COGNITIVE STUDIES | ÉTUDES COGNITIVES, 11
SOW Publishing House, Warsaw 2011

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
(ludmila@cc.bas.bg)
[2]Institute of Slavic Studies, Warsaw, Poland
(amaz1312@gmail.com)
[3]Ľ. Štúr Institute of Linguistics, Bratislava, Slovakia
(garabik@kassiopeia.juls.savba.sk)
[4]Jožef Stefan Institute, Ljubljana, Slovenia
(tomaz.erjavec@ijs.si)
[5]Institute for Information Transmission Problems, Moscow, Russia
(iomdin@iitp.ru)
[6]Ukrainian Lingua-Information Fund, Kiev, Ukraine
(vshirokov48@mail.ru)

MAIN RESULTS OF MONDILEX PROJECT

Abstract

The paper presents the results and recommendations of MONDILEX, a 7FP project that covered six Slavic languages: Bulgarian, Polish, Russian, Slovak, Slovene, and Ukrainian. The paper summarizes the research undertaken on standardisation and integration of Slavic language resources and on the establishment of a virtual organisation supporting research infrastructure for Slavic lexicography The results should be useful for an implementation of a research infrastructure in the coming years.

Keywords: Slavic languages, digital language resources, language technologies, digital lexicography, research infrastructure.

1. Introduction

The MONDILEX project *Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources* is a EU project funded by the European Commission within the 7th FP in the field "Capacities – Research Infrastructures: Design studies for research infrastructures in all sciences and technologies fields".

The main objective of the MONDILEX[1] project was to design a conceptual scheme of a research infrastructure supporting the networking of centres for high-

[1]http://www.mondilex.org

quality research in Slavic lexicography. Research infrastructures in general function as sets of strategic centres of excellence for research, education and training, whose chief aim is facilitating scientific cooperation and public partnership as well as strengthening the interaction between research and applications. As such, research infrastructures greatly contribute to the development of the knowledge society. The MONDILEX project was motivated by the need of a sustainable and scalable infrastructure for institutions involved in creating and supporting a network of multilingual resources of Slavic languages (Dimitrova et al. 2010). Such an infrastructure is necessary in view of the obvious mismatch between the importance of Slavic languages, spoken by a substantial part of Europe's population, and the insufficient number and inadequate quality of digital lexical resources for these languages.

Other important objectives of the MONDILEX project were to study problems in the development, management, and reuse of lexical resources in a multilingual context. The increased the EU participation of countries whose national languages belongs to the Slavic group, as well as intensified communication with non-EU Slavic countries, brings up the issue of standardisation of digital bi- and multilingual resources. This is needed to facilitate exchange and serve in education, business, and research.

In our ever expanding information society, most information systems are now facing the challenges of multilingualism. Lexical resources, which play an essential role in these systems, should provide information on many languages in a common framework and should be reusable in many automatic applications and human practices. Many centres have been involved in national, European or international projects dedicated to building harmonized language resources and creating expertise in the maintenance and further development of standardized linguistic data. These resources include those developed along the lines of best practices and recommendations: corpora (mono- and multilingual, parallel, comparable, and annotated), dictionaries (mono- and bilingual, electronic and online), lexicons, thesauri, wordnets, ontologies etc. However, efforts in evaluating these resources remain the responsibility of local authorities, usually with limited funding and few opportunities for academic assessment and recognition of the achieved results.

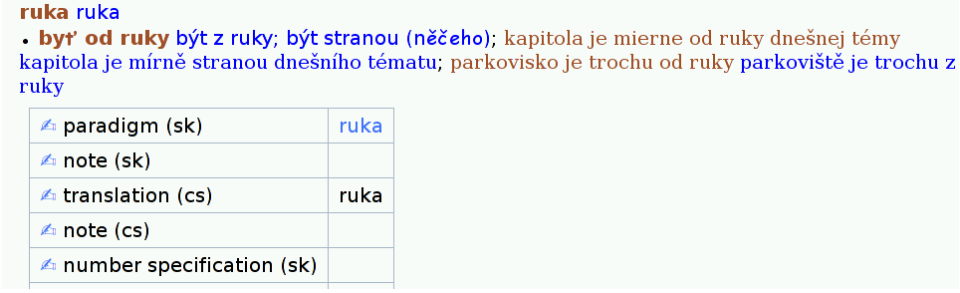
The MONDILEX project examined strategies for the coordination, unification, integration and extension of existing digital lexical resources and the creation of new ones, in accordance with recent advances in the field and international standards. A series of five MONDILEX open workshops investigated these problems. The first workshop analysed the partners' needs for a common infrastructure supporting scientific and applied activities in digital lexicography (Iomdin, Leonid & Dimitrova, Ludmila, Editors. 2008). The second workshop studied the state of the art in digital lexical resources and requirements for their integration (Shyrokov, Volodymyr & Dimitrova, Ludmila, Editors, 2009). The third workshop tackled innovative solutions for lexical entry design in digital Slavic lexicography (Garabík, Radovan, Editor, 2009). The representation of semantics, phraseology, etymology and related matters were discussed in the fourth workshop (Koseska, Violetta, Dimitrova, Ludmila, Roszko, Roman, Editors. 2009). The last workshop focused on the research infrastructure for Slavic lexicography (Erjavec, Tomaž, Editor, 2009).

The MONDILEX project surveyed the fundamental concepts of traditional and digital lexicography and presented a conceptual scheme of a research infrastructure supporting the networking of centres for high-quality research in Slavic lexicography.

2. Evaluation of Slavic language resources for digital lexicography

During the course of the project, **all types of language resources** that should be included in a research infrastructure for Slavic lexicography were discussed in detail and evaluated.

Initially, a set of ***lexical databases*** (LDBs) for Slavic languages were analyzed and discussed, including a Slovak-Czech LDB (Garabík, Špirudová 2009), a Bulgarian-Polish LDB (Dimitrova et al. 2009a), a multilingual corpus linguistics terminology database (Šimková et al. 2009), a Slovak morphology database (Garabík 2008), and a paremiography database (Ďurčo, Garabík 2009). The analysis focused on the problems and difficulties of database support arising due to LDB's internal logical complexity, alignment of the structure and content tags of LDB's structural units to international standards, as well as compatibility with language resources created in other projects and for other languages. In this context, some conceptual models of actual electronic databases were described, among them the Slovak-Czech LDB and multilingual terminology database, both compiled with the MoinMoin wiki system, and Bulgarian-Polish LDB based on CONCEDE[2] model for dictionary encoding. The proposed structure of LDB allows synchronized and unified representation of the linguistic information.

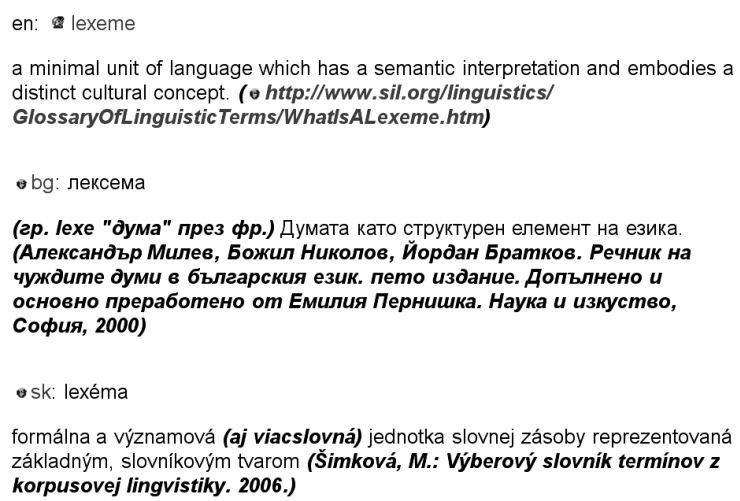

ruka ruka

• **byť od ruky** být z ruky; být stranou (něčeho); kapitola je mierne od ruky dnešnej témy kapitola je mírně stranou dnešního tématu; parkovisko je trochu od ruky parkoviště je trochu z ruky

paradigm (sk)	ruka
note (sk)	
translation (cs)	ruka
note (cs)	
number specification (sk)	

Example from the Slovak-Czech LDB

[2]The EC project CONCEDE *Consortium for Central European Dictionary Encoding* developed lexical databases for six CEE languages. See http://www.itri.brighton.ac.uk/projects/concede/.



