MONDILEX

Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources

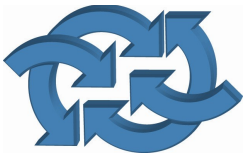
Institute of Mathematics and Informatics,
Bulgarian Academy of Sciences

Ludmila Dimitrova, Violetta Koseska, Radovan Garabík,

Tomaž Erjavec, Leonid Iomdin, Volodymyr Shyrokov

CONCEPTUAL SCHEME FOR A RESEARCH INFRASTRUCTURE SUPPORTING DIGITAL RESOURCES IN SLAVIC LEXICOGRAPHY

Sofia 2010



MONDILEX

Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources

Institute of Mathematics and Informatics,
Bulgarian Academy of Sciences
2010

The volume is the outcome of the efforts of the participants of the project **GA212938 MONDILEX** *Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources* and the financial support of the European Commission: **7th Framework Programme Capacities—Research Infrastructures** (Design studies for research infrastructures in all Sciences and Technologies fields).

2010
ISBN 978-954-8986-33-5

TABLE OF CONTENTS

Foreword

This volume is the outcome of the efforts of the participants in the project **GA212938 MONDILEX** *Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources* and the financial support of the European Commission: **7th Framework Programme Capacities—Research Infrastructures** (Design studies for research infrastructures in all Sciences and Technologies fields).

The MONDILEX project has six participants: (1) Institute of Mathematics and Informatics, Bulgarian Academy of Sciences (IMI-BAS), Sofia, Bulgaria, which coordinates the project; (2) Institute of Slavic Studies, Polish Academy of Sciences (ISS-PAS), Warsaw, Poland, (3) Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences (ĽŠIL), Bratislava, Slovakia, (4) Jožef Stefan Institute (JSI), Ljubljana, Slovenia; (5) Institute for Information Transmission Problems, Russian Academy of Sciences (IITP-RAS); and (6) the Ukrainian Lingua-Information Fund of the National Academy of Sciences of Ukraine (ULIF-NASU). The partners are research organisations from six European countries whose six national languages belong to the Slavic group: four EU members – Bulgaria, Poland, Slovak Republic, Slovenia, as well as the Russian Federation and Ukraine. All partners are national centres for research in linguistics, lexicography, and natural language processing.

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

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 the 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 bilingual 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 bilingual and multilingual dictionaries of Slavic languages. The MONDILEX project contributes to the preservation and support of the multilingual and multicultural European heritage. 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.

Ludmila Dimitrova (IMI-BAS)

MONDILEX coordinator

INTRODUCTION

The main aim of the MONDILEX project was to draft a sustainable and scalable infrastructure for institutions involved in creating and supporting a network of multilingual resources of Slavic languages. The MONDILEX project studied accordingly problems concerning the development, management, and reuse of lexical resources in a multilingual context because these play an essential role in a world of rapidly developing multilingual communication. Lexical resources provide information on many languages in a common framework and should be reusable in many automatic applications and human practices. Such resources include those developed along the lines of best practices and recommendations like monolingual and multilingual, parallel, comparable, and annotated corpora, monolingual and bilingual, traditional, electronic and online dictionaries, lexicons, thesauri, wordnets, ontologies, etc.

Large annotated corpora have recently gained importance as a source of data and especially as a foundation for adequate linguistic description. As they grow in quantity, size and variety, their integration and standardisation on the basis of common concepts and shared frameworks become critical. Automatic annotation tasks such as word alignment or semantic indexing are computationally very expensive. So are the investigations of today's lexicographers, who have to perform complex searches or other operations over large and heavily annotated corpora and so can benefit from sharing resources (storage and computing power), even though, due to copyright and other factors, such sharing must be controlled via a system of access rights and permissions.

As applied technological aspects become top priority for linguistic studies, the lexicographic description of the language system gains importance. The problem of multilinguality in the global information medium raises the question of an integrated lexicographic description of all languages. The effectiveness of linguistic technologies depends on the quantitative and qualitative parameters of the lexicographic description.

MONDILEX emphasised the importance of the developed harmonised lexical specifications in CES format and of the language independence of the tools. The use of annotated Slavic lexicographic resources with unified lexical descriptions is a contribution to the production of new bilingual and multilingual Slavic lexical resources and will open them up to the European academic community.

Another important objective of the MONDILEX project was to present recommendations for standardisation and integration of language resources.

These were the principal founding motives of the project MONDILEX, including the establishment of a highly efficient environment for creative interaction between researchers and practitioners in the linguistic disciplines.

The present volume consists of four parts and concluding remarks.

The first part describes different kinds of language resources – lexical databases, dictionaries, corpora, and grammars. First, some lexical databases are presented, namely a Slovak morphology database, multilingual corpus linguistics terminology database, Slovak-Czech lexical database, paremiography database, and Bulgarian-Polish lexical database. A presentation of Dictionary of Slovak collocations covering collocation profiles of several hundred words of different parts of speech and serving as a base of a modern collocation dictionary, a Bulgarian-Polish On-line Dictionary and a Ukrainian On-line Dictionary follows. Finally, corpora (monolingual corpus SynTagRus and multilingual parallel corpora MULTEXT-East and Bulgarian-Polish) and grammars are also described.

The second part is dedicated to the problems of standardisation of Slavic lexicographic resources and their metadata, among them standards for corpus encoding, machine readable dictionaries and lexical databases. Also, a proposal is made for a lexical encoding concentrating on morphological properties of words, esp. of the strongly inflecting Slavic languages. The format is an application of the new ISO standard LMF; the core lexical structure and morphosyntactic annotation are from MTE, with recent extensions for Slovene. A detailed representation of paradigms, regular derivation, variant spellings, etc. is also given. A universal networking language, a tool for global information exchange in computer networks, is presented.

The third part focuses on software environments for digital lexicography, primary on a conceptual modelling of services for the bilingual lexicographic systems and their integration with other services of the lexicographic systems. In addition, it contains a short presentation of various software environments for creating digital corpora and digital dictionaries (namely, MoinMoin and MediaWiki), and for automated database processing.

The fourth part describes shortly a technological platform for a research infrastructure for digital lexicography. The concept of a virtual lexicographic system is presented in details. Grid infrastructure requirements for supporting research activities in digital lexicography are discussed.

The section **Concluding remarks** discusses the impact of research infrastructure on digital Slavic lexicography.

Some recommendations for corpora annotation, lexical database structure and dictionary entry design and content are presented accordingly.

Part 1. Language Resources in a Research Infrastructure for Slavic Lexicography

1.1 Lexical Databases

1.1.1 Slovak morphology database

Although the primary purpose of the wiki is to keep the data for the automatized NLP processing purposes, the data is useful also as a reference database for dictionary-like queries, and therefore the design of the pages has been made with this goal in mind.

Basic unit of the wiki data is called a page (using MoinMoin terminology). Each page contains data pertaining to one lexeme, i.e. lemma with full paradigm and morphology annotation. Each page name is equal to the lemma, taking into account common capitalization of words in Slovak (proper nouns) (an important point, because by design the final morphology analyser disregards the capital letters and gives all the lemmas in lowercase). In case of lexical homonymy, pages are named by the lemmas with part of speech tag attached in parentheses (e.g. *mať_(V)* for a verb, *mať_(S)* for a noun). The page structure attempts to be both human-readable and human-editable and easily automatically parseable. Page body contains of several sections, the first one is the *Lema*, which contains just one word, the lemma. Then follows the *Paradigma* section, containing the inflectional paradigm spelt out in full. For each grammar category there is one corresponding line, with morphological tag separated from the form by a colon (:). Alternative forms per one grammar category can be either given on a separate line, or on the same line, separated by a comma (,). At the end of a page there is the part of speech category the described word belongs to.

Homonymy

Only the basic homonymy – where lemmas for two different words (two different parts of speech) are identical – is addressed by the database. The other forms of homonymy (inflectional) are automatically taken care of by keeping the homonyms under their corresponding lemmas and morphology tags. In case of part of speech homonymy, there is a special disambiguation page, linking to all the possible lemmas.

In Slovak, reflexive verbs are marked by a special separate morpheme *sa/si*, which is separated from the verb and has relative freedom of movement around the verb (Unlike other languages, e.g. in Russian the reflexive pronoun/particle takes a form of a clitic inseparably bound to the verb). As there exist a reflexive/non-reflexive dichotomy (i.e. reflexive verbs having almost always their non reflexive counter-

part), only the non reflexive parts in the dictionary, without the *sa/si* pronoun. Several singular cases of reflexive verbs without a meaningful standalone non reflexive counterpart (*smiať sa, báť sa, uvedomiť si, čudovať sa*) do not pose any problem – the missing *sa* is confusing only for the uninitiated users.

