COGNITIVE STUDIES | ÉTUDES COGNITIVES, 10
SOW Publishing House, Warsaw 2010

Part 2

Around the Problems of Language Corpora and Electronic Dictionaries in Slavic Languages and Its Cognitive Aspects

COGNITIVE STUDIES | ÉTUDES COGNITIVES, 10
SOW Publishing House, Warsaw 2010

LUDMILA DIMITROVA[1] **VIOLETTA KOSESKA-TOSZEWA**[2]
RADOVAN GARABÍK[3] **TOMAŽ ERJAVEC**[4]
LEONID IOMDIN[5] **VOLODYMYR SHYROKOV**[6]

[1]Institute of Mathematics and Informatics, Sofia, Bulgaria
[2]Institute of Slavic Studies, Warsaw, Poland
[3]Ľ. Štúr Institute of Linguistics, Bratislava, Slovakia
[4]Jožef Stefan Institute, Ljubljana, Slovenia
[5]Institute for Information Transmission Problems, Moscow, Russia
[6]Ukrainian Lingua-Information Fund, Kiev, Ukraine

ludmila@cc.bas.bg amaz1312@gmail.com
garabik@kassiopeia.juls.savba.sk tomaz.erjavec@ijs.si
iomdin@iitp.ru vshirokov48@mail.ru

MONDILEX — TOWARDS THE RESEARCH INFRASTRUCTURE FOR DIGITAL RESOURCES IN SLAVIC LEXICOGRAPHY

Abstract. The paper presents activities of the EU project MONDILEX, *Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and their Digital Resources*. The main objective of MONDILEX is to design the conceptual scheme of a research infrastructure that supports the networking of centres for high-quality research in traditional and digital Slavic lexicography.

1 Introduction

MONDILEX (Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and their Digital Resources) is a two-year (2008–2010) EU funded project[1], which aims to design an extensible infrastructure for institutions involved in creating and supporting a network of multilanguage resources of Slavic languages. The main goal of the MONDILEX project is to design the conceptual scheme for a research infrastructure supporting the networking of centres for research in Slavic lexicography.

The need for such an infrastructure arises from the disparity between the importance of the Slavic languages, spoken by a large part of Europe's population, and the insufficient number and quality of digital lexical resources for these languages.

[1] The project MONDILEX has received funding from the EC's Seventh Framework Programme [FP7/2007-2013] under Grant Agreement 211938.

The project provides strategies for the coordination, integration and extension of existing digital lexical resources and the creation of new ones in accordance with the recent advances in the field and international standards. At the same time, the project provides a venue for networking activities, such as joint management and pooling of resources, implementation of standards for products of digital lexicography, and coordination with relevant international standards and practices. Unified strategies should contribute to reusability and interoperability of such resources so that researchers in the humanities and social sciences as well as business communities could have easy access to bilingual and multilingual dictionaries of Slavic languages. In this way, the project contributes to the preservation and support of Europe's multilingual and multicultural heritage. In addition to lexical resources, MONDILEX also addresses the construction and maintenance of large deeply annotated text corpora, which are gaining in importance as the basis for linguistic research and technologies.

The partners in the project are research organisations from six European countries — four EU members (Bulgaria, Poland, Slovakia, Slovenia), as well as Russia and Ukraine, whose national languages belong to the Slavic group. All partners are national centres for research in linguistics, lexicography, and natural language processing. The partners' experience and methodology in compiling lexical databases are rather different. Some of the partners are primarily oriented toward digital and application-oriented dictionaries and corpora; others develop extensive grammar descriptions, whilst the Ukrainian partner is the main team in its country that develops traditional lexicographic resources using high-end computer technology. Still, all partners have experience in compiling dictionaries or lexical databases. Despite the diversity of design methods and techniques used in these resources, cooperation under the MONDILEX project has revealed that the underlying structure of the dictionaries and overall workflow methodology share many common features. Even more importantly, this cooperation enabled the partners to identify common points of salience in the dictionaries, which lays down an excellent basis for further networking of already existing lexical databases, and for a common lexicographic research project. During the first phase of MONDILEX, we have identified areas where we can deploy a unified scheme to interconnect the databases.

2 Main Tasks and their Implementation

The multidisciplinary character of the MONDILEX project consists in uniting the effort of the linguistic and ICT communities, applying up-to-date techniques of processing dictionary systems (information theory of lexicographic systems) and using state-of-the-art network technologies for information exchange between task groups. The project will lay the foundations for further cooperation, set up and elaborate a methodology of interaction of remote research groups and coordination of formats of lexicographic resources. The work program consists of the following major tasks:

– To examine the state of the art in monolingual, bilingual, and multilingual Slavic digital lexical resources developed by the partners.

- To discuss the applicability of existing methods and work techniques for the creation and maintenance of multilingual Slavic lexical resources and the possibilities for their enhancement.
- To offer expert recommendations: (1) for the standardization and integration of multilingual Slavic lexical resources and their availability to research, education, business, and the general public; (2) for the design of a common encoding scheme, representation of semantics, phraseology and etymology in bilingual and multilingual Slavic dictionaries.
- To develop a conceptual scheme for research networking infrastructure and cooperation of research groups working in digital Slavic lexicography, which should accelerate the preparation of digital and traditional multilingual dictionaries and enhance their quality.
- To outline an architecture and functional characteristics of the MONDILEX Linguistic Grid as a research infrastructure for the implementation of a network of multilingual digital resources.
- To supply the projected network of digital linguistic resources with facilities for opening these resources, making them widely accessible and usable not only by scholars of all disciplines and by education, but also in business and social communications.