Example from the multilingual corpus linguistics terminology database

When dealing with various languages, it is important that all participants agree upon a common terminology for the problem at hand. This is all the more important when Slavic lexicography is concerned, mostly because of two opposite phenomena: first, linguistics studying different languages have traditionally used different ways of analysing (the same) grammar categories, which results in conflicting use of professional terms in different languages; and second, newly emerging branches of linguistics do not yet have their native terminology stabilized across languages. In order to facilitate professional discussion and information exchange, we recommend creating a corpus linguistics terminology database.

MONDILEX designed and implemented a prototype of a multilingual corpus linguistics terminology database, which contains terms in Bulgarian and Slovak, with relevant English equivalents, intended to facilitate collaboration among MONDILEX member institutes.

The aim of this database is to minimize the barriers of internal communication due to the fact that certain terms could be incompatible across Slavic languages or missing altogether, and to unify existing terminology. The database includes corpus linguistics entries from the Slovak Terminology Database (Levická 2007, 2008).

Future extensions shall proceed towards creating a database of all languages of the MONDILEX project, including English (added as a hub language, and also because most terminology originates from it). Such a database can serve as a nucleus of a multilingual terminology database of lexicographic (or even general linguistic) terms.

Second, different kinds of applications were exemplified by the following ***digital dictionaries of Slavic languages***: the dictionary of Slovak collocations (Ďurčo P. et al. 2009), Bulgarian-Polish on-line dictionary (Dimitrova et al. 2009b), Ukrainian online dictionaries (Shyrokov et al. 2009), and the Slovene semantic lexicon (Fišer, Erjavec 2009).

Sub1Nom + Verb [edit]

ak ma oči neklamali | čo oči nevidia, srdce nebolí | div mu oči nevypadli | ide, kam ho oči vedú | má, čo oči vidia | oči ho bolia | oči ho pália | oči ho prezradili | oči ho štípu | oči mu behajú | oči mu blčali | oči mu išli vyskočiť z jamôk | oči mu padli na niečo | oči mu slabnú | oči mu svietia | oči mu tancovali | oči mu zahoreli láskou | oči mu zažiarili | oči mu zvlhli | oči mu žiaria | oči sa im stretli | oči sa mu jagali | oči sa mu naplnili slzami | oči sa mu rozšírili | oči sa mu smejú | oči sa mu zaiskrili | oči sa mu zapichli do niekoho, do niečoho | oči sa mu zatvárajú | oči sa mu zúžili |

Atr + Sub1Gen [edit]

|

Sub2 + Sub1Gen [edit]

farba očí | líčenie očí | pohľad očí | pokožka okolo očí | prevracanie očí | únava očí | vrásky okolo očí | výraz očí | začervenanie očí |

Verb + Sub1Gen [edit]

biť do očí | byť v niekoho očiach nejaký | civieť niekomu do očí | dívať sa niekomu do očí | hľadieť niekomu do očí | hľadieť smrti do očí | Choď / Hybaj / Prac sa / ... mi z očí! | klamať niekomu do očí | mať sto očí | napľuť niekomu do očí | nechať niekomu

Example from the dictionary of Slovak collocations – word ***oko***

Electronic dictionaries are capable of meeting users' requests many times faster than paper dictionaries, as well as of providing the possibility to locate all entries whose headwords satisfy user-defined criteria. Despite the fact that dictionary entries resemble a text on the screen, the internal representation of electronic dictionaries is a database.

The use of modern database technologies for fast access to dictionaries requires careful design and implementation of an underlying data structure and storage. The LDB has to meet the following requirements:

- to be a web based database with queries performed not just on lemmata, but also on inflected wordforms, in order to easily reach the intended audience using existing, standard software components;
- to include links to various entry-related information in external databases (such as morphological paradigms);
- to enable easy online updating and editing by multiple editors;
- to keep track of revision history, with the possibility of roll-back.

These points can be partly met by using advanced wiki-based collaboration editing systems. We recommend unifying the classifiers of the headword in the dictionary entry. The headwords in the dictionary entries of the digital dictionary must be indexed according to the number of lexical meanings, and each meaning must be unambiguously related to the form. In this manner most meanings of the form can be encompassed. Such a description might require more classifiers, but also

provide a more adequate correspondence. We recommend unifying the systems of categories and tags used for annotation in the various systems.

The project also dealt with the ***representation of semantics problems in Slavic digital lexicography*** in the context of the very substantial growth of digital dictionaries of all types. Information technologies offer new possibilities for lexicographers, namely: easy and fast addition of new dictionary entries, enrichment of their content by supplementary information on the headword (grammar, word-formation class, etymology, usage, etc.), and examples (e.g. for clarification of individual usages), phrases and collocations, idioms, etc.

The development of innovative solutions for lexical entry content in Slavic lexicography is a challenging task despite the broad and theoretically well-developed knowledge area of linguistics (Simov, Osenova 2009). Determining the content of a lexical entry in bi- and multilingual digital dictionaries is a complex endeavour which has to deal with the description of linguistic forms with various meanings in the languages concerned (Koseska 2009b).

The difficulty stems from the fact that so far, the starting point for language description has been the form rather than the content. Distinguishing between the form and its meaning in comparing the material from six languages that belong to three different groups of Slavic languages (as is the case in the MONDILEX project) will help avoid numerous substantial mistakes and erroneous conclusions.

To achieve this goal, MONDILEX concluded that a *semantic interlanguage* or a *dictionary/lexicon of concepts* be developed, on which multilingual dictionaries should be based.

To this end, the role of *semantic interlanguage in contrastive studies* was investigated (Koseska 2009a). A language used for comparing two or more natural languages – the *interlanguage* – was proposed and analysed in detail. Contrastive linguistics is a field of synchronous linguistics with both theoretical and practical applications. When contrastive studies deal with analysing differences and similarities for practical purposes (didactic or translation-related ones), we could refer to them as a field of applied linguistics, connected first of all with teaching of foreign languages. On the other hand, theoretical contrastive studies are related to universal linguistic issues and use methods of language studies aimed at isolating from languages the elements which are either common or different for them.

With respect to research methods used, theoretical contrastive studies are close to typological studies, but differ from the latter in the aim of description.

The *interlanguage* is not only related to theoretical contrastive studies (Koseska, Korytkowska, Roszko 2007). For that reason, development of such a language is an extremely difficult task, even if we are comparing only two languages. An equally difficult task is a description leading from analysing the content plane towards formal analysis of the considered languages, but such a description guarantees the maximum advantage for the recipient.

In order to separate descriptive representations of an individual language from contrastive descriptions, it was necessary to clearly distinguish between the notions of a metalanguage describing a single language from that of an interlanguage, which constitutes a tool for comparing at least two language systems. Thus the notion of a metalanguage differs from that of an interlanguage first of all in the fact that

a metalanguage is used for describing one given language, while an interlanguage is a tool for comparing at least two language systems. In this approach, it is also a semantic language, which consists of semantic categories and notions necessary for their description. It is worth noting that an interlanguage keeps developing and acquiring new notions as the research progresses.

The *Universal Dictionary of Concepts* (UDC: Boguslavsky, Dikonov 2008) is a language-independent intermediary lexical tool developed as a part of the effort to create a semantic language for global information exchange. It can evolve into an open and freely available language-neutral resource, a tool to uniformly record and link meanings of words of different languages and help the creation of bi- and multilingual dictionaries.

The making of dictionaries which would link the vocabularies of any natural languages with UDC and the pivot Universal Networking Language (UNL) is explained, using Russian and English as case studies. UDC is a repository of concepts forming the lexicon of the UNL.