Traditionally, *sa* and *si* are called "reflexive pronouns" if semantically there is a discernible action performed on the agent (i.e. they can be seen as contractions of personal pronouns *seba* and *sebe*), otherwise they are considered to be a part of a verb. This is just a convention – they could be called equally well to be particles, indeed this is how they are sometimes classified in the traditional Czech grammars. In the database, they are assigned a special morphology tag **R**, regardless of their semantic use.

Statistics

Currently, the wiki contains 77567 entries (Garabík 2008). Categorised by the POS type, there is the following distribution:

28163	verbs
26061	substantives
13100	adjectives
5069	adverbs
1297	abbreviations
1104	participles
656	interjections
369	particles
369	pronouns
311	numerals
123	prepositions
110	conjunctions
72	citation elements (Note (1))
26	part of multiword expression (Note (2))
2	sa/si
1	*By* (Note (3))
716	disambiguation pages

Table 1: Distribution of parts of speech

Notes: (1) "Citation element" is a foreign language word appearing in Slovak text, e.g. most often in book or movie names, or French or Latin quotations. In this database, only a few such words are included. (2) Used to mark standalone morphemes that are a part of multiword expressions – these are in fact just a remnant of the

tokenization. (3) Special conditional morpheme, traditionally classified as a particle.

Scalability

As the total amount of entries in the database reaches tens of thousands, with the possibility of growth up to several times the number, it is important to achieve reasonable scalability of the wiki engine. Since the MoinMoin stores each page in its own directory and all the directories are stored under one parent directory, it is important for the underlying file system to be able to cope with many thousand entries per directory. All the major modern Linux file systems have no problems with this usage pattern, probably the best file system for this purposes at the moment is ReiserFS, which has also other convenient features, such as tail-packing to conserve disk space, since the files used by the backend storage are predominantly way below file system block size. Total size of the data is 1.2 GB of disk storage.

Basic usage works well, direct searching for a lemma, page editing, revision history and similar actions are performed without noticeable delays. However, the built in full text search engine is unable to cope with the amount of data, basic search for an inflected word form takes typically tens of minutes of 100 % CPU utilization. After the switch to the Xapian search engine, the search for a word form is instantaneous. However, other features that depend on number of pages are difficult to use, e.g. displaying all the pages in one category takes several minutes (much of the time is not due to searching, but to formatting such a huge number of links).

Usage

The wiki can be used directly, as a reference dictionary of inflectional data. However, the main use is mostly as a source of data for a morphology analyser, transforming the data from the wiki into constant database tables for quick retrieval, further independent on the wiki software (Garabík 2008). The data are also converted into a nicer looking format for the DICT server (RFC 2229) for a quick web-based search, integrated with several other Slovak language dictionaries.

1.1.2 Multilingual Corpus Linguistics terminology database

As the corpus linguistics is relatively new in Slavic languages – the development began only after the personal computer boom – there is no unified terminology of this field. The terminology started to develop uncontrollably, either by directly adopting English terms or by calquing the English expressions, or by embracing and extending existing linguistic terminology in each country. This development

lead to widely varied terminology in different countries, and even to different terminology used by different institution in the same country, while sometimes the English terms are considered to be just a part of an informal slang.

The key issue is to harmonise the definitions and thus ensure consistency and clarity of information across the languages, especially when communicating with experts from various countries, where the use of bridge language is often not sufficient, or when dealing with bilingual or multilingual resources, with the consequent need of multilingual documentation. The database has been designed in a way to function as a quick reference source of terms in different languages, which has influenced its overall design (Šimková et al. 2009). The database, once finished, could be also used to compare the usage and acceptance of English terms in various languages.

Implementation

Multilingual terminology database (MLTD) is uses the MoinMoin wiki engine as a backend. The data is kept in plain text files, with one file (MoinMoin page) corresponding to one terminology entry. The technical implementation, and to an extent a terminology entry structure has been inspired by the Slovak Terminology Database design (Levická 2007, 2008). This design allows the internal format of the database entry to be kept very simple, nothing more than a plain text file with a minimal layout, without any special formatting markup. By a design decision, internal page format does not use any immediately visible markup language. The motivation stems from the empirical observation regarding usability – the presence of any, even the most incopious markup distracts the editors, unless they are reasonably well trained in the markup (and discourages them to learn to use the system). The markup is hidden in the overall text structure, using nothing more than strategically placed paragraph breaks, colons and parentheses used in a relatively (hopefully) intuitive way.

Each page consists of several entries (one for each language), separated by an empty line. Each entry starts with a term name, prefixed with an ISO 639-1 language identifier separated by a colon (:), followed by an empty line, followed by a definition, followed (immediately) by a source of the definition. Each page can belong to one or more categories – these are expressed by using the usual category mechanism (adding Category* link to the end of the page). A special parser for MoinMoin has been written to display the entries in a distinct graphical way. Main features of the parser are:

- language entries are separated by a horizontal ruler
- ISO 639-1 language identifiers point to an external URL with more information about the language used

- English term is hyperlinked with the corresponding English Wikipedia entry definition source is emphasized
- URLs in definitions or sources are automatically recognized

The points outlined are implemented in order to make the navigation around the database more efficient – they should be thought of as a visual and formatting aid to the database representation, not as a part of the database itself. In fact, the parser can be very easily modified to accommodate different visual styles and different formatting representations.

Terminology entries have been often described using encyclopædic style and format – under the general headword there are often specified other, narrow meanings (e.g. korpus –- korpus hovorených textov: elektronická databáza hovorenej formy jazyka; – korpus písaných textov: elektronická databáza písanej formy jazyka; – národný korpus: jednojazyčný korpus textov konkrétneho národného (jazykového) spoločenstva; – synchrónny korpus: korpus jazyka v jeho súčasnej vývinovej fáze; – všeobecný korpus: nešpecifický, základný korpus zahŕňajúci široké spektrum jazykových štýlov a žánrov, vecných oblastí (domén), autorských generácií, vydavateľských úzov, regiónov a pod.). However, in the MLTD, each of the meanings has to be entered separately.

1.1.3 Slovak-Czech Lexical Database

The primary design goals of the dictionaries created with the help of the database:

- to be primarily a passive readers' dictionaries
- to be general purpose, “traditional”" middle sized (cca. 20–30 thousand entries) dictionaries, with good coverage of different expressions and false friends
- to contain information on levels of usage

From this it follows that the lexical database had to meet the following requirements:

- to be a web based database with queries performed not just by lemmata, but also by varying wordforms
- to include links into various entry related information (such as morphology paradigm)
- to enable easy, online updating and editing by multiple editors

The last two points are satisfied by using wiki based software. The database uses the MoinMoin wiki engine, because it supports custom page parsers and plugins that can be tailored to the needs of an online lexical database. On the other hand,

MoinMoin full-text search is not really scalable – it is a problem especially concerning the Category pages, which internally use the full-text search mechanism. Therefore category pages are not used in the database design.

Basic structure of the database

Basic building block of the database is an entry, which, using MoinMoin terminology, is called a page. It is is used to cover information pertaining to strictly one word meaning, information about homonyms is delegated to the overlying database structure. Each page is uniquely identified by its name, which by convention corresponds to the lemma, or, in case of homonymy, the page name consists of a lemma and a disambiguation identifier (Roman or Arabic numeral).

Lexical entry microstructure

Each page (database entry) is kept in a tabular form, where each item (row) has a predefined form and/or content. As an aid for the editors, fields that contain primary linguistic information have a language flag that indicates the language of that field (i.e. either sk or cs).

Paradigm (sk)

The paradigm field contains an identification of lemma's inflectional paradigm. Since the morphology is also stored in a MoinMoin wiki, the identifier is formatted and displayed as an inter-wiki link, to allow easy one-click access to the complete word morphology. Since all the word forms are available, the entries do not contain any other inflectional information (traditionally, Czech and Slovak dictionaries contain genitive singular and nominative plural suffixes for nouns, or the 3rd person singular and plural indicative forms for verbs). Similarly, since the paradigm contains a complete morphosyntactic specification including a part of speech category, there is no need to indicate the part of speech separately in the database.

Translation (cs)

The translation field contains direct Czech translation of the Slovak word (or of its particular meaning). The best Czech equivalent is chosen. In case there are two or more equally good possibilities, all of them are used, separated by a semicolon (;). The etymological relation between the words are taken into account, and preferably etymologically related translation is used. (For example, the Slovak word *jazykoveda* is translated by the Czech *jazykověda*, even if it could be equally well translated by Czech *lingvistika*, and the Slovak word *lingvistika* is translated as *lingvistika*, even if the Czech *jazykověda* would be a good translation, too.)

In case there is no direct or indirect Czech equivalent of the Slovak word (e.g., pahreba), this field should contain a description of the semantic content.

