

Bulgarian Digital Resources in a Multilingual Context

Ludmila Dimitrova

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences

ludmila@cc.bas.bg

Introduction

The first Bulgarian electronic corpora and language-specific resources were developed in the the Department of Mathematical Linguistics of the Institute of Mathematics and Informatics (IMI) at the Bulgarian Academy of Sciences (BAS). The IMI – BAS was created in 1947, the Department of Mathematical Linguistics was founded in 1977 (as the Laboratory for Mathematical Linguistics, 1977-1985). It is the successor of the Group for Machine Translation (1964-1976), which worked in the areas of Russian-Bulgarian automatic translation and quantitative and statistical studies of the Bulgarian language.

The Department participated in two large language engineering EC projects:

- COP project 106 **MULTEXT-East** *Multilingual Text Tools and Corpora for Central and Eastern European Languages*, 1995-1997, coordinator Jean Véronis, CNRS, <http://aune.lpl.univ-aix.fr/projects/multext-east/>,
- INCO Copernicus project PL96-1142 **CONCEDE** *Consortium for Central European Dictionary Encoding*, 1998-2000, coordinator Roger Evans, University of Brighton, <http://www.itri.brighton.ac.uk/projects/concede/>.

A recent project in this field is the 7th FP project GA 211938 **MONDILEX**, 2008-2010, coordinator Ludmila Dimitrova, IMI – BAS; www.mondilex.org.

MULTEXT-East (MTE) project is a continuation of the EU project MULTEXT for six Central and Eastern European languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene, three of which are Slavic – Bulgarian, Czech, and Slovene. The CONCEDE project successfully employs the resources developed in MTE. Our experience shows that continuity is an important prerequisite for the success of EC-financed projects.

1. Bulgarian language-specific resources - morphosyntactic specifications

In the framework of the MTE project were developed the morphosyntactic specifications for Bulgarian. They contain the list of defined categories – parts of speech (POS). Each POS is encoded by a letter: noun - N, verb - V, adjective - A, pronoun -P, determiner - D, article - T, adverb - R, adposition - S, conjunction - C, numeral - M, interjection - I, residual - X, abbreviation - Y, particle - Q. A table of attribute-values is defined for each category in order to reflect the characteristic features of each so-called *MTE languages*: Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene. The specific features of each language are marked up additionally by **Is**. The characters following the POS-encoding give the values of the position-determined attributes.

The specifications define, for each part of speech, its appropriate attributes and their values, encoded by one symbol code. If a certain attribute is not appropriate for a language, for the particular combination of features, or for the word, this is marked by a hyphen in the attribute's position.

MTE use the EC project MULTEXT format of lexical description - morphosyntactic description (MSD).

MSD consists of linear strings of characters, representing the morphosyntactic information for each word-form. The string is constructed in the following way:

- the positions of a string of characters are numbered 0, 1, 2, ...;
- the agreed character at position 0 encodes the corresponding part of speech: N for noun, V for verb, A for adjective, etc. ;
- each character at position 1, 2,..., n, encodes the value of one attribute (for nouns the attributes are: type, gender, number, case, definiteness).

For example, the MSD of the word **стената** /the wall/ is **Ncfs-y** that means POS: noun, Type: common, Gender: feminine, Number: singular, no Case, Definiteness: yes.

The proposed formalism for the MSD is not arbitrary (a MSD contains the full description of a lexical item), but has a clear and concrete aim – to be used for specific applications, incl. **corpus annotation** (the process of adding linguistic information in an electronic form to a text corpus). The most common and important type of corpus annotation is morphosyntactic annotation (grammatical tagging or **POS tagging**), where a **label or tag** is associated with each word in the text in order to indicate its grammatical classification. On the basis of these standard MSDs the set of corpus tags were determined. The list of MSDs for Bulgarian contains 326 elements.

2. Corpora

2.1 Bulgarian MTE parallel annotated corpus

MTE is building an annotated multilingual corpus, composed of three major parts:

- **Parallel Corpus**,
- **Comparable Corpus**,
- **Speech Corpus** (a small one) of spoken texts in each of the six languages, comprising forty short passages of five thematically connected sentences, each spoken by several native speakers, with phonemic and orthographic transcriptions.