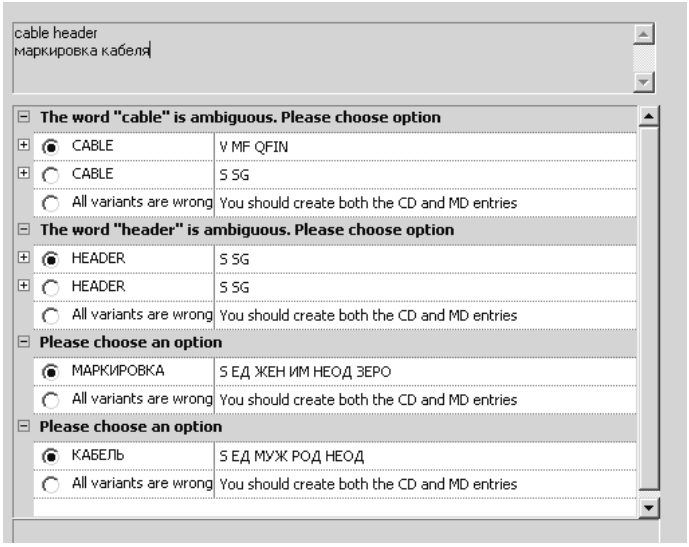
The following examples show representing semantics in the English-Russian Combinatory Dictionary: a universal zone of an English Combinatory Dictionary entry and a universal zone of an Russian Combinatory Dictionary entry.

```
1 ACCUSATION
2 POR:S
3 SYNT:VOC,COUNT
4 DES:‘ФАКТ’,‘ДЕЙСТВИЕ’,‘АБСТРАКТ’
5 D1.1:BY1
6 D1.2:OF,‘ЛИЦО’
7 D2.1:AGAINST
8 D3.1:OF,‘ФАКТ’
9 _V0:ACCUSE
10 _SYN1:CHARGE3
11 _S1:ACCUSER
12 _ANTI:JUSTIFICATION
13 _MAGN:GRAVE3
14 _VER:JUST2/WELL-BASED
15 _ANTIVER:FALSE/GROUNDLESS/UNFOUNDED/BASELESS/UNJUST
16 _OPER1:MAKE1/BRING
17 _FINOPER1:DROP2
18 _REAL1-M:PROVE/SUBSTANTIATE
19 _OPER2:BE<UNDER1>
20 _REAL2-M:DENY/REFUTE/REPUDIATE
21 _ANTIREAL2-M:ADMIT
22 _CAUSFUNC1:LAY1/LEVEL2
```

Universal zone of an English CD entry

```
1  ОБВИНЕНИЕ1
2  COMMENT:‘‘ВЫСКАЗЫВАНИЕ МНЕНИЯ О ЧЬЕЙ-ЛИБО ВИНЕ’’
3  EXAMPLE:‘‘ОБВИНЕНИЕ В ХАЛАТНОСТИ”
4  POR:S
5  SYNT:СРЕДН,ИСЧИСЛ
6  DES:‘ДЕЙСТВИЕ’,‘ФАКТ’,‘АБСТРАКТ’
7  D1.1:ТВОР,‘ЛИЦО’
8  D1.2:РОД,‘ЛИЦО’
9  D2.1:РОД
10 D2.2:ПРОТИВ1
11 D3.1:В2
12 _SYN1:ОБЛИЧЕНИЕ
13 _ANTI:ОПРАВДАНИЕ2
14 _VO:ОБВИНЯТЬ
15 _S1:ОБВИНИТЕЛЬ
16 _S2:ОБВИНЯЕМЫЙ
17 _MAGN:СУРОВЫЙ/ТЯЖКИЙ
18 _VER:ОБОСНОВАННЫЙ/ПРАВИЛЬНЫЙ/СПРАВЕДЛИВЫЙ
19 _ANTIVER:ЛОЖНЫЙ/НЕОБОСНОВАННЫЙ/НАПРАСНЫЙ/ПУСТОЙ
20 _OPER1:ВЫДВИГАТЬ/ПРЕДЪЯВЛЯТЬ/БРОСАТЬ2
21 _SO_INCEPOPER1:ВЫДВИЖЕНИЕ
22 _INCEPOPER1:ВЫДВИГАТЬ
23 _FINOPER1:ОТКАЗЫВАТЬСЯ<ОТ>/СНИМАТЬ1
24 _SO_FINOPER1:ОТКАЗ1<ОТ>/СНЯТИЕ
25 _SO_OPER1:ПРЕДЪЯВЛЕНИЕ
26 _OPER2:ПОДВЕРГАТЬСЯ
27 _REAL1-M:ДОКАЗЫВАТЬ
28 _REAL2-M:ОТВЕРГАТЬ/ОТМЕТАТЬ/ОТКЛОНЯТЬ
29 _ANTIREAL2-M:СОГЛАШАТЬСЯ2<СЗ>/ПРИЗНАВАТЬ
30 TRAF:АГЕНТ.10
31 TRAF:1-КОМПЛ.20
```

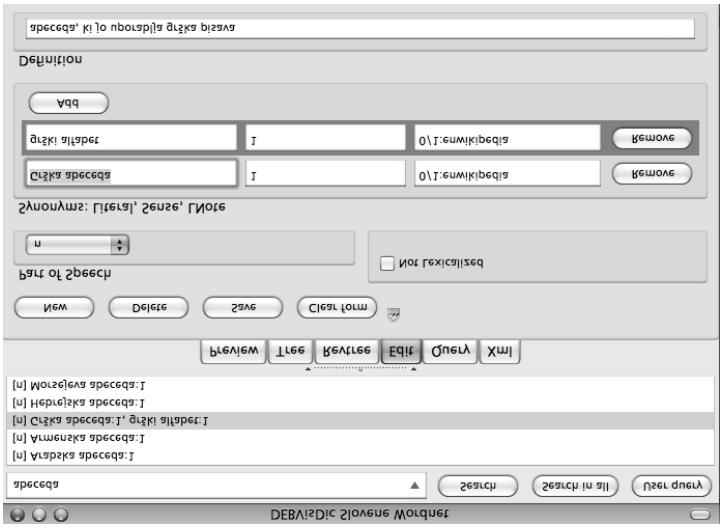
Universal zone of a Russian CD entry



Interactive window of the CD for the user entering the pair
cable header ⇔ *маркировка кабеля*

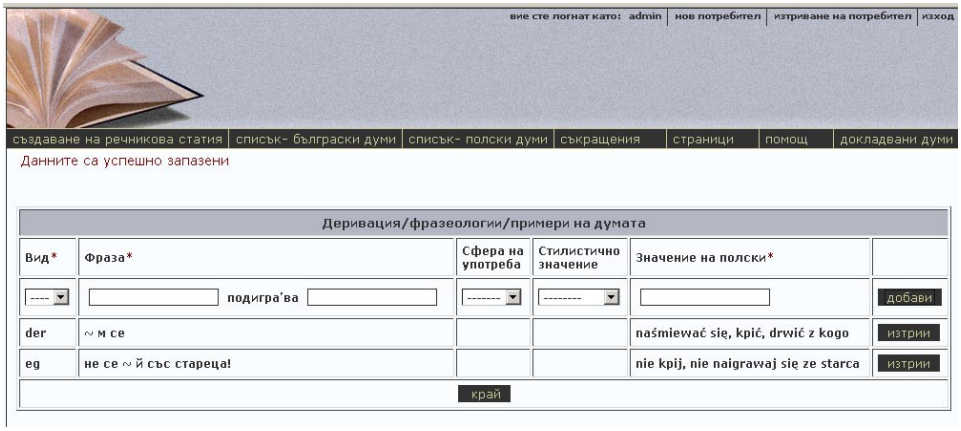
A third semantic problem MONDILEX discussed is the representation of temporal situations and some issues of the description of modality, using the Petri nets formalism. This formalism allows a partial rather than complete ordering of mutually independent events and states and coexisting and mutually exclusive states with different histories in one model. It can show the temporal relations in compound sentences and the complex manifestations of modality in language, and model conditionality better than logical implication does and is useful for creating new classifiers in dictionary entries related to time, so as to render the content as well as the form. A *catalogue of descriptions of temporal and modal situations* (Koseska, Mazurkiewicz 2010), expressed in different languages, was published in Warsaw. The entries in this catalogue are parameterized names of temporal and modal situations, and the corresponding values precise formal descriptions of such situations. The catalogue contains a collection of studies on temporal subjects, analyzed in accordance with the methodology of cognitive linguistics. The catalogue can be used to create a language-independent list of basic temporal situations.

