceedings of the LREC Special Workshop on wordnets*, 35–41, Las Palmas, Spain, 2002.

[115] TUFIŞ, D., BARBU, A.M., ION, R., *TREQ-AL: A word-alignment system with limited language resources*, *Proceedings of the NAACL 2003 Workshop on Building and Using Parallel Texts; Romanian-English Shared Task*, 36–39, Edmonton, Canada, 2003.

[116] TUFIŞ, D., POPESCU, C., ROŞU, R., *Automatic classification of documents by random sampling*, Proceeding of the Romanian Academy, Series A, **1**, no. 2, 18–28, 2000.

[117] TUFIŞ, D., ION, R., *Word Sense Clustering Based on Translation Equivalence in Parallel Texts; A Case Study in Romanian*, in Corneliu Burileanu (ed), *Speech and Dialogue Systems*, 8–22, Romanian Academy Publishing House, Bucharest, 2003.

[118] TUFIŞ, D., BARBU, E., ION, R., MITITELU, V., *Cross-lingual Validations of Wordnets*, in *Proceedings of the 2nd International Wordnet Conference*, pp. 332-340, Brno, Czech Republic, 2004.

[119] TUFIŞ, D., ION, R., IDE, N., *Word Sense Disambiguation as a Wordnets' Validation Method in BalkaNet*, to appear in *Proceedings of the 4^{th} Language Resources and Evaluation Conference* (LREC), Lisbon, Portugal, 2004.

[120] YOKOI, T., *The EDR Electronic Dictionary*, Communications of the ACM, **38**, no. 11, 1995.

[121] * * *, *VisDic*, `http://nlp.fi.muni.cz/projekty/visdic/`

[122] VOSSEN, P., *Right or Wrong: Combining Lexical resources in the EuroWordNet Project*, in M. Gellerstam, J. Jarborg, S. Malmgren, K. Noren, L., Rogstrom, C.R. Papmehl (Eds.), *Proceedings of the Euralex Workshop*, 715–128, Göteborg, Sweden, 1996.

[123] VOSSEN, P., BLOKSMA, L., RODRIGUEZ, H., CLIMENT, S., CALZOLARI N., ROVENTINI, A., BERTAGNA, F., ALONGE, A., PETERS, W., *The EuroWordNet Base Concepts and Top Ontology*, LE-4003, Deliverable D017, D034, D036, University of Amsterdam, 1997.

[124] VOSSEN, P., DIEZ-ORZAS, P., PETERS, W., *The Multilingual Design of EuroWordNet*, in P. Vossen, Calzorari N., Adriaens G., Sanfilippo A., Wilks Y. (Eds.), *Proceedings of the ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid, Spain, 1997.

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

[126] VOSSEN, P., PETERS, W., GONZALO, J., *Towards a Universal Index of Meaning*, in *Proceedings of the ACL-99 SIGLEX Workshop*, University of Maryland, USA, 1999.

[127] VOSSEN, P., *Extending, Trimming and Fusing WordNet for Technical Documents*, in *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, Pennsylvania, 2001.

[128] VERDEJO, F., GONZALO, J., PENAS, A., LOPEZ, F., FERNANDEZ, D., *Evaluating wordnets in Cross-Language Information Retrieval: the ITEM search engine*, in *Proceedings of LREC-2000*, Athens, Greece, 2000.

[129] WOODS, W., *Conceptual Indexing: A Better Way to Organize Knowledge*, Technical Report TR-9761, Sun Microsystems Laboratories, Mountain View, CA, 1997.

[130] ZAJAC, R., *Towards Ontological Question Answering*, in *Proceedings of the ACL-EACL Workshop on Open-Domain Question Answering*, Toulouse, France, 2001.