The project presented these activities in a series of five open workshops, with attendant proceedings. First, the needs of the partners for a common infrastructure supporting scientific and applied activities in digital lexicography were analysed [22]. The state of the art in digital lexical resources and requirements for their integration were studied next [33]. The third workshop tackled innovative solutions for lexical entry design in digital Slavic lexicography [18]. The representation of semantics, phraseology, etymology and related matters were discussed next [24], with the last workshop concentrating on the research infrastructure for Slavic lexicography [17].

The project has now entered its final stage: design of a conceptual scheme for modelling an extensible infrastructure of institutions able to create and support a network of multilingual resources for Slavic languages.

In the next section we outline the main contributions to the project by partner institutions.

3 Partners' Contribution to the Networking Activities

The MONDILEX project has six partners: (1) Institute of Mathematics and Informatics, Bulgarian Academy of Sciences (IMI-BAS), Sofia, Bulgaria, which coordinates the project; (2) Institute of Slavic Studies, Polish Academy of Sciences (ISS-PAS), Warsaw, Poland, (3) Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences (ĽŠIL), Bratislava, Slovakia, (4) Jožef Stefan Institute (JSI), Ljubljana, Slovenia; (5) Institute for Information Transmission Problems, Russian Academy of Sciences (IITP-RAS); and (6) the Ukrainian Lingua-Information Fund of the National Academy of Sciences of Ukraine (ULIF-NANU), the two latter participants belonging to EU's International Cooperation Partner Countries.

3.1 Institute of Mathematics and Informatics, Bulgarian Academy of Sciences

The Mathematical Linguistics Department of the IMI-BAS has developed TEI-compliant digital language resources, among them corpora, lexicons, lexical databases, and electronic dictionaries. The first Bulgarian language-specific resources: morpho-syntactic specifications for encoding and annotating digital corpora and lexica, the electronic corpora: MTE Bulgarian-English parallel, annotated and aligned corpora, MTE Bulgarian annotated comparable corpus and the Bulgarian lexica, were developed in the EC project COP 106 *MULTEXT-East* Multilingual Text Tools and Corpora for Central and Eastern European Languages[2] (MTE for short, [6]). In order to clarify and update the Bulgarian specifications and make them more useful for annotation of corpora and automatic disambiguation of Bulgarian texts, we have proposed some changes to the Bulgarian specifications [13], in particular the treatment of participles, which should bring the morphosyntactic description better in line with the grammatical characteristics of Bulgarian.

The first lexical database (LDB) for integrated multilingual resources for Bulgarian was developed in the EC INCO Copernicus project PL96-1142 *CONCEDE* Consortium for Central European Dictionary Encoding[3]. The first Bulgarian-Polish bilingual electronic resources are being developed in the framework of a bilateral collaboration between IMI-BAS and ISS-PAS. The Bulgarian-Polish digital corpus contains more than 5 million words, and is composed of two corpora: parallel and comparable, [9]. The corpus is collected with the main purpose to ensure the selection of the entries for the first experimental electronic Bulgarian-Polish dictionary, the current version of which consists of approximately 20,000 dictionary entries. Bulgarian-Polish parallel corpus contains more than 3 million words, mostly contemporary fiction: novels, short stories, science fiction and children's literature. A small part comprises official documents of the European Commission available through the Internet. The corpus is composed of two parts: original Bulgarian texts with Polish translations or *vice versa* and texts in other languages translated into both Bulgarian and Polish. The most texts of the parallel corpus are annotated at paragraph level.

The Bulgarian and Polish teams are developing (currently for research purposes) the first *Bulgarian–Polish–Lithuanian experimental parallel corpus* [10]. The corpus contains over one million words.

In the framework of the joint collaborative project between IMI—BAS and ĽŠIL—Slovak AS a small *Slovak-Bulgarian parallel corpus* is currently under development only for research.

A small *parallel corpus* with Bulgarian, Polish, Slovak, Slovene (incl. English as a hub language) texts of official documents of the European Commission available through the Internet is also currently collected.

The Bulgarian comparable corpus includes *fiction* (texts from contemporary Bulgarian novels), *non-fiction* and *newspapers* (newspaper excerpts) subsets. The Bulgarian subset of MTE comparable corpus (approx. 200 000 words) were anno-

[2] `http://aune.lpl.univ-aix.fr/projects/multext-east/`
[3] `http://www.itri.brighton.ac.uk/projects/concede/`

tated manually. The rest of comparable corpus contains approximately 3 million words from works of Bulgarian authors, and Bulgarian translations of novels and short stories of prominent European authors (approx. 2 million words).

The formal model of the LDB [12] supporting the first experimental Bulgarian-Polish online dictionary is the CONCEDE model for dictionary encoding. The hierarchical structure of the dictionary entry is a well-formalised tree-structure. The selection of headwords included in this LDB is based on the Bulgarian-Polish parallel corpus. The main forms (lemmata) of the most frequent word forms in the corpus are selected. The word distribution according to parts of speech follows the CONCEDE model: open parts of speech — no more than 90%, closed parts of speech — minimum 10% of the whole set of lemmata chosen.