The representation of semantics problems in Slavic digital lexicography are discussed widely. Two very different types of linguistic resources, textual corpora and lexical resources can be interrelated and enhanced through *semantic concordances* (Fišer, Erjavec 2009), in which words from the corpus are connected with their meanings specified in a semantic lexicon. Semantic concordances are a useful resource for a wide range of applications, such as automatic word sense disambiguation or corpus-based studies of sense frequency, distribution and co-occurrence, and are also invaluable as an aid for translation as well as for vocabulary acquisition in a foreign language. Some suggestions of simplifying and improving the manual annotation process in the future and further research directions into leveraging manual work in order to eventually automate the semantic annotation of corpora are presented.



Slovene synsets in DEBVisDic

Problems of representation of semantic features of the headword in a digital dictionary entry are discussed in parallel with the description of the design and development of an experimental *Bulgarian-Polish online dictionary* (Dimitrova et al. 2009c, Dimitrova, Koseska 2009a).



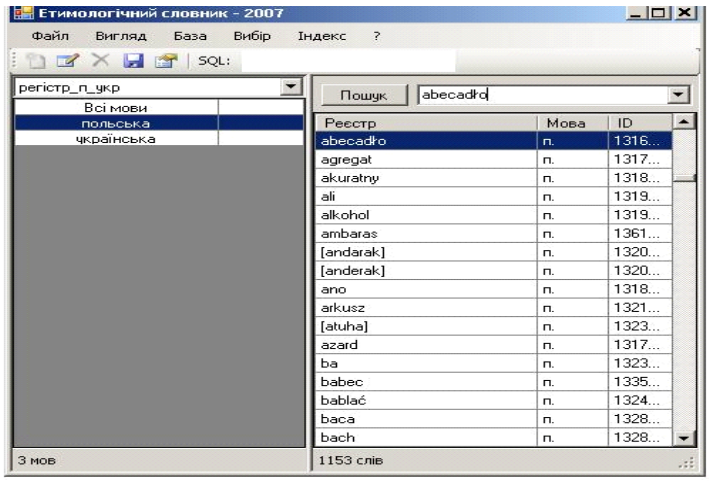
Bulgarian-Polish online dictionary - administrative panel for adding derivations, phrases and examples for the specific headword

Every dictionary entry is a structured object which uses different abbreviations and structural units in order to present the whole information succinctly. The external structure (presentation of text) does not completely determine the internal structure (information content in the database). This makes the database supporting the dictionary logically complex and difficult to create. The structure and content tags of the designed structural unit should fully meet international standards so that the LDB and the electronic dictionaries are made compatible with language resources created in other projects and for other languages.

Further issues are phraseology and etymology as separate domains of language description requiring specific linguistic research. The development of a lexical database of *Slovak language collocations* (Ďurčo 2007, Ďurčo et al. 2009), was presented. This lexical database should cover collocation profiles of several hundred words of different parts of speech and will be a base of a modern collocation dictionary. The database is built using the MediaWiki engine, which offers excellent remote collaboration features along with automated processing possibilities. The standard use of corpora for linguistic research and lexicography is aimed predominantly at the examination of occurrences and co-occurrences of word forms and lemmata. The main goal is to acquire data about semantic, grammatical and combinatorial behaviour of words.

Idiom variability is presented and discussed (Parizoska 2009). In cognitive linguistics most idioms (multiword units which have figurative meanings and relatively stable forms) are considered to be motivated by various cognitive mechanisms which link the meaning of idioms with the meanings of their constituents.

Problems accompanying the development of *digital etymological dictionaries* as a special case of digital lexicography were presented (Ostapova 2009). An etymological dictionary typically contains a large amount of lexical material from many languages with different kind of scripts, and from different periods. A special kind of software for processing such material is needed, including the tool for preparing a register of foreign words and an effective instrument for flexible search in dictionary.



Main user interface for the Etymological dictionary of Ukrainian

MONDILEX described the following set of **corpora** as resources for digital lexicography: *multilingual* parallel and annotated corpora – Bulgarian-Polish (Dimitrova, Koseska 2009b), Polish-Ukrainian (Shyrokov 2008, Shyrokov et al. 2005), and *monolingual* – morphologically and syntactically tagged corpus of Russian SynTagRus (Apresjan et al. 2006, Boguslavsky et al. 2009), Slovene language corpus with semantic annotation (Erjavec, Krek 2008, Krek, Erjavec 2009).

```
<tu tuid="0000000032">
    <tuv xml:lang="Polish">
        <seg>Oczekiwałem zapowiedzi startu, jakichś sygnałów, nakazu przypięcia
            się pasami, nic jednak nie nastąpiło.</seg>
    </tuv>
    <tuv xml:lang="Bulgarian">
        <seg>Очаквах команда за старт, някакви сигнали, заповед за стягане
            на коланите - нямаше нищо подобно.</seg>
    </tuv>
</tu>
<tu tuid="0000000033">
    <tuv xml:lang="Polish">
        <seg>Po matowym suficie zaczęły biec z przodu w tył niewyraźne cienie,
            jakby sylwetki wystrzyżonych z papieru ptaków.</seg>
    </tuv>
```

```
<tuv xml:lang="Bulgarian">
    <seg>По матовия таван започнаха да прелитат отпред-назад някакви
        неясни сенки, сякаш силуети на изрязани от хартия птици.</seg>
</tuv>
</tu>
```

Example from the Bulgarian-Polish aligned corpus

MONDILEX evaluated the application potential of various software environments for digital lexicography (for creating digital corpora and digital dictionaries).

Since modern dictionaries are almost universally collaborative projects involving many contributors, the choice of the working environment is subject to several requirements – easy remote editing, access control list, revision history, communication between editors. These requirements can be easily met by deploying wiki based software. The most relevant required features of a wiki system are:

- efficient indexing and searching,
- full Unicode support, with only some limitations concerning right-to-left scripts (irrelevant for Slavic languages) acceptable,
- full editing history with backup of page revisions, allowing to see the complete history of previous entry versions,
- review of differences between arbitrary page versions, using user-friendly output,
- multiuser support with full access control list,
- warnings to avoid editing conflicts, in case when two users intend to edit the same entry simultaneously.

There are many different wiki engines in use, mostly available under OpenSource license, but two of them are actually deployed for lexicographic purposes. One of them is MediaWiki, software that stands behind well known Wikipedia project. It is a complete and full featured, though rather complex system, with a difficult installation process and heavy software dependencies. MediaWiki is written in the PHP programming language and has many attractive options for the intended purposes, among them the possibility to use templates (a kind of macro) for better handling of repeating text parts. Templates are basically predefined text snippets in wiki-format with additional specialized markup for accommodating passing of arguments which are dynamically loaded inside another page.

The other is MoinMoin, very successful software written in the Python programming language, and as such particularly interesting because of the ease of installation, usage and customisation. MoinMoin is a wiki written completely in the Python programming language, using flat text files as a storage backend, rather than a database. This makes it particularly attractive for the needs of digital lexicography, because of the programming language involved and the ease of making

various data modifications and extraction, using just common text processing tools. MoinMoin is also fully Unicode aware, and all the stored data, output and input is invariably in UTF-8 encoding.

Among the described tools, there is a platform for research infrastructure in digital lexicography, namely the so-called *virtual lexicographic system.* Aspects of Web presentation and the impact of research infrastructure for digital lexicography are discussed.

3. Standardisation of Slavic Lexicographic Resources

Slavic languages are well known for their complex inflectional morphology. In order for Slavic digital lexicography to be made operational in a unified framework, it is desirable that a harmonised set of morphosyntactic features and morphosyntactic descriptions be used for all languages. On the one had such features are used to describe lexical and the inflectional properties of lemmas and their paradigms in lexica of Slavic languages, on the other, corpora of Slavic languages are annotated with tagsets of morphosyntactic descriptions.

3.1. Morphosyntactic Annotation in Slavic Digital Lexicography

MONDILEX discussed morphosyntactic annotations in Slavic digital lexicography. MULTEXT-East[3] (MTE) morphosyntactic specifications, and especially standardisation of Slavic lexicographic resources and their metadata were discussed and described in full with an emphasis on the importance of the developed harmonised lexical specifications in CES format (Ide 1998) and of the language independence of the tools. The MTE language resources, a multilingual dataset for language engineering research and development, focused on the morphosyntactic level of linguistic description. This standardised and linked set of resources covers a large number of mainly Central and Eastern European languages and includes the EAGLES-based morphosyntactic specifications; morphosyntactic lexica; and annotated parallel, comparable, and speech corpora. The MTE morphosyntactic specifications are a TEI P5 document that provides the definition of the attributes and values used by the various languages for word-level syntactic annotation, i.e. they provide a formal grammar for the morphosyntactic properties of the languages covered. In addition to the formal parts the specifications also contain commentary, bibliography, etc. The MTE specifications define 12 categories (mostly corresponding to parts-of-speech), each of which then defines its attributes and their values and the languages that each particular attribute-value pair is appropriate for. The morphosyntactic specifications also define the mapping between the feature-structures and morphosyntactic descriptions (MSDs), which are compact strings used in the morphosyntactic lexica and for corpus annotation.