Multilingual parallel corpus, based on George Orwell's novel "1984" in the English original and the six translations in Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene of the novel, was developed. The parallel corpus is produced as a well-structured, lemmatized, CES-corpus.

The texts were automatically annotated for tokenization, sentence boundaries, and part of speech annotation, using the project tools, and validated for sentence boundaries and alignment.

The alignment between the English version and a translation in each of the six CEE languages ensures six pair-wise alignments.

For **Bulgarian**, the alignment was made by the Vanilla aligner - 6699 bilingual links in total:

Aligned pairs:	2-2	2	0.030017%
	2-1	23	0.345190%
	1-2	36	0.540297%
	1-1	6637	99.074487%
	0-1	1	0.014970%

The next examples show excerpts of the *Bulgarian-English aligned 1984 texts*:

1-1 Aligned sentences:

<Obg.1.1.2.4> Уинстън се запъти към стълбите.

<Oen.1.1.2.4> Winston made for the stairs.

<Obg.1.1.2.5> Излишно бе да вика асансьора.

<Oen.1.1.2.5> It was no use trying the lift.

1-2 Aligned sentences:

<Obg.1.1.23.16> Не беше много вероятно и въпреки това винаги, когато тя бе наоколо, той изпитваше странно чувство на неудобство, примесено със страх, дори враждебност.

<Oen.1.1.24.16> That, it was true, was very unlikely.<Oen.1.1.24.17> Still, he continued to feel a peculiar uneasiness, which had fear mixed up in it as well as hostility, whenever she was anywhere near him.

<Obg.1.1.24.8> Изпитваше дълбок интерес към него не само защото беше заинтригуван от контраста на изисканите маниери с телосложението му на борец, а много повече заради стаената увереност -- или навярно не толкова увереност, колкото надежда, -- че политическата правоверност на **О'Брайън** не е изрядна.

<Oen.1.1.25.8> He felt deeply drawn to him, and not solely because he was intrigued by the contrast between **O'Brien's** urbane manner and his prize-fighter's physique.<Oen.1.1.25.9> Much more it was because of a secretly held belief -- or perhaps not even a belief, merely a hope -- that **O'Brien's** political orthodoxy was not perfect.

2.2 Bulgarian-Polish parallel corpus

The *first Bulgarian–Polish corpus* is currently under development in the framework of the joint collaborative project “Semantics and contrastive linguistics with a focus on a bilingual electronic dictionary” between IMI-BAS and ISS-PAS, coordinated by L. Dimitrova and V. Koseska. It contains approx. 5 million words and comprises two corpora: parallel and comparable.

Bulgarian–Polish parallel corpus contains more than 3 million words mainly in works of Bulgarian and Polish authors – short stories, novels, children’s literature, science fiction. 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.

2.3 Other parallel corpora with Bulgarian

In the framework of the joint collaborative project „Electronic corpora – contrastive study with focus on design of Bulgarian-Slovak digital language resources“ between IMI—BAS and LŠIL—Slovak AS, coordinated by L. Dimitrova and R. Garabík a small *Slovak-Bulgarian parallel corpus* is currently under development.

The first *Bulgarian-Polish-Lithuanian corpus*, currently under development only for research, contains more than 3 million words. It comprises two corpora: parallel and comparable. The *parallel Bulgarian-Polish-Lithuanian corpus* contains more than 1 million words.

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.

2.4 Bulgarian comparable corpora

For each of the six MTE CEE languages, a comparable corpus was developed. It included two subsets of at least 100 000 words each, consisting of fiction, comprising a single novel or excerpts from several novels; newspapers.

The data was comparable across the six languages, only in terms of the number and size of texts. The entire MTE multilingual comparable corpus was prepared in CES format, manually or using ad-hoc tools.

The *Bulgarian MTE comparable corpus* includes *Fiction* (texts from contemporary Bulgarian literature) and *Newspapers* (newspaper excerpts) subsets. The texts were annotated at paragraph level.