One of the main problems of the development of digital dictionaries is the choice of classifiers. Whenever the development of a system of bilingual dictionaries (serving as a future basis for a system of multilingual dictionaries) is concerned, there arises the issue of unification of the classifiers in the dictionary entry. In order to harmonise the classifiers for various languages, we need to present a unified selection of classifiers and a standard form of their presentation. In a broader sense, the issue of unifying classifiers in the dictionary entry is close to the issue of a new part-of-speech classification oriented towards the specifications of a digital dictionary. For example, the unification of classifiers in the proposed structure of the LDB that support the Bulgarian-Polish online dictionary allows synchronisation and unified representation for the data on the two languages [8, 11]. The work under the MONDILEX project demonstrates the potential for developing useful lexicographic reference works (both digital and hardcopy) by using the format of a LDB and an adequate mathematical foundation. Various parameters of classification of the lexicon are likely to emerge in the process of developing a lexical database. As this will possibly occur through distributed effort, it highlights the importance of an interface to the lexicographic system. The LDBs should be brought in line with one another by sharing theoretical concepts and platforms. Synchronisation and unification of bilingual dictionaries entails a uniform structure of the dictionary entry; the unification of classifiers for presenting headwords; a synchronous presentation of morpho-syntactic features, and a uniform presentation of the content.

Common suggestions of the Bulgarian and the Polish teams regarding the unification of classifiers can be grouped around the mode of classification of forms and the mode of denoting the meanings of verb tense forms (two types with exact definition that can be "translated" in a formal language).

3.2 Institute of Slavic Studies, Polish Academy of Sciences

The Department of Semantics of ISS-PAS tackles issues of linguistic confrontation of several Slavic languages. The team has elaborated a semantic interlanguage used for contrasting languages and worked on the distinction between a form and its meaning in dictionary entries. For the first time, a formal description of the meanings of tenses and aspects in Bulgarian, Polish, Russian and English has been proposed, together with a catalogue of meaning-related situations to be used for processing temporal semantic phenomena.

The team's work aims at constructing a Catalogue of Temporal Situations. Establishment of the proper correspondence between linguistic forms of temporality in various languages is of primary interest in building multilingual computerized dictionaries. Usually, this is done by comparing linguistic forms used in different languages, and trying to find out their similarities and differences. It may happen that two different forms in two different languages denote the same meaning; or two apparently identical forms can denote different situations. To avoid such cases, the team relies on temporal situations rather than on linguistic forms. In dictionaries, such situations are usually given as examples supplemented by informal comments. In multilingual dictionaries this approach is inconvenient and difficult to manage.

In a series of papers, the team proposes an approach which consists in executing the following steps:

1. prepare a list of temporal situations occurring in typical dictionaries as explanations of the usage of some verbal forms specific for the languages in question;
2. find a formal way of describing the listed situations, uniquely determining their specifications;
3. for any given situation and any eligible language, find the proper way of describing the relevant situation in the chosen language.

The papers give samples of situations and verbal forms used in 4 languages: English (as the reference), Bulgarian, Polish and Russian. Such descriptions are not intended for computer processing, but serve for understanding the meaning of sentences referring to chosen situations. When introducing a catalogue of situations, one can assign selected entries of such a catalogue to some sentences to be processed and then create a formal basis for their comparison in different languages.

There are several possibilities of defining the meaning of temporal properties of sentences. The team has chosen a description based on Petri Net theory [29]. Petri nets consist of three basic elements: events, states, and the flow relation. Any finite structure consists of these elements, with some of them joined by the flow relation in such a way that a state is connected to an event, or an event to a state (but neither two states nor two events are directly connected by the flow). An alternating sequence of states and events connected directly by the flow relation is called a path through the net, and indicates the sequence in which these elements appear in time. Any two elements of the same path are time-ordered. Nets with places which have only one incoming or outgoing arrow are called deterministic. Otherwise, nets are nondeterministic, and describe different, mutually exclusive temporal histories.

The team focuses on creating a catalogue of temporal situations that can be useful for comparison, analysis, processing, or translation of phrases in different languages containing temporal dependencies; and) distinguishing verbal forms from their temporal meanings. The findings described above were presented in [25, 26, 28].

A theoretical foundation for morphosyntactic specifications for Polish was proposed in the context of an extension of the MTE language-specific resources [30, 31].

3.3 Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences

Ľ. Štúr Institute of Linguistics is the central linguistic institution in the Slovak Republic. Its main areas of research and activities comprise research on contemporary Slovak language, traditional and computer-aided lexicography, Slovak dialectology, interaction with general and comparative linguistics in the fields of etymology, corpus linguistics and NLP research concerning the Slovak language

In the lexicography field, ĽŠIL is active in compiling traditional dictionaries, both its flagship — Short Dictionary of Slovak language [39] and its companion The Rules of Slovak Orthography [40], currently chosen by the Ministry of Culture as two of the publications defining the literary Slovak language; and multi-volume Dictinary of Slovak Language [41], marking an attempt to create a descriptive dictionary of modern standard Slovak, based on real linguistic data (mostly from the Slovak National Corpus). There are also other dictionary projects currently carried on, for example traditional Czech-Slovak dictionary and a wiki-based Slovak-Czech dictionary, produced in collaboration with the Czech Language Institute of the Czech Academy of Sciences [20].