MONDILEX discussed also the Text Encoding Initiative recommendations, an XML-based framework suitable for encoding a wide variety of text types, from those constituting digital libraries, to machine readable dictionaries, and annotated corpora; e.g. a TEI based encoding for linguistic annotation of corpora is now being

[3]The EU COP Project 106 MULTEXT-East *Multilingual Text Tools and Corpora for Central and Eastern European Languages,* http://nl.ijs.si/ME/

proposed in the scope of CLARIN[4] initiative. TEI is also suitable for encoding machine readable dictionaries, however, TEI does not have a module for lexical databases, but a model for those has been recently proposed as the ISO standard LMF, "Lexical Markup Framework".

In addition, MONDILEX made a proposal for lexical encoding concentrating on morphological properties of words, esp. of the strongly inflecting Slavic languages. The format of this encoding is an application of the new ISO standard LMF; the core lexical structure and morphosyntactic annotation are from the COP Project 106 MULTEXT-East, with recent extensions for Slovene.

The first version (realised 17 December 1997) – Specifications and Notation for Lexicon Encoding – was prepared in the framework of the MTE project. The specifications covered Bulgarian, Czech, Estonian, English, Hungarian, Romanian, and Slovene. Version 2 added morphosyntactic specifications for Serbian, Croatian, and the Resian dialect of Slovene. Version 3 of MULTEXT-East resources, *TELRI-CONCEDE edition*, brings together TELRI and CONCEDE projects' releases, makes them available in TEI P4 XML, and introduces further extensions. The fourth release of these resources was recently developed and introduces XML-encoded morphosyntactic specifications, using the latest version of the Text Encoding Initiative Guidelines, TEI P5 (TEI, 2007). This edition adds Macedonian, Polish, Russian, Slovak, Ukrainian, and Persian. The specifications now cover 10 Slavic languages, providing a good basis for a unifying morphosyntactic framework for digital Slavic lexicography (Erjavec 2010). The resources are available at http://nl.ijs.si/ME.

4. Recommendations

Lexicographic resources, in particular machine readable dictionaries, lexical databases, and mono- or multilingual annotated text corpora are developed and stored in a variety of formats, which makes them difficult to process on a common platform and to achieve interchange between programs and applications. The effectiveness of language technologies ultimately depends on the quantitative and qualitative parameters of the lexicographic description of units, relations and levels of language on which these technologies are based. This section proposes several mutually reinforcing recommendations which can serve to overcome this obstacle. All the proposed frameworks have already been extensively tested in practice and, in certain cases, further developed in the scope of the MONDILEX project.

The work of the project demonstrates the potential for developing useful lexicographic reference works (both digital and hardcopy) by using the format of the lexical data base and an adequate mathematical foundation. Various parameters of classification of the lexicon are likely to emerge in the process of developing the lexical data base, possibly through distributed effort, which highlights the importance of the interface to the lexicographic system. The lexical data bases forming the foundation of the dictionaries should be brought in line with one another by sharing theoretical concepts and platforms.

[4] http://www.clarin.eu

4.1. Morphosyntactic specifications

The MTE specifications provide a well-defined and powerful framework for expressing morphosyntactic features, which is now also instantiated for most Slavic languages. The MTE attributes and their values could sensibly be linked to other related attempts at standardisation of morphosyntactic features, in particular the ontology for descriptive linguistics GOLD[5] and the ISOcat Data Category Registry[6]. Given that this effort is well advanced, and that (morphosyntactic) terms are extensively documented, also with references to literature, it would be interesting to link the categories, attributes and their values from the MTE specifications to GOLD, providing an explanation of their semantics.

4.2. Corpus Storage and Processing

Regarding the storage and processing of corpora, there are several issues that need to be addressed. Corpora can be rather large – a medium-sized corpus today represents between 50 and several hundreds of gigabytes, either monolithic or (typically) split into many individual files with their own metadata sections.

While it is planned that each contributing organization will store the original versions of contributed corpora on their servers – either on one machine or in a distributed fashion, using metadata servers to find and access the correct files – a system of data pools and replica servers must be established to alleviate the load on the servers and provide for data consistency and availability, enabling uninterrupted access to the data.

For the purpose of corpus processing, the data from corpora must be transformed and often both intermediate and final versions of the data have to be stored on disk at least temporarily. This poses two problems: individual computing nodes have to have several gigabytes of storage available and an additional considerable amount of possibly temporary grid storage has to be available for the final datasets.

While the amounts of data needed for **H**uman **L**anguage **T**echnologies (HLT) tasks are entirely manageable using existing middleware and grid practices, a simple but powerful method for streamlining this procedure has to be put in place to simplify the process and to maintain integrity and availability of the data using central metadata servers, data pools and replicas.

The following recommendations are made:

- Unification of the classifiers of the headword in the dictionary entry. The headwords in the dictionary entries of the digital dictionary must be indexed according to the number of meanings, and each meaning must be related unambiguously to the form. In this manner most meanings of the form can be encompassed. Such a description might require more classifiers, but also provide a more adequate correspondence.

- Unification of the systems of categories and tags used for annotation in the various systems.

[5] http://linguistics-ontology.org/gold.html
[6] http://www.isocat.org/

- A uniform presentation of the lexical entry content.
- Creation of a corpus linguistics terminology database of all languages of the MONDILEX project (including English). The database should contain entries in Bulgarian, English (added as a hub language, and also because most terminology originates in English), Polish, Russian, Slovak, Slovene, and Ukrainian. The database aims to unify existing terminology. It can serve as a nucleus of a multilingual terminology database of lexicographic (or even general linguistic) terms.
- Creation of a special digital lexicographic environment adapted to the LDBs and digital dictionary entry structures and oriented to the creation of a multilanguage index in the automatic mode is necessary.
- Standard format of available corpus data. Additionally, linguistic annotations, such as morphosyntactic (or POS) tagging, alignments, chunking etc., have to be documented and standardized to the point where transformations between language-specific features of different corpora are possible. This compatibility is crucial for any advanced application, such as for parallel evaluation, compilation of WordNets, multi-language corpus alignment etc.

5. Grid – a technological platform for a future implementation

MONDILEX investigated the features of Grid as a technological platform for implementation of a network of centres for research in Slavic lexicography and their digital linguistic resources according to the specific requirements of its functionalities. This task is related to innovative technological solutions, which can be attained by the consortium's joint effort and will contribute to conceptual design studies for new research infrastructures of European character and relevance. The motivation was based on the fact that Human Language Technologies and related disciplines such as digital lexicography increasingly rely on large annotated corpora as a basic source of data, serving such needs as datasets for training and testing language models or for lexical investigations based on naturally occurring data. In view of the above, it is quite natural that albeit slowly and with some time lag as compared to other areas, the application of the grid paradigm has started to the area of HLT, especially to subareas that deal with the processing of large amounts of data (corpora).

MONDILEX concluded that the dynamic nature of the dictionary admits a relatively easy adaptation of the lexical database to any updated model of dictionary entry such as: addition of new types of information, improvement of the system of classifiers used for structuring the dictionary entry in order to describe the headword optimally, acquisition of digitally presented information for the creation of a new digital dictionary (e.g. a multilingual one). In addition to requiring large amounts of storage and computing power, lexicographers can also benefit from sharing the resources, corpora included. Of course, due to copyright and other factors, such sharing must be controlled via a system of access rights and permissions. So the grid aspects of enabling a distributed infrastructure for corpus processing

should include the establishment of a virtual organisation, rights and metadata management and corpus storage and processing.