The *Bulgarian comparable corpus in Bulgarian-Polish corpus* contains approximately 3 million words from works of Bulgarian authors, including:

- prose: Dimitar Talev, Dimitar Dimov, Pavel Vezhinov, Yordan Radichkov,
- non-fiction: Zhelyu Zhelev’s „Fascism“,
- Bulgarian translations of novels and short stories of prominent European authors.

3. Bulgarian Lexical Databases

3.1 Bulgarian Lexical Databases for CONCEDE

The first lexical database (LDB) for Bulgarian was developed in the framework of the **CONCEDE** project. The lexical databases of the project CONCEDE were developed on the basis of the MTE parallel multilingual corpus. The CONCEDE project suggested a model for dictionary encoding containing a lexical database with standardized and well-understood structure and semantics.

The CONCEDE project has developed lexical databases (LDBs) in a general-purpose document-interchange format for the same six MTE CEE languages: 3000-headword lexical databases for Bulgarian, Czech, Estonian, Hungarian, Romanian, and a 500-word one from the English-Slovene dictionary.

Under the CONCEDE project was developed an *encoding scheme for lexicographic specifications*

of the Bulgarian language, according to the standards for electronic dictionary encoding. This encoding scheme served to create the Bulgarian dictionary in the LDBs of CONCEDE. The choice of dictionary entries follows the method accepted by CONCEDE. The entries are equipped with lexicographic specifications for Bulgarian language in TEI-conformant SGML.

The electronic dictionary is based on the *Bulgarian Explanatory Dictionary*. Each entry in BDB is represented as a tree-structure.

For example, the entry with headword “име” in the *Bulgarian Explanatory Dictionary* is:

име *ср.* Отличително название на човек, животно и др. прен. Известност. *Той има голямо име.* грам. Категория думи, които означават предмети, качества, числа. *Съществително име. Прилагателно име. Числително име.* ◇ В името на предл. Въз основа на, заради. В името на закона. В името на свободата.

The corresponding entry in the Bulgarian LDBs has the following structure:

```
<entry><hw>|име</hw>
<gen>ср.</gen>
<struc type="Sense" n="1">
<def>Отличително название на човек, животно и др.</def></struc>
<struc type="Sense" n="2"><usg type="register">прен.</usg>
<def>Известност.</def>
<eg><q>Той има голямо име.</q></eg></struc>
<struc type="Sense" n="3"><usg type="register">грам.</usg>
<def>Категория думи, които означават предмети, качества, числа.</def>
<eg><q>Съществително име.</q></eg>
<eg><q>Числително име.</q></eg></struc>
<struc type="Phrases">
<struc type="Phrase" n="1"><orth>В името на</orth><pos>предл.</pos>
<def>Въз основа на, заради.</def>
<eg><q>В името на закона.</q></eg>
<eg><q>В името на свободата.</q></eg></struc></struc>
</entry>
```

3.2 Bulgarian Lexical Databases supporting Bulgarian-Polish digital dictionary

We use the CONCEDE model for dictionary encoding during the process of design and development of the LDB supporting the Bulgarian-Polish digital dictionary. The tagset consists of two types of tags, namely:

- **Structural Tags:** alt, entry, struc,
- **Content Tags:** case, def, domain, eg, etym, gen, geo, gram, hw, itype, lang, m, mood, number, orth, person, pos, q, register, source, subc, time, tns, trans, usg, xr.

In order to obtain a full correspondence between the specific characteristics of Bulgarian and the formal description of the dictionary entries in the database we introduced new content tags for Bulgarian verbs:

- conjugation (to represent the conjugation of verbs)
- type (for the type of conjugation)

We introduced new additional information for Bulgarian verbs: in the tag <subc> - to express transitivity/intransitivity and in the tag <gram> - to express perfect aspect and progressive aspect.

Realization of homonyms: the meanings of homonyms are entered in the dictionary as different DB records. In the page for entering the words there is a field where the user must specify a homonym index - a number which shows the order of the meanings.