Slovak National Corpus is a representative corpus of contemporary written texts, containing about 550 million words with automatic lemmatisation and morphological tagging. A smaller, balanced subcorpus consists of one third of journalistic texts, one third of specialised texts and one third of fiction, amounting to 200 million words. Another subcorpus contains manually lemmatised and annotated texts of about 700 000 words. A manually syntactically annotated corpus contains about 70 000 sentences, and a corpus of spoken Slovak contains about 400 000 words.

The Russian-Slovak, French-Slovak and Czech-Slovak sentence-aligned parallel corpora project is intended for linguistic research, teaching, translation, cross-linguistic studies and applications in natural language processing, primarily for machine translation, as well as dictionary compilation.

The Slovak Terminology Database Project is intended to provide a reference tool for domain experts as well as for linguists, translators, and lexicographers and general public as the database will represent a unique source of information both of linguistic and scientific character. It also enables to monitor terminology development in particular fields.

The whole corpus and other linguistic resources are publicly available in the Internet (with some licensing restrictions).

ĽŠIL's effort in the framework of the MONDILEX project has been to explore the possibilities of collaborative compiling and editing of mono- and bilingual dictionaries and other lexical databases, using, whenever possible, existing Open Source software resources. Towards this goal, the partner described implementation of several wiki-based lexicography databases: a database of Slovak language morphology, a Slovak-Czech bilingual lexicography database, and a database of Slovak collocations, including a system of mutual linking and references to the existing electronic dictionaries and text corpora.

The second area of ĽŠIL's involvement relates to corpus linguistics terminology, where there is no generally accepted terminology across the languages involved, and often not even among scientific institutions within the same language. The termi-

nology started to develop uncontrollably, either by directly adopting English terms or by calquing English expressions, or by adopting and extending existing linguistic terminology. The key issue is to harmonise and describe the definitions, ensuring consistency and clarity of information across the languages, especially when communicating with experts from various countries, where the use of a bridge language (usually English or Russian) is often insufficient or awkward. ĽŠIL designed and implemented a multilingual terminology database of corpus linguistics terms, with the goal of describing terminology of all the MONDILEX languages. The database has been tested with several Slavic language entries [38].

3.4 Jožef Stefan Institute, Ljubljana, Slovenia

The Department of Knowledge Technologies at the Jožef Stefan Institute has long-standing experience in the development of language resources, including research in automatic annotation techniques and encoding standardisation. In MONDILEX, the JSI partner has concentrated on two connected issues, the establishment of a Grid infrastructure, primarily for lexicography oriented corpus processing, and standards of encoding digital resources, with a focus on describing the morphosyntactic properties of words in lexica and annotated corpora.

A blueprint for the establishment of a Grid infrastructure for multilingual corpus processing, including the establishment of a virtual organization, rights and metadata management, was presented [15, 23]. Digital lexicography increasingly relies on large annotated corpora as the basic data source for lexical investigations. And while computational power of even personal computers has increased dramatically over the years, the sizes of corpora, and the computation resources needed to annotate, store, and investigate them have increased even more. Corpora for various languages can nowadays contain over a billion words, and annotation can easily increase their size by a factor of ten. Similarly, various automatic annotation tasks, such as word-alignment or word-sense disambiguation can be computationally very expensive, both in the training phase, for inductive methods, as well as in the application phases; computationally expensive are also the sophisticated investigations of today's lexicographers, where complex searches or other computations need to be performed over such large and heavily annotated corpora. We proposed methods to operationalise Grid processing over corpora, and performed several corpus annotation and analysis experiments, to establish speed advantages of Grid environments.

The Grid infrastructure puts an emphasis on rigorous user authentication and authorization, using PKI infrastructure to ensure that any task submitted to the Grid network originated by an authorized user and has access only to approved resources. However, in the NLP context, this is just the foundation of the necessary security environment, since the corpus linguistics deals with rather sensitive data (text corpora), for which the research institutions in questions usually do not have any distributions rights for the copyrighted texts. As the Grid should be regarded as a potentially hostile environment (the node where the computing takes place could get compromised or stolen, or even the node administrator might, despite his obligations, overstep his authority in accessing the data), the need of secure

data storage is paramount. We have analysed various possibilities of using symmetric or asymmetric encryptions that could provide the required level of protection. Some inspiration can be taken from CryptoSRM (cryptographic storage resource manager) and Hydra Key Storage (a distributed fragmented encryption key storage system). Implementing careful security and fine grained access control model we suppose to accommodate all the necessary requirements for differentiated user activities [19].

The Grid infrastructure is based on the Scientific Linux CERN distribution (SLC), whereas "usual" NLP tasks require often various other packages not present in the SLC, or specialised software often tailored to run inside specific environment. The standard and supported way is to put all the necessary software inside user environment, this however sometimes causes problems with modifying the required software to be installed into user defined directories (and solving the dependencies on other libraries or programs not present in the standard SLC, which have to be installed manually inside the user environment as well). We have investigated the possibility to use existing virtualization techniques to address the issue; installing desired GNU/Linux distribution and and all the necessary support software inside the standard POSIX chroot environment turned out to be straightforward and sufficient, although it requires the cooperation of Grid node administrator(s) to install the chroot environment. Using chroot in the Grid infrastructure is a non-standard feature, it is however widely supported.

In addition to requiring large amounts of computing power, lexicographers can also benefit from sharing resources, such as corpora (once the legal issues regarding access to the copyrighted files are sorted out). Additionally, common standards for file formats, corpora annotation and tools are necessary for such exchanges. We addressed these needs in the context of Grid computing, by presenting the underlying standards and evolving Grid infrastructures that need to be taken into account, giving a blueprint for IPR and access management, and discussing the language technology tasks useful in lexicography-oriented corpus processing, and how they could function in a Grid environment.