Number specification (sk)

This field contains the classification of typical or prevalent number or gender characteristics of the word (for nouns). Possible values are:

- usually plural
- usually masculine or feminine
- masculine or feminine
- feminine or neuter
- feminine, usually plural
- masculine, usually plural
- neuter, usually plural
- exclusively plural
- exclusively singular

Qualifier (sk)

This field contains a terminological and/or style qualifier(s), or a special keyword denoting a phrase. The qualifiers are taken out of a fixed set of abbreviated words. When editing this field, the lexicographer is provided with a checkbox entry for each of the qualifiers.

Gloss 1 & 2

Gloss 1 narrows down the semantics – shade of meaning of the entry word or its semantic and functional equivalent. Gloss 2 comments on the typical usage of the word.

Exemplification

The exemplification is not a single field, but consists of a variable number of Slovak-Czech exemplification pairs. The Slovak exemplification is primary, the Czech exemplification should be an appropriate translation of the Slovak one. The table displays all the non-empty exemplifications, plus an empty input field for the last Slovak one (to enable the editor to add another exemplification pairs).

Note

The note contains assorted notes for the dictionary user, relevant to the entry. There is a magic word *viz* (Czech for cf.) to denote a reference to another entry (such as a close synonym, an antonym, comments on significant style characteristics of the Czech equivalents or other related word).

False friends

This field contains a list of false friends, separated by a semicolon. The database does not distinguish between variants of false friends (originating in Slovak or Czech, with a similar meaning, with a completely different meaning...)

Comment

This field is intended for any other comments by the editors – as such, it will not be displayed in the final entry form.

Sense disambiguation mesostructure

There is no place in the entry microstructure to be filled in with hints concerning homonymy disambiguation. Instead, this information is encoded into the overlaying database nomenclature of entries instead, following to some extent the usual lexicographic classification. At the lowest level, an entry is identified by its headword (MoinMoin page name), which – as its first function – directly encodes the lexeme's lemma. If there are two or more closely related, functionally and pragmatically identical word variants (e.g. spelling variations, such as mliekar; mliekár}), a headword can contain more variants, separated by a semicolon (;) as a convenient shortcut. This should be thought of as a shorthand for database compilers, nothing more – functionally, such an entry is equivalent to describing both (or more) variants in full.

A headword can have a trailing uppercase Roman numeral, separated by a space. This is used to mark off major homonyms (or even homographs – such as part of speech homonymy, or a completely – even etymologically – unrelated meaning).

An entry can be created as a subpage of an already existing entry, by using MoinMoin's mechanism for subpages. A subpage XX of a page YY is an ordinary page, with a special name written as YY/XX (i.e. the subpage name follows the main page, separated by a slash). Subpages of a given page are logically clumped together, in the formatted entry output they are displayed nested with the primary page. Subpages are used to connect diminutives, augmentatives and phrasal units to the principal word. Although MoinMoin allows for the whole hierarchy of subpages, only the first level subpages are used (with the exception of sense disambiguation, as outlined the following paragraph).

A headword can have a trailing slash and an Arabic numeral. While technically a subpage, this is used as a weaker variant of a Roman numeral disambiguation in cases, where the words are related and the meaning does not diverge that much. A Roman numeral major disambiguation can be combined with an Arabic numeral minor one (e.g. čap I/1 – a pivot, journal (mechanical device), čap I/2 – a hinge, čap II/1 – a splash, čap II/2 – a catch (act of catching)).

A headword can contain parenthesized reflexive pronouns (sa), (si) (note that *sa* can be added to almost any transitive Slovak (and as *se* to a Czech) verb to express reflexivity, and *si* can be added to almost any verb). This is used with those cases which are either very frequent, or where the reflexive form diverges in its meaning from the non-reflexive one.

Also, this is used with words which do not have straight one-to-one Czech equivalent, in case the presence of the reflexive does not change the basic meaning and usage of the word (e.g. dopukať (sa) – to crack (about skin)).

Technical implementation

The dictionary has been pre-filled with a bilingual glossary of about 60 thousand word pairs and with links into the morphology analyser wiki, in order to ease the initial editing and to enhance the usefulness of the database by offering at least the first-guess translation and morphology paradigm of the words that would not get into the "core" (Garabík, Špirudová 2009).

A page is internally stored as a flat plain text file, with each line corresponding to one table row, with the field name followed by a colon (:), followed by a field value (which can be empty). There is a special MoinMoin formatter plugin that displays the table in a human-friendly way, together with a final, streamlined formatted entry, together with a custom MoinMoin action that is used to edit just one specific table row. The action code has hardwired fields that can contain only a fixed set of values (number specification and qualifier) and provides the editor with checkboxes for all the possible values. The tabular format of the dictionary entries displays the information in a clear and obvious way, however it is quite unsuitable for the intended published (paper) dictionary, and there is also the need to present the information in a more compact, concise form also for the internet-based version. Therefore the table is parsed and formatted into a traditionally looking entry.

Licensing

The database is publicly accessible and editable under a triple license, GNU Free documentation license v. 1.2 and Creative commons Contribution-Share alike (CC-BY-SA) license v. 3.0 for the use in text document, and under Affero GNU Public license v. 3 for use in computer programs (where by "linking" as specified in the license text is understood any use of the dictionary data by a computer program).

1.1.4 Paremiography database

The database is build using MoinMoin engine. Since the most of the data has been obtained via OCR, the most common sources of errors stemming from scanning, converting and parsing the texts are discussed. A paremiography dictionary (or a database) spreads lexicographic description of a language into a broader realm of commonly used expressions, and as such, it extends and complements the (better researched and described) dictionaries of idioms.

Concerning Slovak language, so far unsurpassed paremiography collection is a compilation by Adolf P. Záturecký (Záturecký 1896), first published in 1896. It

contains over 10 000 different proverbs (not counting variants). The influence of this work on any subsequent paremiography compilations was immense, since no other collection came even close to the volume of this work, and there was virtually no need to engage in additional field research – following compilations just upgraded and refined selected subsets of Záturecký's collection. The collection itself has been reprinted several times (with the orthography and language progressively converted to ever increasingly modern Slovak, acquiring additional notes and comments), the most recent edition was published as late as in 2006 (Záturecký 2006).

The core of the collection is made up of proverbs, sayings and locutions. However, there are also some more indefinite units (pieces of weather-lore, rhymes etc.) as well as other types of phraseologisms (similes, figurative expressions). Although the collection does not record phraseology in its entire extent but concentrates on one type of idioms – proverbs and sayings, i. e. stable sentences. Záturecký divided the entire material into 20 thematic groups (man, one's age, sex, family and home, human body, its needs, disease and death, social circumstances, social classes, status, descent and employment, possession and nourishment, food, clothes, cleanliness and dance, human intellect, general rules of wisdom and carefulness etc.). The collection includes immensely valuable material which is however only insufficiently exploited and explored from the point of view of linguistic theory and interdisciplinary research. Záturecký tried to solve the problem of variability of proverbs. His correspondence with other scholars gives also evidence of his interest in the semantics and etymology of proverbs. Záturecký, together with Dobšinský dealt also with paremiological terminology and they attempted to elaborate optimal taxonomy of thematic concepts. Záturecký combined an alphabetical order of statements within the thematic groups. He also applied the formal criterion of division within particular groups and elaborated the index of key words.

Technical implementation

The database has been implemented as a straight, unmodified MoinMoin installation (http://moinmo.in). Since the database is expected to be pre-filled with the data, it will be used mostly in passive mode (searching the data) and the editing will be limited to occasional fixing of typos and OCR errors, there was no need to design an additional user-friendly data visualization and/or editing. The database micro- and macro-structure is implemented only in a set of guidelines for the users, concerning article structure and components, while keeping standard MoinMoin syntax (in fact, only a tiny subset of it, to facilitate further automatized article parsing). The database maps one (semantic) locution into one wiki page. The page starts with locution variants, separated by an empty lines (visualised as separate paragraphs), followed by an optional comment (currently used to note the locution

number in Záturecký's collection, if applicable), followed by a list of categories the locution belongs to (see Tab. \ref{tbl:formalism}). Initially, the core of the database consisted of proverbs from the published subset of Záturecký collection (Mlacek, Profantová 1996), extended by selected proverbs from two other sources (Miko 1989, Smiešková 1988). To these first 2828 entries, was then added Chapter 3 of Záturecký's collection.

Deriving a page name

The database uses carefully designed "semantic hash" for its page names – trying to reduce the locution down to as little words as possible, while keeping a hint of the meaning in the resulting name.

The page names are constructed by eliminating "unimportant" words from the locutions. Not only lexical words (such as nouns, verbs, adverbs, adjectives) are kept in the names, but also prepositions and two words *sa* and *si*. The presence of preposition is necessitated by not lemmatising the nouns – the case is often governed by prepositions and excluding the preposition would lead to markedly ungrammatical sentences. *Sa* and *si* form (among other possibilities) a part of reflexive verbs, and leaving out an obligatory reflexive marker would again emphasise ungrammaticality.