For the representation of the homonym it is necessary to fill in the value of the attribute n (homonym index) in the tag <entry>:

- <entry n="1">
- <entry n="2">

4. Bulgarian-Polish Digital dictionaries

4.1 Bulgarian-Polish electronic dictionary

The experimental version of the first Bulgarian–Polish electronic dictionary is prepared in WORD-format and consist approximately 20 thousand dictionary entries till now. This dictionary provides a part of the language material for the LDB of the web-based application that supports Bulgarian-Polish online dictionary.

4.2 Bulgarian-Polish online dictionary

The Bulgarian-Polish online dictionary pursues so far experimental purposes. A LDB provides the language material for the dictionary.

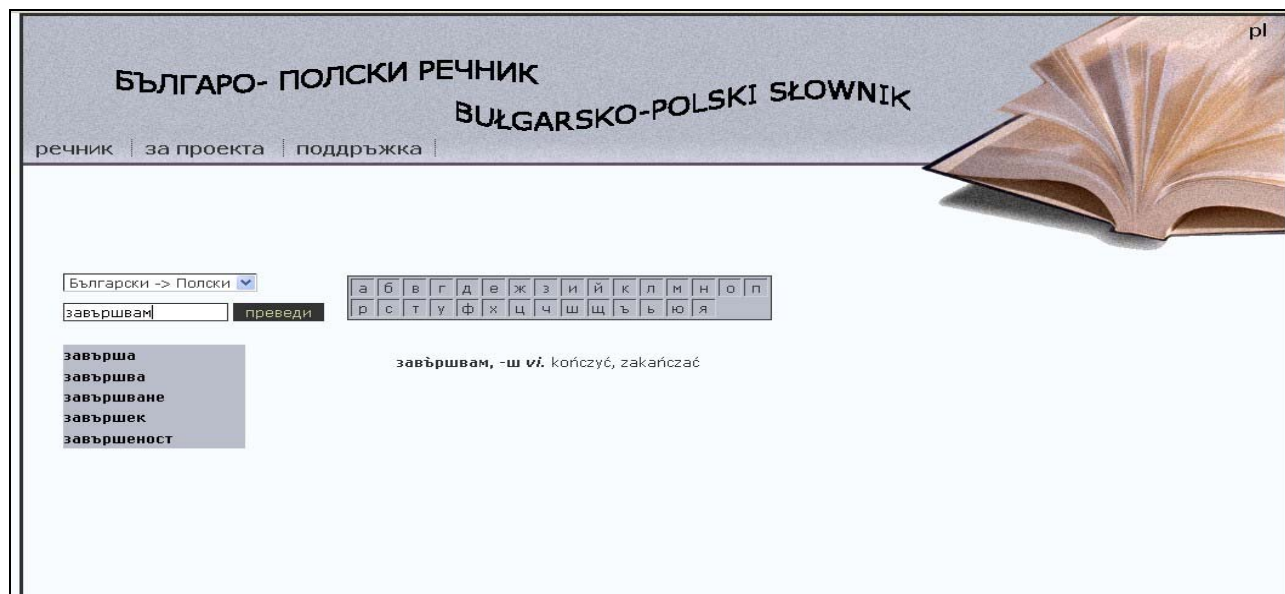
The web-based application representing the Bulgarian-Polish online dictionary consists of two primary modules: administrator module and end-user module.

The administrator module is intended for the person updating the dictionary, and is accessible only for authorized users. The end-user module is aimed at presenting correct and up-to-date information to the user.

- To be convenient and easy for searching and finding the meanings of words:
- An opportunity for translation from Polish to Bulgarian.
- To allow the end-user to report missing words.
- To create a user interface in both languages – Bulgarian and Polish