For Grid processing and storage, as well as for any interchange of resources, either between programs or humans, it is advantageous to have a common encoding scheme for digital lexica. We elaborated [27] a proposal which concentrates on the description of the morphological layer, and its connection to lexical and corpus resources; this is especially relevant for Slavic languages, which exhibit very rich inflection.

The presented encoding format for lexica is an application of the ISO standard LMF (Lexical Markup Framework), for which we detail the representation of inflectional paradigms, regular derivational relations, variant spellings, etc.

The framework and vocabulary of proposed common morphosyntactic annotations come from the MULTEXT-East project [6]. The MULTEXT-East standardised and linked set of language resources covers a large number of mainly Central and Eastern European languages and includes harmonised morphosyntactic resources consisting of the specifications, lexica and a parallel corpus. The MULTEXT-East resources, currently at Version 3 [14], are freely available for research use from `http://nl.ijs.si/ME/` and have been used in numerous studies

connected to language technologies. Most importantly, the MULTEXT-East resources were extended, for instance, specifications, lexicon and corpus were developed for Slovak. Furthermore, the differences between the MULTEXT-East specifications for Bulgarian and Slovak languages were analysed [7].

In the scope of MONDILEX and concurrent projects, we further standardised the specifications to express the features and define the tagsets used for lexical and corpus annotation, so that now the complete framework is encoded in XML, and compliant with the Text Encoding Initiative Guidelines P5 [16]. This new encoding enables more flexible language-particular encodings, localisations of feature names and codes, easy generation of derived formats (HTML, tabular, XML libraries), and simplifies the addition of new languages.

3.5 Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences

The Laboratory of Computational Linguistics of the Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, has developed a multipurpose linguistic processor, ETAP-3, which includes, among other things,

- a machine translation system operating between Russian and English, with small prototypes for other language pairs (French-Russian, Russian-German, Russian-Korean, Russian-Spanish, and Arabic-English) [1];
- a system of synonymous and quasi-synonymous paraphrasing of natural language utterances (in English and Russian),
- a module that enables computer-assisted translation of texts from UNL (Universal Networking Language, a semantic interlingua specially designed to facilitate multilingual communication in Internet) to natural languages and vice versa.

Another major project is SynTagRus, a deeply annotated corpus of Russian texts, in which every sentence is supplied with morphological tagging and a full syntactic structure represented in the dependency formalism as a tree of labelled syntactic dependencies between words (see e.g. [2]). A recent innovation in SynTagRus is the so-called lexical functional annotation, where arguments of lexical functions and their values are marked if these elements occur in sentences. The corpus is about 40,000 sentences (600,000 words) and constantly growing.

Both the ETAP-3 processor and SynTagRus rely on large digital dictionaries, including a Russian morphological dictionary with 130,000 entries and a Russian combinatorial dictionary (100,000 entries) that contains versatile and highly sophisticated information on lexical units.

In the combinatorial dictionary, every entry is divided into several zones. The first, universal, zone contains linguistic data relating to the source language while all the remaining zones present information referring to specific options: machine translation into a specific target language, paraphrasing etc. The universal zone is subdivided into several fields: 1) entry name field; 2) part of speech; 3) syntactic features; 4) semantic features, or descriptors; 5) government pattern, or subcategorisation frame, and 6) lexical functions. Additionally, an entry may have operational

fields where specific rules referring to particular stages of text analysis, or references to such rules, are given. Optional fields of discriminative comments can be used to distinguish between word senses of polysemous lexemes, or homonyms.

Many of the solutions implemented in the combinatorial dictionary of Russian can be extrapolated to digital dictionaries of other languages, first of all Slavic ones. This concerns, primarily, semantic features, lexical functions, semantically motivated syntactic features, and government patterns, many of which are quite similar across the languages [21]. Considering the fact that government patterns are developed with utmost care, and checked for consistency across lexicographic types and classes, they can serve as basis for developing similar resources for other Slavic languages.

Another suggestion by the IITP-RAS team is to use the UNL dictionary, which may be regarded as a universal dictionary of concepts [3], as a reference resource for contrasting digital dictionaries of the languages involved in the project, which could eventually lead to the creation of a large semantically focused Slavic multilingual lexicographic resource.

3.6 Ukrainian Lingua-Information Fund of the National Academy of Sciences of Ukraine

The ULIF-NASU is a repository of the National Dictionary Base of Ukraine. The institution's efforts are focused on computer technologies for creating the monolingual, bilingual, and multilingual dictionaries, and natural language processing systems. ULIF publishes a series of academic dictionaries "Dictionaries of Ukraine", which now numbers more than 70 volumes. As a member of TEI, ULIF develops the national standards for electronic text processing. The Fund also develops the Ukrainian National Linguistic Corpus [5, 32].