To keep the page names short, there are at most two words that are either noun or verb (with the exception of forms of verbs mať, byť and jesť ("to eat", 3rd person singular *je* is homonymous with the same Slovak morphology database is kept in a MoinMoin wiki system, with a complete paradigm for each word present in the database. The database covers all the words present in the Short Dictionary of the Slovak Language, 4th edition (over 60 000 entries). Although the primary purpose of the wiki is to keep the data for the automatized NLP processing purposes, the data is useful also as a reference database for dictionary-like queries, and therefore the design of the pages has been made with this goal in mind.

Basic unit of the wiki data is called a page (using MoinMoin terminology). Each page contains data pertaining to one lexeme, i.e. lemma with full paradigm and morphology annotation. Each page name is equal to the lemma, taking into account common capitalization of words in Slovak (proper nouns) (an important point, because by design the final morphology analyser disregards the capital letters and gives all the lemmas in lowercase). In case of lexical homonymy, pages are named by the lemmas with part of speech tag attached in parentheses (e.g. *mať_(V)* for a verb, *mať_(S)* for a noun). The page structure attempts to be both human-readable and human-editable and easily automatically parseable. Page body contains of several sections, the first one is the *Lema*, which contains just one word, the lemma. Then follows the *Paradigma* section, containing the inflectional paradigm spelt out in full. For each grammar category there is one corresponding line, with morpholo-

gical tag separated from the form by a colon (:). Alternative forms per one grammar category can be either given on a separate line, or on the same line, separated by a comma (,). At the end of a page there is the part of speech category the described word belongs to.

Homonymy

Only the basic homonymy – where lemmas for two different words (two different parts of speech) are identical – is addressed by the database. The other forms of homonymy (inflectional) are automatically taken care of by keeping the homonyms under their corresponding lemmas and morphology tags. In case of part of speech homonymy, there is a special disambiguation page, linking to all the possible lemmas.

In Slovak, reflexive verbs are marked by a special separate morpheme *sa/si*, which is separated from the verb and has relative freedom of movement around the verb (Unlike other languages, e.g. in Russian the reflexive pronoun/particle takes a form of a clitic inseparably bound to the verb). As there exist a reflexive/non-reflexive dichotomy (i.e. reflexive verbs having almost always their non reflexive counterpart), only the non reflexive parts in the dictionary, without the *sa/si* pronoun. Several singular cases of reflexive verbs without a meaningful standalone non reflexive counterpart (*smiať sa, báť sa, uvedomiť si, čudovať sa*) do not pose any problem – the missing *sa* is confusing only for the uninitiated users.

Traditionally, *sa* and *si* are called "reflexive pronouns" if semantically there is a discernible action performed on the agent (i.e. they can be seen as contractions of personal pronouns *seba* and *sebe*), otherwise they are considered to be a part of a verb. This is just a convention – they could be called equally well to be particles, indeed this is how they are sometimes classified in the traditional Czech grammars. In the database, they are assigned a special morphology tag **R**, regardless of their semantic use.

1.1.5 Bulgarian-Polish Lexical Database

Unification of classifiers

One of the main problems of the development of digital dictionaries is the choice of classifiers. Whenever the development of a system of bilingual dictionaries (serving as a future basis for a system of multilingual dictionaries) is concerned, there arises the issue of unification of the classifiers in the dictionary entry. In order to harmonise the classifiers for various languages, we need to present a unified selection of classifiers and a standard form of their presentation. In a broader sense, the issue of unifying classifiers in the dictionary entry is close to the issue of a new

part-of-speech classification oriented towards the specifications of a digital dictionary. For example, the unification of classifiers in the proposed structure of the lexical database (LDB) that support the Bulgarian–Polish online dictionary allows synchronisation and unified representation for the data on the two languages (Dimitrova, Koseska 2008b, 2009a).

An important classifier of the verb which must be included in the dictionary entry refers to the transitivity or intransitivity of the verb. The tendency of including more classifiers in the dictionary entry confirms the necessity of a classifier reflecting transitivity or intransitivity of the verb. It is a common practice to list as a headword in the dictionary entries the infinitive of the verb. In Bulgarian the infinitive has disappeared and has been functionally replaced by the "da-construction", which connects the particle "da" to the present tense forms. In this respect Bulgarian is more similar to other Balkan languages (Modern Greek, for example), but differs from Polish where the infinitive is preserved. This is an important example for the requirement of distinguishing a form from its function and meaning. The present tense form in this case does not have "present tense" meaning. In the Bulgarian verb entries it is accepted to list as headword the 1st person singular form of the present tense.

The classifier "aspect" of a verb is universally accepted. So the "aspect" classifier in the dictionary entry for a Slavic language is obligatory. The aspect in Slavic languages is a well-formed grammatical category whose meaning boils down to the expression of events – by the perfective aspect, and states – by the imperfective aspect, where "event" and "state" as described in the net description of temporality in a natural language were interpreted. In languages such as Polish, Czech, Slovak, Ukrainian and Russian, in which "aspect" is a strongly developed semantic and grammatical category, there are few tense forms. This is not the case in South Slavic languages, in which, for example, in Bulgarian, has a high number of tense forms as well as a strongly developed semantic and grammatical category "aspect". As we know, the languages which lack the grammatical category "aspect", such as Latin, French, Italian or Spanish, has a high number of tense forms. As mentioned in (Koseska 2009b), there are two distinct tendencies in the South Slavic languages – the first towards reduction of tense forms (Croatian/Serbian), the second one towards reduction or extinction of the aspect. So it should happen in Bulgarian, but does not! In Bulgarian the development of category "aspect" does not lead to a reduction of the tense forms.

The work under the MONDILEX project demonstrates the potential for developing useful lexicographic reference works (both digital and hardcopy) by using the format of a LDB and an adequate mathematical foundation. Various parameters of classification of the lexicon are likely to emerge in the process of developing a lexical database. As this will possibly occur through distributed effort, it highlights the importance of an interface to the lexicographic system. The LDBs should be

brought in line with one another by sharing theoretical concepts and platforms. Synchronisation and unification of bilingual dictionaries entails a uniform structure of the dictionary entry; the unification of classifiers for presenting headwords; a synchronous presentation of morpho-syntactic features, and a uniform presentation of the content. Common suggestions of the Bulgarian and the Polish teams regarding the unification of classifiers can be grouped around the mode of classification of forms and the mode of denoting the meanings of verb tense forms (two types with exact definition that can be "translated" in a formal language).

Structure of a traditional paper dictionary entry:

```
Headword
Formal Features – phonetics, grammar, morphology, syntax,
etymology, style
Semantic information
Quotations
Additional information:
1. Derivatives
2. Phrases
3. Examples - phrasal and sentence usages, illustrations
```

Formal Model

The formal model for dictionary encoding should be developed in accordance with the complex structures of the dictionary entries. These structures reflect to the content of the dictionary entries, which are very different and depend of the grammatical features of the headwords.

The starting point for the formal model of lexical database (LDB) of the first Bulgarian-Polish experimental online dictionary (Dimitrova et al. 2009b) is the CONCEDE model for dictionary encoding. This model was developed in the framework of the EC project CONCEDE (Consortium for Central European Dictionary Encoding[1]).

The tagset for LDB of the Bulgarian-Polish online dictionary contains 3 structural tags and a set of content tags.

(1) The structural tags are:

alt – a tag indicates alternation, though generally for use in quite different contexts,
entry - a tag, contains the dictionary entry,
struc- a tag indicates separate independent part in the dictionary entry.

[1] http://www.itri.brighton.ac.uk/projects/concede/

(2) The set of content tags includes all other 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.***

The **hw** tag contains the headword and is used for alphabetization and indexing, access. The **pos** tag indicates the part of speech assigned to a dictionary headword (noun, verb, adjective, etc.): <hw>**свобод|а'**</hw><pos>noun</pos>.

The **xr** tag uses to indicate a cross reference with the pointer:

<hw>**построя'ва|м**</hw> <xr>**постро|я'**<xr>.

The **gram** tag contains grammatical information relating to a word other than gender, number, case, person, tense, mood, itype, as these all have their own element, for example, perfective aspect and imperfective (progressive) aspect: <gram>imperfective</gram>. The **subc** tag contains sub-categorization information (transitive/intransitive for verbs, countable/non-count for nouns, etc.): <subc> transitive </subc>.

For a more adequate description of the Bulgarian verbs, two new tags are being introduced to represent the verb's conjugation (Bulgarian verbs are divided into 3 conjugations): **conjugation -** *a new tag* is added to represent the conjugation of verbs; its structure allows the subtag **type** for the possible types of conjugations of Bulgarian verbs. Furthermore, it is allowed to input additional information in the **gram** tag for the aspect – *perfect and progressive* of verbs, and in **subc** tag – for *transitivity/intransitivity* of verbs. The value "NILL" in order to represent empty corresponding values was introduced.