5.1. Virtual research infrastructure

The Grid computing technology, as a form of distributed computing where a "virtual supercomputer" is composed of a cluster of networked, has been applied to computationally-intensive problems, requiring the storing and sharing of large amounts of data, in many areas of science. Some domains of applications such as processing data from medical records) demand a high level of data protection and controlled access. User authentication and digital rights management is part of the grid infrastructure. Because of this overlap of requirements, this paradigm has started being applied to the area of Human Language Technologies, especially to areas which deal with large amounts of data, i.e., with corpora (Javoršek, Erjavec 2009). While virtual organizations in modern grids are self-contained infrastructure elements, they must be included in the common infrastructure of all sites supporting the virtual organization.

The key points of *Grid infrastructure requirements* needed for supporting research activities in digital lexicography that could be mentioned here, are: virtualization techniques, specific legal issues (the data to be processed are in most cases copyrighted, and the research institutions either have very strict legal agreements governing the use of the data, or are operating entirely on copyright law sections allowing scientific and research use of the data), security measures used in the Grid infrastructure, such as public key infrastructure, virtual organizations, proxy certificates, and data protection (Erjavec, Javoršek 2008).

5.2. Establishment of the virtual organisation supporting human language technologies on grid

In order to provide the power of grid computing to researchers in the domains of digital lexicography, corpus processing and human language technologies in general, the technology needs to be accessible as a part of dedicated grid infrastructure (Garabík et al. 2009). Luckily, modern grid infrastructures support this approach in the form of Virtual Organizations (VOs), self-contained infrastructure elements that provide authorization management, software distribution, tools development and organizational support for a project or disciplinary community in the grid.

A number of steps are described here that should be taken to provide this service to the community.

5.2.1. Creation of Core Services

To support the Human Language Technologies Virtual Organization (HLT VO), a Virtual Organization Membership Service (VOMS) server to provide VO user and service access control has been set up. This is the central server for the Virtual Organization user and server access control, including accreditation, authentication and authorization. To use the server, a user (organization or person) has to get a grid digital certificate for authentication and use the server to apply for accreditation.

To support the virtual organization, any organization can include the HLT VO VOMS configuration in its authorization control set-up, thus allowing a combination of local and VO controls to govern access to data and services of HLT VO members. HLT VO VOMS is supported by the SiGNET cluster. Any organization wanting to participate in the HLT VO can enroll with the VOMS to use the infrastructure and include its configuration in the local set-up to support the infrastructure locally.

In order to support distributed data management and access, a central metadata server will have to be established. While existing solutions for grid infrastructure can be used for mappings from grid names to local file names and distributed data management, a solution for extensive corpora metadata management and mapping will have to be evaluated and developed to enable meaningful querying and access to corpora from linguistic tools.

5.2.2. Registration of the VO

While Virtual Organizations in modern grids are self-contained infrastructure elements, they have to be included in the common infrastructure of all sites supporting the Virtual Organization (VO). Two different grid middleware solutions shall be supported: NorduGrid and gLite.

NorduGrid ARC is a good match for applications that, in grid terms, are not very resource intensive and is also easier for setting up new sites due to much simpler installation and integration procedures.

gLite from the EGEE project is, on the other hand, the most widely used and supported middleware and therefore has to be supported by the HLT VO. As soon as HLT VO is registered, it will be discoverable using the central services of EU Grid infrastructure (i.e. with the EGEE and NorduGrid projects). It is also expected to become one of the supported VOs in the future European Grid Initiative (which is to start its operations in 2010).

After the VO is registered as a member of the EGEE project, support for the widely used gLite grid middleware should be included in the system – so far only the easier-to-use and more efficient NorduGrid ARC has been supported. For NorduGrid ARC, sites that already use it can start supporting the new VO simply by editing the relevant setup files and installing the software base for the job execution environment from the VO repository.

5.2.3. Data and Metadata

Due to frequently imposed restrictions on the use of corpus data according to contracts regulating the use of copyrighted and other non-free materials, it is essential that a managed distributed data access be provided with a central metadata server and full support for VO-based access control and authorization. While no such solution has been implemented, it is an essential element of international collaboration.

A number of existing solutions for grid infrastructure has been tested and we recommend a metadata service on the base of the Arda Metadata Catalogue Project (AMGA) as a viable solution that could allow us to leverage rich metadata services

and grid access controls to enable linguistic researches to use the available resources while enforcing the legal restrictions in place.

5.2.4. Web interfaces and central services

A dedicated web site for information, documentation and user management of HTL VO should be set up. It will provide the central grid services for the virtual organization, such as basic task and job reporting, statistics of usage and meta-data access. The central infrastructure will be sufficient for initial testing and evaluation for Human Language Technologies Grid, but additional services will have to be developed to support web based job submission and control, data-set upload (including corpus upload, transformation etc.) and data retrieval from finished jobs. A number of these techniques have been already tried in the experiments.

We recommend expanding this effort to provide research community with a reliable basis for resource intensive NLP tasks in an EU Grid computing environment. One of the major attractions of the new system, next to the flexibility, compatibility of tools and the sheer computing and storage power, will be to provide a single method (and programming API) to many resources in different languages, and to resolve the difficulties inherent in different legal, technical and practical restrictions that make any multilingual research rather difficult today.

Some web-based interfaces to the resources incorporated in the grid shall also be added. The first of such planned services will be a grid-aware concordancer, accessible both as a web service and from grid jobs. The service will enable the user to access the available grid-based corpora according to user's authorization. For testing purposes, a set of command-line tools for submitting typical linguistic grid jobs should be developed, based on a basic set of tools that will be prepared for the use on the grid (gridified) for testing purposes. These tools will have either the form of dedicated scripts or specialized makefiles and will be able to perform a resource-intensive task using distributed corpus data and distributed computing resources in the HLT VO.

For enabling grid-based corpus processing, at least partially in the scope of the MONDILEX project, some central services shall be gridified, incl. annotation of corpora and term extraction.

5.3. Two examples of annotation tasks that can be easily helped by transfer to the Grid platform

When building a (huge) text corpus, one of the basic foundations (at least for synthetic languages) is a morphological (or part-of-speech) tagging and lemmatization. Usually very language specific, there are different algorithms and different methods applied. As a proof-of-concept, we run such annotation systems as used in Slovak and Slovene languages. Since form the application point of view the Grid is just an ordinary GNU/Linux distribution, there were no problems whatsoever (however, we used the installation in a chrooted environment).

Corpus annotation with ToTaLe:

The automated multilingual annotator ToTaLe, used for Slovene (Erjavec et al., 2005), is the program written in Perl, which implements the following annotation steps, in a multilingual setting:

1. tokenisation
2. part-of-speech tagging
3. lemmatisation

A plain Unicode (UTF-8) text is first tokenised, the word tokens (word-forms) are then tagged with their context-disambiguated part-of-speech, or, more accurately, morphosyntactic description (MSD), and the word-forms, given their MSD, are lemmatised to arrive at the canonical form of the word. The program can produce the output in several formats, in particular in tabular form or encoded in TEI-compliant XML. The tool has been extensively tested with TEI P5 encoded corpora and MTE tag sets.

Morphosyntactic annotation with morče:

Morphosyntactic tagging of the Slovak National Corpus consists of morphosyntactic analysis, where each word in the input texts is assigned a set of possible morphosyntactic tags by looking up the possibilities of lemma/tag combinations in a constant database table using the wordform as a key, with an additional step for unknown words, where the list of possible tags is derived from the similarities of word endings to the ones present in the database tables, and by following disambiguation, where one of the lemma-tag pairs is selected. The analysis is implemented in the Python programming language and is quite fast. On a reasonably recent hardware it is able to process over 10 000 words per second.

The disambiguationan averaged perceptron model (originally used for the Czech language tagging, re-trained on the Slovak manually annotated corpus) is used. Disambiguation speedreaches on average only about 300 words per second. Parallelization at the application level is not possible without some redesign of the *morče* itself, but the nature of tagging makes it easy to split the input data into as many chunks as desirable and run morphology analysis and disambiguation in many instantiations in parallel.

6. Socio-economic impacts of the project

Integration of the new EU countries and smaller economies within a European e-infrastructure framework promotes their involvement in European development and enables them to profit from the wide range of competencies across Europe. This process will also democratize the research and enable innovation independent of physical location. MONDILEX developed and promoted best practices and tools for Slavic languages resources exchange for the stimulation of sustainable collaboration and business models for research infrastructure utilization in the future.