To create a unified language for dictionary structure description, a theory of lexicographic systems (L-systems) was developed [36, 37]. The theory, which combines the features of several formal structures for data description (data models, logical-linguistic calculi), was used to create an integrated L-system that tackles the phenomena of inflection, orthoepy, synonymy, antonymy, and phraseology of the Ukrainian language. An electronic dictionary based on this system can be accessed at the Ukrainian Linguistic Portal (`http://ulif.org.ua/`). ULIF's experience in creating the computer repository of national lexicography can be used in the MONDILEX project, as the repository can be regarded as a prototype for the lexicographic infrastructure in Slavistics. An important advantage of this repository is the fact that it relies on a general and standardized architecture of ANSI/X3/SPARK information systems, which distinguishes three levels of data abstraction (conceptual, internal and external). The theory of L-systems provides sufficient means for developing a well-formalized conceptual model for the network of digital lexicographic resources [35].

The theory of L-systems was implemented for creating a number of fundamental lexicographic databases that support the morphological models for different languages. One of them is a grammatical L-system of the Ukrainian language that supports paradigmatic classification of the Ukrainian vocabulary (more than

2500 inflectional classes). The total volume of the lexicographic database, used for this classification, is more than 550 thousand lexical units. A similar classification and the corresponding L-system were created for the Russian language. They were the basis for creating online computer paradigmatic dictionaries located on the Ukrainian Linguistic Portal that was created and supported by ULIF-NASU. The inflectional classifications for several other languages (German, French, Turkish, and Georgian) and the appropriate computer dictionaries were also developed in ULIF-NASU.

The research and development in the field of explanatory lexicography, creation of the explanatory dictionaries, computer semantic analyzers and tools for semantic marking of the text are now carried out in ULIF-NASU [4, 5, 34]. A considerable part of ULIF's activity is devoted to virtual systems of professional interaction in linguistics that enable the development of common lexicographic projects by researchers from different organizations or countries. Based on these achievements, a computer technology for compiling the explanatory dictionaries has been created in ULIF-NASU, which is implemented in the form of so-called virtual lexicographic laboratory (VLL). Thus, the VLL «Ukrainian Language Dictionary», «Russian Language Dictionary» and «Turkish Language Dictionary» have been created in ULIF-NASU. Using VLL «Ukrainian Language Dictionary» allowed creating a new 20-volume explanatory Ukrainian Language Dictionary in a fairly short time, which is now being prepared for publication.

Based on similar technology, the VLL «Polish-Ukrainian Dictionary» was created by ULIF-NASU and ISS-PAS, and the VLL «Russian-Ukrainian Dictionary» was created by ULIF-NASU and V.V. Vinogradov Institute of the Russian Language, RAS.

Practical testing of the virtual lexicographic laboratories created in ULIF-NASU has demonstrated high efficiency of the systems engineering. Currently virtual bilingual lexicographic laboratories are developed for the languages, whose representatives take part in the MONDILEX project. At the same time the systems engineering foundations for integration of the virtual lexicographic laboratories into a single integrated virtual lexicographic system capable of providing single research infrastructure for the entire MONDILEX project, are developed.

In future, the network of such virtual lexicographic laboratories may also function within the linguistic Grid.

4 Conclusion and Future Development

Finally, we underline the fact that three participants in the MONDILEX project — IMI-BAS, Bulgaria, ISS-PAS, Poland, and JSI, Slovenia — are members of CLARIN: a pan-European infrastructure that covers a variety of areas from language resources and technology to creation of a landscape of services at the European level, standardisation and application of language resources and technology to the humanities and social sciences, and aims to integrate national research centres and national resources. In all, MONDILEX has only 6 members, while CLARIN is a community of more than 70 institutions from 33 countries (31 European institutions are members of the project CLARIN consortium; there are no MONDILEX partners

among them). The scope of MONDILEX — conceptual modelling of networking of six centres for research in Slavic lexicography with a focus on bilingual and multilingual digital resources — is more specific than that of CLARIN. While CLARIN ultimately aims towards a common infrastructure of national language resources, MONDILEX is more concerned with collaboration between partners and eventual linking of their already existing diverse multilingual digital resources. Thus, the two projects are concerned with topics of mutual interest, namely:

1. creation of an advanced infrastructure on a European or global level to attract the best researchers
2. facilitating collaboration between research groups in a multilingual and multicultural society.

The MONDILEX participants agreed on the need of a common, unified part-of-speech and morphology classification system by taking as a basis the MULTEXT-East specifications. This should by no means supplant existing classifications widely used on national levels; rather it should be seen as an extension allowing for fast and rather effortless automated cross linguistic comparisons and applications.

Inclusion of extensive resources for Slavic languages not present in the MONDILEX project is planned in a new proposal, primarily, Belarusian, Croatian, and Czech. We also envisage the creation of a general lexical database with the possibility of searching for entries in any Slavic language, over diverse information, such as etymology, and correspondences between all Slavic languages, English, and, possibly, the UNL Universal dictionary of Concepts. The database could provide a ramified system of links between forms and senses of words in synchrony and diachrony. An interactive Web portal would enable the supervised extension of the database by the end-users, especially if the end-user will have possibilities (e.g. supported by a Wiki-style engine) to report missing words, errors etc. and to import their own lexical resources, via Grid-enabled software and storage, to ensure a fast growth and relevance to the user needs and requirements.

References

[1] Juri Apresjan, Igor Boguslavsky, Leonid Iomdin, Alexander Lazursky, Vladimir Sannikov, Victor Sizov, Leonid Tsinman. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT. In: *MTT 2003, First International Conference on Meaning — Text Theory.* Paris, École Normale Supérieure, Paris, June 16–18 2003, 279–288.