The selection of headwords included in this LDB is based on the Bulgarian-Polish parallel corpus. The main forms (lemmata) of the most frequent word forms in the corpus are selected. The word distribution according to parts of speech follows the CONCEDE model: open parts of speech - no more than 90 %, closed parts of speech – minimum 10% of the whole set of lemmata chosen.

Let us consider an entry of the Bulgarian–Polish LDB, whose respective dictionary entry of the Bulgarian–Polish printed dictionary is:

сп|я, -иш *vi*. spać; **~и ми се** chce mi się spać, ogarnia mnie senność

The grammatical features of this Bulgarian verb ***спя*** /sleep/ are:

aspect - imperfect (progressive) /*несвършен вид/,* this verb is **intransitive** /*непреходен*/*,* its conjugation is a **II type** */II спрежение*/.

The structure of the entry with headword ***спя*** /sleep/ in Bulgarian–Polish LDB follows:

```
<entry>
<hw>сп|я</hw>
<pos>verb</pos>
<gram>imperfect</gram>
<conjugation><orth>-иш</orth>
                    <type>II</type>
</conjugation>
<subc>intransitive</subc>
<struc type="Sense" n="1">
<trans> spać </trans>
</struc>
<struc type="Derivation" n="1">
<orth>~и ми се</orth>
<struc type="Sense" n="1">
<trans> chce mi się spać </trans>
<alt><trans> ogarnia mnie senność </trans></alt>
</struc>
</struc>
</entry>
```

Realization of homonyms

The meanings of homonyms are entered in the dictionary as different database records. On the word entry page, 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"><hw>|ясен</hw>
<gen>м.</gen>
<struc type="Sense" n="1">
<def>Широколистно дърво с перести назъбени листа и
яка, трайна и еластична дървесина, Fraxinus;
осен.</def></struc>
</entry>

<entry n="2"><hw>|ясен</hw>
<pos>прил.</pos>
<struc type="Sense" n="1">
<def>За небе, време и под. - който не е покрит с облаци, във
или през който няма облаци, мъгла; ведър, светъл. Прот.
мрачен, облачен.</def>
<eg><q>Ясно небе.</q></eg></struc>
<struc type="Sense" n="2">
```