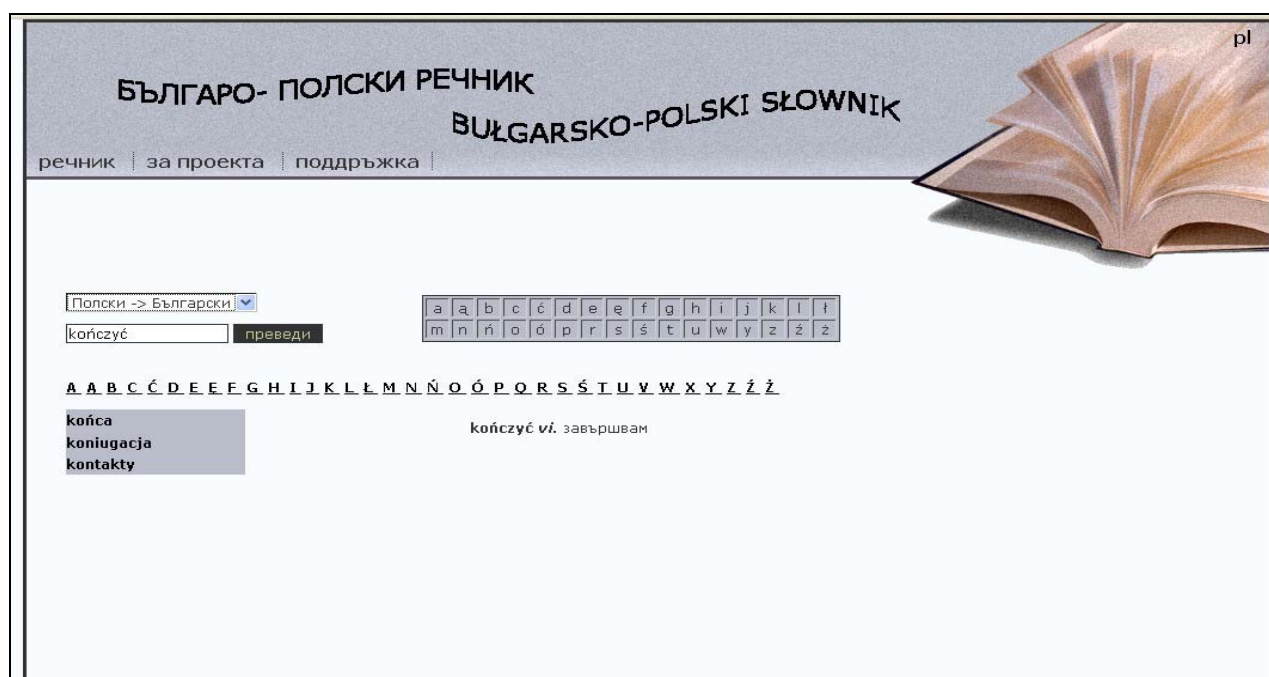
For the program realization of the web-based application the IC technologies Apache, MySQL, PHP and JavaScript have been used; these are free technologies originally designed for developing dynamic web pages with a lot of functionalities. The current version of the Bulgaria-Polish online dictionary works optimally with Internet Explorer 6.0+ (Windows), and with Firefox 2.0.1+ (Windows, Linux).

The window shown below illustrates the translation of the Bulgarian verb “завършвам” /to finish/ into Polish:



The screenshot displays the web interface of the Bulgarian-Polish dictionary. At the top, the title "БЪЛГАРО- ПОЛСКИ РЕЧНИК" (BULGARO-POLSKI SŁOWNIK) is shown in both Bulgarian and Polish. Below the title, there are navigation links: "речник", "за проекта", and "поддръжка". The main content area features a search bar with a dropdown menu set to "Български -> Полски". The search input contains the Bulgarian word "завършвам" and a "преведи" button. To the right of the search bar is a keyboard layout with letters in Bulgarian and Polish. Below the search bar, a list of Bulgarian forms for "завършвам" is shown: "завърша", "завършва", "завършване", "завършек", and "завършеност". To the right of this list, the Polish translation "завършвам, -ш *vł.* kończyć, zakańczyć" is displayed. An image of an open book is visible in the top right corner of the interface.