[2] Juri Apresjan, Igor Boguslavsky, Leonid Iomdin, Boris Iomdin, Andrei Sannikov, Victor Sizov. A Syntactically and Semantically Tagged Corpus of Russian: State of the Art and Prospects. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006).* Genoa, 2006, 1378–1381.

[3] Igor Boguslavsky, Vyacheslav Dikonov. Universal Dictionary of Concepts. In: Iomdin, Dimitrova (Eds. 2008), *Lexicographic Tools and Techniques.* 31–41, ISBN 978-5-990813-6-9.

[4] Bugakov, O. Definitions of Prepositions, Conjunctions and Particles in the Explanatory Dictionaries. In: Koseska, Dimitrova, Roszko (Eds. 2009), *Representing Semantics in Digital Lexicography.* Warsaw, 2009, 194–197. ISBN 978-83-89191-87-8.

[5] Bugakov, O. Using Ukrainian National Linguistic Corpus in Lexicography. In: Erjavec (Eds. 2009), *Research Infrastructure for Digital Lexicography.* Ljubljana, 2009, 120–124. ISBN 978-961-264-012-5/ISSN 1581-9973.

[6] Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.-J., Petkevič, V., and Tufis, D. MULTEXT-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In: *Proceedings of COLING-ACL '98. Montréal, Québec, Canada, 1998.* 315–319.

[7] Dimitrova, Ludmila, Radovan Garabík, Daniela Majchráková. Comparing Bulgarian and Slovak Multext-East morphology tagset. In: Shyrokov, Dimitrova (Eds. 2009), *Organization and Development of Digital Lexical Resources.* Kyiv, 2009, 38–46, ISBN 978-966-507-252-2.

[8] Dimitrova, Ludmila, Violetta Koseska-Toszewa. The Significance of Entry Classifiers in Digital Dictionaries. In: Iomdin, Dimitrova (Eds. 2008), *Lexicographic Tools and Techniques.* 89–97, ISBN 978-5-990813-6-9.

[9] Dimitrova, Ludmila, Violetta Koseska-Toszewa. Bulgarian-Polish Corpus. In *International Journal* Cognitive Studies | Études Cognitives. 9, SOW, Warsaw, 2009. 133–141, ISSN 2080-7147.

[10] Dimitrova, Ludmila, Violetta Koseska, Danuta Roszko, Roman Roszko. *Bulgarian-Polish-Lithuanian Corpus — Current Development.* In: Erjavec (Eds. 2009), *Research Infrastructure for Digital Lexicography.* Ljubljana, 2009, 72–86, ISBN 978-961-264-012-5.

[11] Dimitrova, Ludmila, Violetta Koseska-Toszewa, Joanna Satoła-Staśkowiak. Towards a Unification of the Classifiers in Dictionary Entry. In: Garabík (Ed. 2009), *Metalanguage and Encoding Scheme Design for Digital Lexicography.* Bratislava, 2009, pages 48–58. ISBN 978-80-7399-745-8.

[12] Dimitrova, Ludmila, Rumiana Panova, Ralitsa Dutsova. Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In: Garabík (Ed. 2009), *Metalanguage and Encoding scheme Design for Digital Lexicography.* Bratislava, 2009, 36–47. ISBN 978-80-7399-745-8.

[13] Ludmila Dimitrova, Peter Rashkov. A New Version for Bulgarian MULTEXT-East Morphosyntactic Specifications for Some Verbal Forms. In: Shyrokov, Dimitrova (Eds. 2009), *Organization and Development of Digital Lexical Resources.* Kyiv, 2009, 30–37, ISBN 978-966-507-252-2.

[14] Erjavec, Tomaz. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004.* Paris: ELRA, 1535–1538.

[15] Erjavec, Tomaz, Jan Jona Javoršek. Grid Infrastructure Requirements for Supporting Research Activities in Digital Lexicography. In: Iomdin, Dimitrova (Eds. 2008), *Lexicographic Tools and Techniques.* Moscow, 2008, 5–14, ISBN 978-5-990813-6-9.

[16] Erjavec, Tomaz. MULTEXT-East Morphosyntactic Specifications: Towards Version 4. In: Garabík (Ed. 2009), *Metalanguage and Encoding Scheme Design for Digital Lexicography*. Bratislava, 2009, 59–70. ISBN 978-80-7399-745-8.

[17] Erjavec, T. (Ed. 2009), *Research Infrastructure for Digital Lexicography. Proceedings of the MONDILEX Fifth Open Workshop, Ljubljana, 14–15 October, 2009*. ISBN 978-961-264-012-5/ISSN 1581-9973.

[18] Garabík, R. (Ed. 2009), *Metalanguage and Encoding Scheme Design for Digital Lexicography. Proceedings of the MONDILEX Third Open Workshop, Bratislava, 15–16 April, 2009*. ISBN 978-80-7399-745-8.

[19] Garabík, R, Javoršek, J.J, Erjavec, T. Evaluating Grid Infrastructure for Natural Language Processing. In: Proceedings of the 5th International Conference *NLP, Corpus Linguistics, Corpus Based Grammar Research.* Smolenice, 25–27 November 2009. pp. 193–105. ISBN 978-80-7399-875-2.

[20] Garabík, R., Špirudová, J. Design of a New Slovak-Czech Lexical Database. In: Garabík (Ed. 2009), *Metalanguage and Encoding Scheme Design for Digital Lexicography*. Bratislava, 2009, 71–76. ISBN 978-80-7399-745-8.