The full spectrum of e-infrastructure, including data, networks, software and related competences, has to be supported in a balanced way to achieve efficiency in

building the ICT system supporting access to research infrastructures and sharing their research functions. MONDILEX concluded that closer collaboration between research communities and providers of e-infrastructure and related services needs to be promoted.

Tools and processes to manage data, promote interoperability, integrate databases and ensure access rights require significant development effort in order to promote sustainable services. European collaboration in the NLP area is very important because its contribution to improving the quality of language communication of EU citizens. In this respect the promotion of resource and tools exchange among member or non-member states should be piloted.

MONDILEX observed that managing and providing efficient access to data represent a major challenge and a crucial step for resolving the issue is a clear policy of access. Access to specific databases and repositories for research and development purposes and innovative aims should be considered attentively. Efficient transnational access to online digital content should be promoted.

The contributions of the European research community to the activities of MONDILEX project were presented in a series of five open MONDILEX workshops. The Proceedings of these events were first published on-line on the project Web site and subsequently printed and circulated to the libraries of institutions participating in the project, libraries of national academies of sciences, national and university libraries, as well as disseminated among the scholarly community, universities, business, potential partners and users of the future research infrastructure.

7. Conclusion

The project MONDILEX provided 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. It demonstrated that 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.

The implementation of a Research infrastructure for Slavic lexicography will contribute to the development of a knowledge society, not only by carrying out research, but also through the combination of various expertises from different backgrounds, from development of communication capacities and strengthening the interaction between research and society. Access to and use of technologically well-equipped facilities or databases enables young researchers and students to undertake complex problems as part of high-level interdisciplinary teams, and qualifies them, in an outstanding manner, for tasks in science or industry, and fostering their career mobility.

Participation in the MONDILEX consortium enables the sharing of services for data processing and data collections, the coordinated extension and further development of bi- and multilingual lexical resources, so that researchers in the humanities and social sciences as well as education and business will be provided with an easy access to digital bi- and multilingual dictionaries of Slavic languages.

The MONDILEX project contributes to the preservation and support of the multilingual and multicultural European heritage.

The project also emphasized the important role of scientific collaboration in the development of digital language resources, online accessibility and digital preservation of European cultural heritage and collective memory.

It has laid foundations for further cooperation, setting up and elaborating a methodology of interaction of remote research groups and coordination of formats of lexicographic resources.

8. Acknowledgements

The MONDILEX Consortium thanks the European Commission for support through FP7 Grant Agreement 211938 *Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources*.

We would like to thank all colleagues from the six MONDILEX participants' teams from IMI–BAS (Sofia, Bulgaria), ISS–PAS (Warsaw, Poland), LSIL–SAS (Bratislava, Slovakia), J. Stefan Institute (Ljubljana, Slovenia), IITP–RAS (Moscow, Russia), and ULIF–NANU (Kiev, Ukraine) with whom we worked throughout the two years of the duration of the project.

We also acknowledge the valuable contribution of the MONDILEX experts Antoni Mazurkiewicz (Institute of Computer Sciences, Polish Academy of Sciences), Igor Boguslavsky (Russian Academy of Sciences), Jan Jona Javoršek (Jožef Stefan Institute, Slovenia), and Peter Ďurčo (St. Cyril and Methodius University, Slovakia).

Our special thanks to the members of MONDILEX Advisory Board and Consultative Body: Adam Przepiórkowski (Polish Academy of Sciences, CLARIN), Maciej Piasecki (Wrocław University of Technologies, CLARIN), Kiril Simov (Bulgarian Academy of Sciences, CLARIN), Juri Apresjan (Russian Academy of Sciences), Maria Šimková (Slovak Academy of Sciences), and Oleksandr Palagin (National Academy of Sciences of Ukraine).

Many thanks to all colleagues from the European research community that attended MONDILEX events in Moscow (2008), Kiev (2009), Bratislava (2009), Warsaw (2009), Ljubljana (2009), and Sofia (2008 & 2010).

References

Apresjan et al. (2006): Apresjan, Juri, 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, pages 1378–1381.

Boguslavsky, Igor, Dikonov, Vyacheslav (2008). Universal Dictionary of Concepts. In: Iomdin, Leonid, & Dimitrova, Ludmila (Editors, 2008). Lexicographic Tools and Techniques. *Proceedings of the MONDILEX First Open Workshop, 3-4 October 2008, Moscow.* Moscow, IITP–RAS, pages 31–41.

Boguslavsky et al. (2009): Boguslavsky, I., Iomdin, L., Frolova, T., Timoshenko, S. Development of a Russia Tagged Corpus with Lexical and Functional Annotation. In: Garabík, Radovan (Editor, 2009). Metalanguage and Encoding Scheme Design for Digital Lexicography. *Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009.* Tribun, Brno, pages 83–90.

Dimitrova et al. (2010): Dimitrova, L., Koseska–Toszewa, V., Garabík, R., Erjavec, T., Iomdin, L., Shyrokov, V. MONDILEX – Towards the Research Infrastructure for Digital Resources in Slavic Lexicography. In: *International Journal* Cognitive Studies/Études Cognitives. Vol. 10, SOW, Warsaw, pages 147–162.

Dimitrova et al. (2009a): Dimitrova, Ludmila, Panova, Rumyana, Dutsova, Ralitsa. Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In: Garabík, Radovan (Editor, 2009). Metalanguage and Encoding Scheme Design for Digital Lexicography. *Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009.* Tribun, Brno, pages 36–47.

Dimitrova et al. (2009b): Dimitrova, Ludmila, Koseska, Violetta, Dutsova, Ralitsa, Panova. Rumyana. Bulgarian-Polish online Dictionary – Design and Development. In: Koseska, Dimitrova, Roszko (Editors, 2009). Representing Semantics in Digital Lexicography. *Proceedings of the MONDILEX Fourth Open Workshop, 29 June – 1 July 2009, Warsaw.* SOW, Warsaw, 2009, pages 76–88.

Dimitrova et al. (2009c): Dimitrova, Ludmila, Koseska-Toszewa, Violetta, Satoła-Staśkowiak, Joanna. Towards a Unification of the Classifiers in Dictionary Entry. In: Garabík, Radovan (Editor, 2009). Metalanguage and Encoding Scheme Design for Digital Lexicography. *Proceedings of the MONDILEX Third Open Workshop, 15-16 April 2009, Bratislava.* Tribun, Brno, pages 48–58.

Dimitrova, Ludmila, Koseska, Violetta (2009a). Classifiers and Digital Dictionaries. In: *International Journal Cognitive Studies/Études Cognitives.* Vol. 9, SOW, Warsaw, 2009, pages 117–131.

Dimitrova, Ludmila, Koseska, Violetta (2009b). Bulgarian-Polish Corpus. In: *International Journal Cognitive Studies/Études Cognitives.* Vol. 9, SOW, Warsaw, 2009, pages 133–141.

Ďurčo, Peter (2007). Zásady spracovania slovníka kolokácií slovenského jazyka. Online documentation. Available from http://www.vronk.net/wicol/images/Zasady.pdf

Ďurčo, P. et al. (2009): Peter Ďurčo, Radovan Garabík, Daniela Majchráková, Matej Ďurčo. Dictionary of Slovak Collocations. In: Koseska, Dimitrova, Roszko (Editors, 2009). Representing Semantics in Digital Lexicography. *Proceedings of the MONDILEX Fourth Open Workshop, 29 June – 1 July 2009, Warsaw.* SOW, Warsaw, pages 128–137.

Ďurčo, Peter, Garabík, Radovan (2009). Slovak Paremiography Database. In: Erjavec (Editor, 2009), Research Infrastructure for Digital Lexicography. *Proceedings of the MONDILEX Fifth Open Workshop, 14-15 October 2009, Ljubljana.* Informacijska družba Publ. House, Ljubljana, 2009, pages 20–26.

Erjavec, Tomaž (Editor) (2009). Research Infrastructure for Digital Lexicography. *Proceedings of the MONDILEX Fifth Open Workshop, 14-15 October 2009, Ljubljana.* Informacijska družba Publ. House, Ljubljana, 2009, 131 pages.