```
<def>Светъл, блестящ, сияен.</def>
<eg><q>То не било ясно слънце, най ми била сама Неда.
Нар.п.</q><q>Ясни звезди.</q></eg></struc>
<struc type="Sense" n="3"><usg type="register">прен.</usg>
<def>За глас, звук - звънлив, чист, бистър,
приятен.</def></struc>
<struc type="Sense" n="4"><usg type="register">прен.</usg>
<def>Който се чува, вижда или разбира добре; отчетлив,
разбран.</def>
<eg><q>Ясен говор.</q><q>Ясно писмо.</q><q>Ясна
мисъл.</q></eg></struc>
</entry>
```

Technical implementation

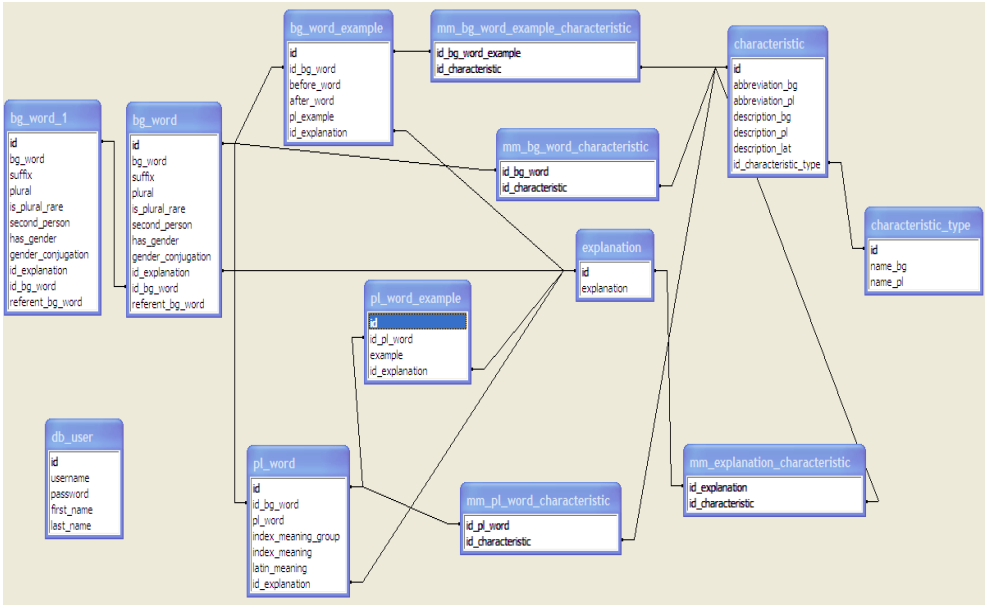
To enable Internet access to the Bulgarian-Polish dictionary, a relational database is used. The lexical database is converted to the relational database with the help of tables containing search data and indices. This organization allows an automatic creation of a dictionary entry for a Polish word, whenever there is a one-to-one translation equivalent.

Relational Database

The LDB serves to design and develop the relational database, which is the basis for the subsequent development of the web-based application for support of the Bulgarian-Polish dictionary.

The model of a relational database is based on lexical entries. An option enabling the translation from Polish to Bulgarian was also provided in the relational database's design. The translation will be automatically made only from the main meanings of the Bulgarian headwords. All additional information, like senses, quotations, derivations, phrases, etc. should be updated by an authorized human editor. Of course, the input of information about the Polish word must be done additionally.

The structure of the relational DB is given in the following figure:



Relational database upon the lexical database of the Bulgarian-Polish-Bulgarian Dictionaries

The proposed relational model can be used for all database management system. For the particular realization of the dictionary the system MySQL is used, which is one of the most popular. MySQL is an open source code and provides interface for the programming languages C, C++, Eiffel, Java, Perl, PHP and Python.

The MySQL server is frequently used for web-based applications and is one of the best choices for building database systems due to its high flexibility, and it is free. The management of MySQL databases is based on phpMyAdmin, which is programmed to manage MySQL via the web. It is free and available in 47 languages. Its functionalities include creation, deletion and editing of tables; adding, deletion and editing of columns; management of keys and columns; management of privileges; SQL query processing; visualisation of data in different formats.

LDB is transformed into a relational database with the help of XML syntactic parser that checks syntax of a given XML file and processes the file's elements. The implementation of the parser for data transfer from the LDB to the relational DB uses the DOM technology Java Development Kit version 1.6. The parser has four principal parts: *help-classes* representing the structure of tables in the relational DB; a *help-class* for link to the MySQL DB; a *class* with the main syntactical analysis logic and the storing procedure for the DB; an *entry-point class* for the program.

Transformation of the Lexical Database to the Relational Database is carried out with the help of tables, into which the search data and indices are input. This organization allows an automatic creation of a dictionary entry for a Polish word, whenever the translation equivalence is one-to-one. Of course, the input of information about the Polish word must be done additionally.

One of the main tables is table ***bg_word***, where the headwords of Bulgarian language and their main characteristics are stored. This table is the entry point to the web-based application that supports Bulgarian-Polish online dictionary. The table ***pl_word*** contains the information for a Polish word that is automatically extracted from a Bulgarian entry. The table ***mm_bg_word_characteristic*** contains the indices of the Bulgarian word characteristics. The tables are presented in the following figures:

id	id_bg_word	pl_word	sense_ index	alternative_ sense_index	latin_ translation	id_explanation
1117	668	podkreślać	1	1		
1118	669	podkreślony	1	1		

Table pl_word

Column / Word	завъ’рш|а	завъ’ршва|м	завъ’ршен
id	662	663	664
homonym_index			
bg_word	завъ#рш	завъ#ршва	завъ#ршен
suffix	а	м	
bg_word_search	завърша	завършвам	завършен
plural			
is_plural_rare			
conjugation	иш	ш	
conjugation_type	2	3	
has_gender			
gender_feminine			
gender_neuter			
id_explanation			
id_bg_word	582		
referent_bg_word	завъ#ршвам		

Table bg_word

id_bg_word	id_characteristic
668	17
668	57
669	44
670	18
670	57

Table mm_bg_word_characteristic

Recommendations for designing a common encoding scheme for Slavic multilingual dictionaries:

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. 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 by lemmata, but also by 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 paradigm)
- to enable easy online updating and editing by multiple editors.
- to keep track of revision history, with the possibility of rollback.

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 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. We recommend unifying the systems of categories and tags used for annotation in the various systems.

When dealing with various languages, it is important that all participants agree upon a common terminology for the problem at hand. This is doubly important when Slavic lexicography is concerned, mostly because of two opposite phenomena: first, 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: (1) of two Slavic languages, in order to serve as a testbed for a bilingual database of corpus linguistics terminology, (2) of all languages of the MONDILEX project (including English). The database shall 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.

We recommend creating 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.

The synchronisation and unification of bilingual dictionaries shall involve:

- Uniform structure of the dictionary entry.
- Unification of the classifiers for presenting headwords in the entries.
- Synchronous presentation of morpho-syntactic descriptors (core and specific features).
- Uniform presentation of the content.

1.2 Dictionaries

1.2.1 Dictionary of Slovak Collocations

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.

For the Slovak language, the only existing collocation dictionary was published in 1931, with a revised edition in 1933 (the author called this book "a dictionary of phrasemes", but in fact it was a dictionary that contained not only phrasemes, but also common word collocations) (Tvrdý 1931, 1933). Since then, the language has undergone immense changes in almost all of its parts, starting with the whole sociolinguistic situation and ending with substantial changes in the vocabulary and orthography. As of today, the dictionary is mostly of diachronic importance, and there is a notable gap in Slovak language lexicography with regard to collocations – modern approaches in lexicography, especially the use of large language corpora partially fill the gap, but they still cannot replace a well-documented, systematically built dictionary of collocations.

The described electronic dictionary of Slovak collocations is being compiled at the University of St. Cyril and Methodius, Trnava, in cooperation with the Slovak National Corpus department of the Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava (Ďurčo P. et al. 2009). The project on Slovak collocations that started in 2007 is the first of its kind in Slovakia and is aimed at the registration and description of selected multiword lexemes and phrasemes as well as typical collocations with restricted collocability. The dictionary provides an overview of the combinatorial behaviour of words, in the first phase the most frequent nouns extracted from the Slovak National Corpus database, with the intention to include also verbs, adjectives, adverbs, and particles. Currently, the database contains information about nouns and (as a separate subproject) particles. Description models on the basis of collocational matrices are also elaborated for verbal, adjectival, adverbial and partical collocations.

Obtaining collocation profiles

An efficient tool for modelling semantic proximity of words and their collocation profiles in large lemmatized corpora is the sketch engine (http://www.sketchengine.co.uk/) – a corpus tool which generates word sketches, i. e. corpus based summaries of a word's grammatical and collocational behaviour. Disadvantages of the sketch engine are long lists of isolated lemmata and too many automatically

generated redundant data in the results, obtained through fixed set of unary, dual, symmetric and trinary rules, which do not always correspond to natural collocational clusters in the language. The basic tool for searching collocations for each entry is the corpus manager client Bonito which provides searching, sorting and statistical evaluation of collocations. By using this tool it is possible to view each given word, extract concordances for each word to get an overview of its behaviour in context, get statistical information like absolute frequency, MI-score, t-score, , MI3, log likelihood, min. sensitivity and salience to recognize word co-occurrences.

Despite these new language technological analysis, scepticism still prevails regarding the possibility of capturing and describing the examined data completely. In particular, this scepticism results from two problems. Word co-occurrences represent a diffuse continuum of semantically connected elements, some of which are linked less closely than the others. The borders between "free" and "bound" cannot be clearly specified. On the other hand, the main problem of the statistical approach is that the frequency and semantic firmness of word combinations do not correlate directly. Not all highly frequent word combinations are also bound. One finds typical collocations in all ranks of the frequency distribution. In the lexical database, the (meaningful) collocations are manually selected from the first 500 occurrences of each grammatical structure listed by the Sketch Engine and cross-checked against the Slovak National Corpus concordances. The statistical results vary, they depend both on the used statistical method and the quality and accuracy of taggers and lemmatisers, the precision rates whereof are different.

Technical implementation of the lexical database

The database macrostructure is simple – all entries are equal, each entry corresponds to one MediaWiki page, neither subpages nor redirects are used. A page is named by an entry lemma, Slovak lexical entries are differentiated from other pages (system pages, user discussions) by the category they belong to (Slovak Nouns, Slovak Adjectives, Slovak Verbs, Slovak Particles).

Structure of an entry

An entry page consists of three main sections: *Významy* (Meanings), *Kolokácie* (Collocations), *Externé odkazy* (External links). While the structure of *Významy* and *Externé odkazy* is the same for all the parts of speech and these sections do not have any substructure, the structure of *Kolokácie*, the most important section, is more complicated (Ďurčo 2007).

Významy

This section ("meanings") contains a bullet list of descriptions of different definitions of the lexeme. The collocations are not split according to polysemy (or

homonymy) of the base noun inside one part of speech category at all, neither there is a distinction between homonyms in collocations. This was a deliberate design decision, based on two observations: First, often a collocation is not clearly attributable to a specific meaning; let alone trying to define and distinguish meanings, which is traditionally a very cumbersome task, where no general consent could be achieved. This was not seen as a task for this project and would unnecessarily slow down the dictionary construction and open door to endless discussions inside and outside the project team about the distinction of individual meanings.