[21] Iomdin, Leonid. Representing Semantics in the Digital Combinatorial Dictionaries of the ETAP-3 System: New Developments. In: Koseska, Dimitrova, Roszko (Eds. 2009), *Representing Semantics in Digital Lexicography.* Warsaw, 2009, 69–75. ISBN 978-83-89191-87-8.

[22] Iomdin, Leonid, Ludmila Dimitrova. (Eds. 2008), *Lexicographic tools and techniques. Proceedings of the MONDILEX First Open Workshop* Moscow, *3–4 October, 2008*. ISBN 978-5-990813-6-9.

[23] Javoršek, J.J, Erjavec, T. Empowering Human Language Technologies with Grid. In: Erjavec (Ed. 2009), *Research Infrastructure for Digital Lexicography.* Ljubljana, 2009, 13–19, ISBN 978-961-264-012-5/ISSN 1581-9973.

[24] Koseska, Violetta, Ludmila Dimitrova, Roman Roszko. (Eds. 2009), Representing Semantics in Digital Lexicography. *Proceedings of the MONDILEX Fourth Open Workshop Warsaw, 29–30 July, 2009*. Warsaw, 2009. ISBN 978-83-89191-87-8.

[25] Koseska Violetta, Antoni Mazurkiewicz. Net representation of sentences in natural languages. *Advances in Petri Nets*, 1988, LNCS 340, Springer Verlag, 249–259.

[26] Koseska Violetta, Antoni Mazurkiewicz. Net Based Description of Modality in Natural Language (on the Example of Conditional Modality). In: Shyrokov, Dimitrova (Eds. 2009), *Organization and Development of Digital Lexical Resources.* Kyiv, 2009, 98–105, ISBN 978-966-507-252-2.

[27] Krek, S., Erjavec, T. Standardised Encoding of Morphological Lexica for Slavic languages. In: Shyrokov, Dimitrova (Eds. 2009), *Organization and Development of Digital Lexical Resources*. Kyiv, 2009, 24–29, ISBN 978-966-507-252-2.

[28] Mazurkiewicz, A. A Formal Description of Temporality (Petri Net approach). In: Iomdin, Dimitrova (Eds. 2008), *Lexicographic Tools and Techniques*. Moscow, 2008, 98–108. ISBN 978-5-990813-6-9.

[29] Petri, C. A. Fundamentals of the Theory of Asynchronous Information Flow, *Proceedings of IFIP'62 Congress.*1962, North Holland Publ. Comp., pp 386–390.

[30] Roszko, Roman. Morphosyntactic specification for Polish. Theoretical Foundations. In: Garabík (Ed. 2009), *Metalanguage and Encoding Scheme Design for Digital Lexicography.* Bratislava, 2009, 140–150,
Roszko, Roman. Description of Morphosyntactic Markers for Polish Verbs within MULTEXT-East Morphosyntactic Specifications. In: Koseska, Dimitrova, Roszko (Eds. 2009), *Representing Semantics in Digital Lexicography.* Warsaw, 2009, 159–165. ISBN 978-83-89191-87-8

[31] Shyrokov Volodymyr, Oleg Bugakov, T. Griaznukhina, etc. *Corpus Linguistics.* Kiev, 2005. ISBN 966-507-189-0. (In Ukrainian).

[32] Shyrokov, Volodymyr, Ludmila Dimitrova. (Eds. 2009) *Organization and Development of Digital Lexical Resources. Proceedings of the MONDILEX Second Open Workshop, Kiev, 2–4 February.* 2009. ISBN 978-966-507-252-2.

[33] Shyrokov Volodymyr. *Elements of Lexicography.* Kiev, 2005. ISBN 966-507-187-4. (In Ukrainian)

[34] Shyrokov Volodymyr. Experience in Creating a National Dictionary Depository of Ukraine and its Use in Conceptual Modelling of Networking of Centres for High-quality Research in Slavic Lexicography and their Digital Resources. In: Shyrokov, Dimitrova (Eds. 2009), *Organization and Development of Digital Lexical Resources.* Kyiv, 2009, 5–8. ISBN 978-966-507-252-2.

[35] Shyrokov, V. Theory of Lexicographic Systems. Part 2. In: Koseska, Dimitrova, Roszko (Eds. 2009), *Representing Semantics in Digital Lexicography.* Warsaw, 2009, 89–105, ISBN 978-83-89191-87-8.

[36] Shyrokov, V. Theory of Lexicographic Systems. Part 3. In: Erjavec (Eds. 2009), *Research Infrastructure for Digital Lexicography.* Ljubljana, 2009, 98–119. ISBN 978-961-264-012-5/ISSN 1581-9973.

[37] Šimková, Maria, Radovan Garabík, Ludmila Dimitrova. Design of a multilingual terminology database prototype. In: Koseska, Dimitrova, Roszko (Eds. 2009), *Representing Semantics in Digital Lexicography.* Warsaw, 2009, 123–127, ISBN 978-83-89191-87-8.

[38] [39] Krátky slovník slovenského jazyka. Ed. Považaj, M., Bratislava: Veda 2003.

[39] [40] Pravidlá slovenského pravopisu. 3., upravené a doplnené vydanie. Ed. Považaj, M., Bratislava: Veda 2000.

[40] [41] Slovník súčasného slovenského jazyka. A–G. Eds. Buzássyová, K., Jarošová, A., Bratislava: Veda 2006.