Erjavec, Tomaž (2010). Multext-East: Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *LREC'10.*

Erjavec et al. (2005): Erjavec, T., Ignat, C., Pouliquen, B., Steinberger, R. Massive multi-lingual corpus compilation: Acquis Communautaire and totale. In: *Arch. Control Sci.*, vol. 15, pages 529–540.

Erjavec, Tomaž, Javoršek, Jan Jona (2008). Grid Infrastructure Requirements

for Supporting Research Activities in Digital Lexicography. In: Iomdin, Leonid & Dimitrova, Ludmila (Editors, 2008). Lexicographic Tools and Techniques. *Proceedings of the MONDILEX Open Workshop, Moscow, 3–4 October 2008.* Moscow, IITP–RAS, pages 5–14.

Erjavec, Tomaž, Krek, Simon (2008). The JOS morphosyntactically tagged corpus of Slovene. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation. Marrakech, Morocco, May 26-June 1, 2008. (LREC 2008)*, ELRA 2008.

Fišer, Darja, Erjavec, Tomaž (2009). Towards Semantic Concordances in Slovene. In: Koseska, Dimitrova, Roszko (Editors, 2009). Representing Semantics in Digital Lexicography. *Proceedings of the MONDILEX Fourth Open International Workshop, Warsaw, Poland, 29 June–1 July 2009,* SOW, Warsaw, 2009, pages 106–114.

Garabík, Radovan (2008). Storing morphology information in a wiki. In: Iomdin, Leonid & Dimitrova, Ludmila (Editors, 2008). Lexicographic Tools and Techniques. *Proceedings of the MONDILEX First Open Workshop, Moscow, 3-4 October 2008.* Moscow, IITP–RAS, pages 55–59.

Garabík, Radovan (Editor) (2009). Metalanguage and Encoding Scheme Design for Digital Lexicography. *Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009.* Tribun, Brno, 192 pages.

Garabík, Radovan, Špirudová, Jana (2009). Design of a New Slovak-Czech Lexical Database. In: Garabík, Radovan (Editor, 2009). Metalanguage and Encoding Scheme Design for Digital Lexicography. *Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009.* Tribun, Brno, pages 71–76.

Garabík et al. (2009): Garabík, Radovan, Javoršek, Jan Jona, Erjavec, Tomaž. Evaluating Grid Infrastructure for Natural Language Processing. In: Levická, J., Garabík, R. (Editors, 2009). NLP, Corpus Linguistics, Corpus Based Grammar Research. Tribun 2009, pages 93– 105.

Ide, Nancy (1998). Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In *First International Conference on Language Resources and Evaluation, LREC'98*, Granada, 1998, pages 463-470, ELRA. http://www.cs.vassar.edu/CES/

Iomdin, Leonid, Dimitrova, Ludmila (Editors) (2008). Lexicographic Tools and Techniques. *Proceedings of the MONDILEX First Open Workshop, 3-4 October 2008, Moscow.* Moscow, IITP–RAS, 109 pages.

Javoršek, Jan Jona, Erjavec, Tomaž (2009). Empowering Human Language Technologies with Grid. In: Erjavec (Editor, 2009). Research Infrastructure for Digital Lexicography. *Proceedings of the MONDILEX Fifth Open Workshop, 14-15 October 2009, Ljubljana.* Informacijska družba Publ. House, Ljubljana, 2009, pages 13-19.

Koseska-Toszewa, Violetta (2009a). Many-volume Contrastive Grammar of Bulgarian and Polish. In: Shyrokov, Volodymyr & Dimitrova, Ludmila (Editors, 2009). Organisation and Development of Digital Lexical Resources. *Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009.* Dovira Publ. House, Kiev, 2009, pages 87–97.

Koseska-Toszewa, Violetta (2009b). Form, Its Meaning, and Dictionary Entries. In: Garabík, Radovan (Editor, 2009). Metalanguage and Encoding Scheme Design for Digital Lexicography. *Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009.* Tribun, Brno, pages 105–111.

Koseska, Violetta, Dimitrova , Ludmila, Roszko, Roman. (Editors) (2009). Representing Semantics in Digital Lexicography. *Proceedings of the MONDILEX Fourth Open Workshop, 29 June–1 July 2009, Warsaw.* SOW, Warsaw, 2009,224 pages.

Koseska-Toszewa, Violetta, Korytkowska, Małgorzata, Roszko, Roman (2007). Polsko-bułgarska gramatyka konfrontatywna. Warszawa: Wydawnictwo Akademickie Dialog. (In Polish)

Koseska, Violetta, Mazurkiewicz, Antoni (2010). *Time Flow and Tenses.* SOW, Warsaw. 223 pages.

Krek, Simon, Erjavec, Tomaž (2009). Standardised Encoding of Morphological Lexica for Slavic Languages. In: Shyrokov, Volodymyr & Dimitrova, Ludmila (Editors, 2009). Organisation and Development of Digital Lexical Resources. *Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009.* Dovira Publ. House, Kiev, 2009, pages 24–29.

Levická, Jana (2007). Terminology and Terminological Activities in the Present-Day Slovakia. In: Computer Treatment of Slavic and East European Languages. *Proceedings of the Conference Slovko 2007.* Tribun, Brno, pages 139–151.

Levická, Jana (2008). Analysis of "classical" and legislative definitions for the term records of the Slovak terminology database. In: *Proceedings of the Third Conference on Translation, Interpreting and Comparative Legi-Linguistics.* Poznań, Poland.

Ostapova, Irina (2009). Digital Etymology (Illustrated by the example of the Etymological Dictionary of Ukrainian language): In: Shyrokov, Volodymyr & Dimitrova, Ludmila (Editors, 2009). Organisation and Development of Digital Lexical Resources. *Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009.* Dovira Publ. House, Kiev, 2009, pages 68–72.

Parizoska, Jelena (2009). Idiom variability in Croatian: the case of the container schema. In: Koseska, Dimitrova, Roszko (Editors, 2009). Representing Semantics in Digital Lexicography. *Proceedings of the MONDILEX Fourth Open International Workshop, Warsaw, Poland, 29 June–1 July 2009.* SOW, Warsaw, 2009, pages 123–127.

Simov, Kiril, Osenova, Petya (2009). Syntactic-Semantic Treebank for Domain Ontology Creation. In: Koseska, Dimitrova, Roszko (Editors, 2009). Representing Semantics in Digital Lexicography. *Proceedings of the MONDILEX Fourth Open International Workshop, Warsaw, Poland, 29 June–1 July 2009.* SOW, Warsaw, 2009, pages 115–122.

Shyrokov, V. A. (2008). Integral Slavic Lexicography in the Linguotechnological Context. In: Iomdin, Leonid & Dimitrova, Ludmila (Editors, 2008). Lexicographic Tools and Techniques. *Proceedings of the MONDILEX Open Workshop, Moscow, 3–4 October 2008.* Moscow, IITP–RAS, pages 23–30.

Shyrokov et al. (2005): Shyrokov V., Bugakov O., Griaznukhina T., Kostishin, O., Krygin, M., Lyubchenko, T., Rabulets, A., Sidorenko, O., Sidorchuk, N., Shevchenko, I., Shipnivska, O., Yakimenko, K. *Corpus Linguistics.* Kiev, 2005. 471 pages. (In Ukrainian).

Shyrokov et al. (2009): Shyrokov, V.A., Rabulets, O.G., Shevchenko I.V., Yakimenko, K.M. Integrated Lexicographic System "Dictionaries of Ukraine". CD ROM edition, Kiev, 2009. (In Ukrainian)

Shyrokov, Volodymyr, Dimitrova Ludmila (Editors) (2009). Organisation and Development of Digital Lexical Resources. *Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009.* Dovira Publ. House, Kiev, 2009, 128 pages.

Šimkova et al. (2009): Šimková, M., Garabík, R., Dimitrova, L. Design of a multilingual terminology database prototype. In: Koseska, Violetta, Dimitrova, Ludmila, Roszko, Roman (Editors, 2009). Representing Semantics in Digital Lexicography. *Proceedings of the MONDILEX Fourth Open Workshop, Warsaw, Poland, 31 May–2 June 2009,*

SOW, Warsaw, 2009, pages 123–127.
TEI (2007): TEI P5: Guidelines for Electronic Text Encoding and Interchange.

COGNITIVE STUDIES | ÉTUDES COGNITIVES, 11
SOW Publishing House, Warsaw 2011

Varia