Kolokácie

All the collocation data are contained in this section. The detailed structure is differentiated according to part of speech the entry stands for. For nouns, it is divided into two subsections for the singular and plural, reflecting the fact that collocates often exhibit different phenomena according to the grammatical number of the base noun. Each of these subsections is further divided into many subsubsections, each for a specific collocation combination.

The subsubsections' naming scheme encodes some human readable information about the collocations, with the base noun marked by the string Sub1Xxx, where Xxx is the abbreviation of the noun's case (so the whole string will be one of Sub1Nom, Sub1Gen, Sub1Dat, Sub1Aku, Sub1Lok, Sub1Ins). Vocatives are conflated with the nominative case, to avoid the controversy about Slovak vocative existence – fortunately, it just happened that none of the nouns chosen for the collocation dictionary is from the set of those few Slovak words that have a morphological vocative.

The other part of the subsubsection name reflects describes the neighbouring word part of speech, so it can be one of Sub2, Verb, Atr (another noun, verb, attribute). Atr subsumes adjectives, pronouns, particles or numerals. This string is positioned either to the left or to the right of the previous base noun string, depending on the predominant position of the word in collocations (but including also the collocations with a different word order). The strings are concatenated with a plus sign, so e.g. the whole subsubsection name Verb + Sub1Aku indicates that the subsubsection contains collocation of verb and base noun in acusative (not necessarily in this order).

Externé odkazy

This section is populated by several macros (templates), providing links to external resources. Each macro has one parameter, equal to the identification of given word in the target database – mostly the same as the lemma, different only in case of homonyms (differentiated at the target). The macros construct an URL pointing to an external resource and insert it as an http hyperlink into the rendered page. The

macros in use are {{ma|...}} to link to morphologic database (this macro is intended to record relations between full word paradigms and the collocation dictionary entries, both for the end user and for eventual computer processing), {{slovnik|...}} to link to dictionaries published at the Ľ. Štúr of Linguistics WWW page (http://slovniky.juls.savba.sk), {{linky|..}} to point to several search engines, such as Google, Ask, Yahoo, Cuil, as well as the Slovak National Corpus. The latter two templates are meant for human consumption, not for computer parsing (due to the somewhat unpredictable nature of the target data). If a need to either add or remove an external data source (e.g. a search engine) arises, or if the form of URL parameters changes, only the template needs to be modified, and the change will be automatically reflected across all the database entries.

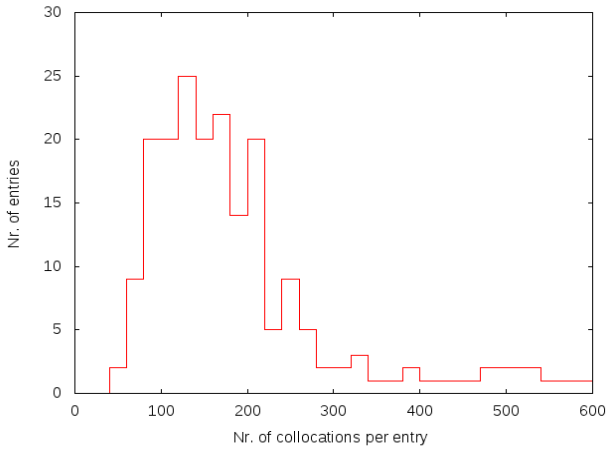
Collocation entry microlanguage

The lexical database has been designed with a goal of a human readable collocation dictionary in mind, published both online and in printed form. However, the entry microformat is designed to be computer readable, except of some minor exceptions, where the (complete) readability stands in the way of human interaction.

Each collocation can be though of as consisting of two units: the base noun and the collocate. The collocates are normalised (lemmatised), and the collocation is written with the base in its corresponding case/number. The exception is only for the combination Atr + Sub1Nom, which is so frequent that the base in nominative is omitted, if it follows the attribute. Auxiliary particles/pronouns are sometimes rearranged, to fit the syntactical requirements of the base (this applies mainly to the reflective pronouns sa, si in combination with infinitives). From this follows that the parser must include the morphology generator in order to recognise the base noun in other forms than nominative singular, and a complete automatised parsing is difficult without including some sort of syntactical rules into the parser. Collocate is terminated by the | (U+007C VERTICAL LINE) character surrounded by whitespace. The vertical line has to terminate also the ultimate collocate in the subsubsection. If there are no collocates for a given collocation pattern, the entry consists of a single vertical line character in a separate line. Optional words (which are sometimes present in a given collocation) are enclosed in parentheses, separated by the rest of collocation by a whitespace or punctuation. Parentheses adjoined to a word specify optional prefixes or suffixes (mostly verb negation or aspect modifier). Variants in words (two or more words that do not change the collocation meaning and are approximately equally frequent) are separated by a slash, three dots (ellipsis, …) denote incomplete variant enumeration (signalling that there are more variants occurring in the corpus than given, usually these variant components belong to a specific lexico-semantic group). Special indefinite pronouns (*niekto*, *niečo*, …) serve as wildcard valency markers which stand for a general class of

aminate/inanimate nouns (and thus signal that the collocation is too broad to be automatically parsed).

There are on average 173 collocations per entry. The symmetry is slightly skewed in favour of small number of bigger sized entries (the median is 157). The entry with least number of collocations is *kára* (cart, barrow), with 40 collocations, the highest number has the word *svet* (world) – 584 collocations. However, the exact number of collocations per entry is subject to several arbitrary conditions, among them the level of detail in describing collocation variants, inclusion of otherwise optional ellipsis and indefinite pronouns, and in general subjective evaluation of collocation candidates by a lexicographer compiling the entry.



Distribution of number of collocations per noun

1.2.2 Bulgarian-Polish online dictionary

The experimental version of the first Bulgarian–Polish electronic dictionary is prepared in WORD-format and at present contains approximately 20 thousand dictionary entries. This dictionary provides a part of language material for the lexical database of the web-based application that supports the Bulgarian-Polish online dictionary. The Bulgarian–Polish online dictionary pursues so far experimental purposes. A lexical database provides the language material for the dictionary.

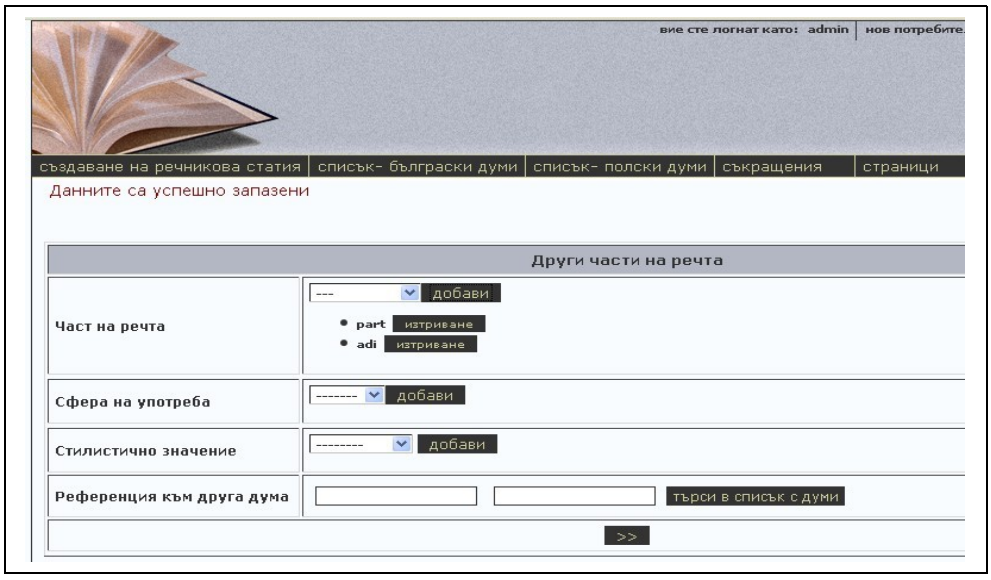
Web-based application for the representation of the Bulgarian-Polish online dictionary consists of two basic modules: an administrator module and an end-user module (Dimitrova et al. 2009d).

The ***administrator module*** is intended for the person updating the dictionary, and is accessible only for authorized users. There are possibilities to create more than one

user with different passwords and usernames. The administrator module is used to fill in the database and to offer user-friendly interface to the user who will be responsible for word management: for adding, editing, deleting and searching words.

After the user's username and password have been verified, the user is redirected to the administrative module where there are several sections – section for entering a new word, sections for searching Bulgarian or Polish words, section where the user can enter new abbreviations, section for setting translations of the user alerts and messages so the user can change the both Polish and Bulgarian translations.

Administrative panel – choosing the type of the word
which will be added: a noun



Administrative panel – 2nd step of adding the participle

There is a common part for each part of speech that ensures the possibility to add unspecified number of derivations, phrases and examples for each headword. At the end of each page for entering headword there is a button "Add derivation/phrase/example". When the user clicks on it a new window is opened in order to add as many as needed derivations, phrases and examples for this headword. Realization of the homonyms in the web-based application: the meanings of the homonyms are entered in the dictionary as separate database 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.

The ***end-user module*** is aimed at presenting correct and up-to-date information to the user. For convenience and ease of searching and finding the meanings of words the end-user module offers:

- An option for translation from Polish to Bulgarian,
- Means that enable the end-user to report missing words,
- User interface in both languages – Bulgarian and Polish.

The end-user module is bilingual, the user can choose the input language (Bulgarian or Polish) and according to his/her choice, a virtual Bulgarian or Polish keyboard is displayed. In this way the user can choose special Bulgarian or Polish characters if they are not supported by the keyboard useds. After making a search for a word on the left site of the screen a list of words, starting from the given entry, are displayed. When clicking on any of these words in the list the translation is visualized in the right frame.

If we translate from Bulgarian to Polish, the whole information saved in the RDB is displayed. In this application there are three sections – section for translating a word, information section and section for reporting a missing word. The end users may report words that are missing in the dictionary into a provided "Contact" form. In this case the administrators will add the reported missing words into the database at a later session. Both modules have "Help" panels. The program realizing the web-based application for representation of the Bulgarian–Polish online dictionary allows expanding the dictionary volume by adding new words, enriching the content of the dictionary entries from the LDB by adding new examples for clarification of the meaning, etc.

The screenshot illustrates the translation of the Bulgarian verb “завършвам” /to finish/ into Polish

The screenshot illustrates the translation of the Polish verb “kończyć” /to finish/ into Bulgarian

Technical implementation

The web-based application for the representation of the Bulgarian-Polish online dictionary, developed by IMI-BAS – the Bulgarian participant of the project, is **an example of software product for creating digital dictionary**. The technologies used for the implementation of the web-based application are Apache, MySQL, PHP and JavaScript. These are free technologies originally designed for developing dynamic web pages with a lot of functionalities. With the help of HTML and CSS the designs of both administrative and end user modules were created. The current version of the Bulgarian-Polish online dictionary works optimally with Internet

Explorer 6.0+ (Windows), and with Firefox 2.0.1+ (Windows, Linux). The website resolution is 1024×768 pixels.

Furthermore, the structures of the developed Bulgarian-Polish LDB and of the web-based application allow a replacement of the Polish translations (texts) by texts in another language L2. Thus, the LDB and the web-based application can be useful for the development of a new bilingual Bulgarian-L2 online dictionary.

1.2.3 Dictionaries of Ukraine on-line

“Dictionaries of Ukraine on-line” (http://lcorp.ulif.org.ua/dictua/) is one of the front-ends to the lexicographic system "Dictionaries of Ukraine" (Shyrokov 2009a, Shyrokov et al. 2009), designed to serve the needs of wide audience and provide the basic reference and search functionality. The technological core is special software that runs in the local network of ULIF-NASU. The server part is implemented using a web-service that provides a program interface to access the lexicographic database of the system and the modules for automatic construction of paradigm, word stemming etc.