The window shown below illustrates the translation of the Polish verb “kończyć” /to finish/ into Bulgarian:



5. EC FP7 Project MONDILEX *Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources*

The participants are:

- 1) Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
- 2) Institute of Slavic Studies, Polish Academy of Sciences
- 3) Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences
- 4) Jožef Stefan Institute, Ljubljana, Slovenia
- 5) Institute for Information Transmission Problems, Russian Academy of Sciences
- 6) Ukrainian Lingua-Information Fund, National Academy of Sciences of Ukraine

The partners in the project are research organisations from six European countries whose six national languages belong to the Slavic group: Bulgaria, Poland, Russia, Slovakia, Slovenia and Ukraine. All partners are national centres for high-quality research in lexicography and digital resources. Each partner is responsible for coordinating a part of the work matching their specialisation and experience.

The main objective of the MONDILEX project is to design the conceptual scheme of a **research infrastructure** that supports the networking of centres for high-quality research in digital 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. 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 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 and using state-of-the-art network technologies for information exchange between the participants. 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 its activities in a series of five open *MONDILEX* workshops:

- 1) *Lexicographic tools and techniques*. Moscow, 3-4 October, 2008,
- 2) *Organization and Development of Digital Lexical Resources*. Kiev, 2-4 February, 2009,
- 3) *Metalanguage and Encoding Scheme Design for Digital Lexicography*. Bratislava, 15-16 April, 2009,
- 4) *Representing Semantics in Digital Lexicography*. Warsaw, 29-30 July, 2009,
- 5) *Research Infrastructure for Digital Lexicography*. Ljubljana, 14-15 October, 2009.

The needs of the partners for a common infrastructure supporting scientific and applied activities in digital lexicography were analysed in the beginning [7]. The state of the art in digital lexical resources and requirements for their integration were studied next [9]. The third workshop tackled innovative solutions for lexical entry design in digital Slavic lexicography [6]. The representation of semantics, phraseology, etymology and related matters were discussed next [8], with the last workshop concentrating on the research infrastructure for Slavic lexicography [5].

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.

References

- [1] 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. pp. 315-319.
- [2] Dimitrova, Ludmila, Violetta Koseska-Toszewa. Bulgarian-Polish Corpus. In *International Journal Cognitive Studies/Études Cognitives*. 9, SOW, Warsaw, 2009. pp. 133-141. ISSN 2080-7147.
- [3] Dimitrova, Ludmila, Violetta Koseska, Danuta Roszko, Roman Roszko. Bulgarian-Polish-Lithuanian Corpus—Current Development. In: *Proceedings of the International Workshop “Multilingual resources, technologies and evaluation for Central and Eastern European languages” in conjunction with International Conference Recent Advance in NPL’2009*. Borovec, Bulgaria, 17 September 2009. pp. 1-8.
- [4] Dimitrova, Ludmila, Rumiana Panova, Ralitsa Dutsova. Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In: Garabik (Ed. 2009), *Metalanguage and Encoding scheme Design for Digital Lexicography*. pp. 36-47.
- [5] Erjavec, T. (Ed. 2009) *Proceedings of the MONDILEX Fifth Open Workshop: Research Infrastructure for Digital Lexicography*. Ljubljana, 14-15 October, 2009. ISBN 978-961-264-012-5. 124 pp.
- [6] Garabík, R. (Ed. 2009) *Proceedings of the MONDILEX Third Open Workshop: Metalanguage and Encoding Scheme Design for Digital Lexicography*. Bratislava, 15-16 April, 2009. ISBN 978-80-7399-745-8. 191 pp.
- [7] Iomdin, Leonid, Ludmila Dimitrova. (Eds. 2008) *Proceedings of the MONDILEX First Open Workshop: Lexicographic tools and techniques*. Moscow, 3-4 October, 2008. ISBN 978-5-990813-6-9. 109 pp.
- [8] Koseska, Violetta, Ludmila Dimitrova, Roman Roszko. (Eds. 2009) *Proceedings of the MONDILEX Fourth Open Workshop: Representing Semantics in Digital Lexicography*. Warsaw, 29-30 July, 2009. ISBN 978-83-89191-87-8, 218 pp.
- [9] Shyrovkov, Volodymyr, Ludmila Dimitrova. (Eds. 2009): *Proceedings of the MONDILEX Second Open Workshop: Organization and Development of Digital Lexical Resources*. Kiev, 2-4 February, 2009. ISBN 978-966-507-252-2. 129 pp.