The web interface was built using the ASP.NET technology. This choice is determined by the following factors:

- ASP.NET is a technology closely linked with the .NET Framework and is aimed at creation of dynamic web applications. ASP.NET technology is the optimal choice because the technological core of the "Dictionaries of Ukraine" system is implemented on the .NET platform;
- ASP.NET makes it easy to interact with the web services;
- The ASP.NET pages (web forms) are compiled, providing better performance compared to script-based applications;
- The process of creating web forms is quickened by using standard components, such as GridView, DetailsView, etc.;
- ASP.NET provides the infrastructure for creating reliable and stable applications that are easily scalable.

The positive features of the integrated system are:

- display of the full registry;
- ability to reload individual parts of a page;
- the set of inputs to the system is not limited to a registry row, and covers the right parts of the entries too.

The L-system ‘Dictionaries of Ukraine’ includes four subsystems: ‘Inflexion’, ‘Phraseology’, ‘Synonymy’ and ‘Antonymy’.

The general registry (over 256 thousand words) of the system ‘Dictionaries of Ukraine on-line’ is based on the registry of the Ukrainian Language Spelling Dictionary, which is almost fully replicated and expanded.

The subsystem ‘Inflexion’ is created on the basis of the inflectional classification of the Ukrainian vocabulary developed at ULIF-NASU. It contains over 2000 paradigmatic classes for all parts of speech defined by formal features. Due to this classification and the software implemented (paradigmatisation – creating a full inflectional paradigm based on the canonical (dictionary) form of the lexeme), a full list of all grammatical forms for all lexical items listed in the registry was created. It enables the visualization of the word forms in all grammatical meanings.

The total number of all word forms in the registry of over 186 thousand units is approximately 3.4 million. The subsystem provides a mapping of the table of all word forms for a registry unit specifying their grammatical parameters.

The subsystem ‘Synonymy’ reflects the synonymic richness of the Ukrainian language. The ‘Dictionary of Synonyms of the Ukrainian language’ (Buryachok A. A. (Ed. 1999)) in 2 volumes was the source of linguistic information.

The software provides presentation of the synonymous rows (about 9200), consisting of the words or their individual meanings, as well as idioms. The core of each synonymous row is its dominant lexical unit with the broadest set of semantic features for the row.

The elements of the synonymous rows are marked with semantic, grammatical and stylistic characteristics. The use of synonyms is illustrated with their typical contexts – quotations from fiction, newspapers, magazines, scientific literature, etc. and with word combinations.

The home page of the ‘Dictionaries of Ukraine on-line’

The subsystem ‘Antonymy’ is based on the ‘Dictionary of Antonyms of the Ukrainian language’ (Polyuga 2001), which consists of 253 entries representing about 2200 components of antonymic pairs.

About 56 thousand phraseological units represented in the ‘Dictionary of Phraseologisms of the Ukrainian language’ (Ukrainian Phraseology 2003) became the linguistic source for filling the lexicographic database of the subsystem ‘Phraseology’. The common phraseology of the Ukrainian language is fully represented in this dictionary. It also contains full lexicographic description of Ukrainian phraseologisms.

The online dictionary presented here can be used as a prototype for the future bilingual lexicographic resources which will be designed within the MONDILEX project.

1.3 Corpora

The great achievements of the information technologies offer numerous methods of natural language processing, especially for the development and use of corpora, both monolingual and multilingual. The availability of high-quality text corpora for the languages concerned is of utmost importance for the task of any lexicographic research. Simple (untagged) monolingual or multilingual corpora can be used for relatively simple lexicographic tasks like registering regular collocations, or, in the case of multilingual corpora (in particular, bilingual ones) finding translation equivalents. Tagged corpora can serve as basis for much more sophisticated research, both in the course of primary dictionary creation and past the point when the bulk of the dictionary is ready. Lexicographers using such corpora are able to establish and validate non-trivial properties of lexical units, e.g. subcategorization frames, complex syntactic features, semantic properties and lexicographic classes.

The higher the level of corpus annotation, the more elaborate research challenges can be issued and addressed.

1.3.1 Monolingual corpus SynTagRus

A good example here is SynTagRus (Boguslavsky et al. 2000, Boguslavsky et al. 2002, Apresjan et al. 2006, Boguslavsky et al. 2008, Nivre et al. 2008, Iomdin et al. 2009), a deeply tagged corpus of Russian texts developed by IITP-RAS, the Russian partner to the MONDILEX project, which offers, for each sentence,

(1) fully disambiguated morphological annotation, i.e. the lemma, part of speech and the list of inflexional morphological features of every word;

(2) a complete syntactic structure represented in the dependency formalism as a tree whose leaves correspond to every word of the sentence and whose branches are labeled with names of syntactic relations: in all, there are about 75 different syntactic relations that account for syntactic links like (a) the predicative one, connecting the verbal predicate of the sentence as syntactic head with its nominal subject as syntactic daughter, as in *korova mychala* ‘the cow was mooing’, (b) the 1st completive one, connecting a predicate word of any part of speech as syntactic head and a word representing the first complement thereof as syntactic daughter, as in *zhevala travu* ‘was chewing grass’ etc.;

(3) partial lexical functional annotation, identifying arguments and values of collocation-type lexical functions as proposed in the Meaning – Text linguistic theory developed by Igor Mel’čuk;

(4) partial semantic annotation that (1) ascribes, to some words of the sentence, their semantic features and (2) identifies semantic roles of predicate words and

their instantiation. The inventory of semantic roles and features comes from the new fundamental classification of predicates recently proposed by Juri Apresjan.

At the moment, SynTagRus includes over 42,000 sentences amounting to ca. 600,000 words.

At the moment, there are no other corpora of comparable size and depth of annotation for the languages of MONDILEX participants; however, a well-known Prague Dependency Treebank for the Czech Language (see e.g. Hajič et al. 2006) can serve similar purposes and meet the same challenges.

It is highly desirable that other Slavic Languages could resort to the resources of a similar kind.

An example of application of monolingual corpora in contrastive studies: onfluence of the dative and Middle Voice in Croatian and Polish. In Croatian and Polish various constructions with the reflexive marker *se/się* may or may not involve a noun in the dative case. In Croatian one may say *govori se o ovome problemu* 'this problem is discussed' as well as *stalno im*-DAT *se govori o tom problemu* 'they are being told about this problem all the time'. Other examples include, for instance, *Kto wie, co się zdarzy za dziewięć miesięcy* (Polish) 'Who knows what will happen in nine months' as opposed to *A jeżeli zdarzy im*-DAT *się coś złego?* 'And what if something bad happens to them?'. The way in which the *se/się* construction interacts with the dative case in the construction of meaning is discussed (Stanojević, Kryżan-Stanojević 2009). A corpus study was conducted on the IPI PAN corpus of Polish[2] and the Croatian National Corpus[3] to find examples where the *se/się* construction coincided with the dative construction. The results show that there are two basic semantic groups: the allative/competitor group and the transfer group, which partially corresponds to semantic groups found for various dative senses. In these senses both the dative and the *se/się* construction are grammaticalized in respect to their other senses, and are hence semantically bleached. Therefore, in those senses a new constructional meaning occurs, which is not present in any senses of the two components taken alone: dative as the experiencer of its internal change of state. Constructional meaning is possible only in the bleached senses, which are less detailed in respect to the "basic", diachronically older senses.

1.3.2 Multilingual parallel corpora

Parallel corpora are bilingual in the least and this fact distinguishes them fundamentally from monolingual corpora. Language material in parallel corpora, unlike

[2] http://korpus.pl/
[3] http://www.hnk.ffzg.hr

the one in monolingual corpora, has to be at the synchronous level and must reflect the current state of the two (or more) languages.

Keeping in mind the richness and diversity of natural languages, we point out that the selection of texts in a parallel corpus is essential, especially for linguistic purposes.

MULTEXT-East parallel and annotated corpus

Here we want to mention some well-known multilingual corpora that were created in recent decades in the field of corpus linguistics, namely, the MULTEXT corpus (Ide, Véronis 1994), initially in seven West European languages (Dutch, English, French, German, Italian, Spanish and Swedish, with more in later editions, including Bambara, Catalan, Kikongo, Occitan and Swahili), and the MULTEXT-East corpus[4] (Dimitrova et al. 1998), initially in six Central and East European languages (Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovenian, plus English as a "hub" language, in later editions including Croatian, Lithuanian, Resian, Russian and Serbian).

The project MULTEXT-East[5] (MTE for short) is an extension of the project MULTEXT, one of the largest EU projects in the domain of the language engineering prepared useful language tools and resources.

The MTE project has developed a multilingual corpus that contains annotated parallel and comparable corpora, in which three languages: Bulgarian, Czech and Slovene, belong to the Slavic group. 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 (Ide 1998). The corpus contains four parts, corresponding to the different levels of annotation: the original text of the novel, the CesDOC-encoding (SGML mark-up of the text up to the sen-

[4] http://nl.ijs.si/ME/

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

tence-level), the CesANA-encoding (containing word-level morpho-syntactic mark-up), and the aligned versions in CesAlign-encoding (containing links to the aligned sentences). 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 translations in each of the six CEE languages produces six pairwise alignments comprising the MTE aligned corpus. Several different software tools, incl. MULTEXT aligner and Vanilla aligner, were used for producing such corpora.

Bulgarian-Polish parallel corpus

The MTE model for corpus design and development is being used in the design of the first Bulgarian-Polish corpus (Dimitrova, Koseska 2009b). This bilingual corpus supports the lexical database (LDB) of the first experimental online Bulgarian-Polish dictionary (section 1.2.2).

The Bulgarian–Polish corpus consists of two corpora: a parallel and a comparable. All collected texts in the corpus are texts published in and distributed over the Internet and were downloaded from existing digital libraries. Currently the corpus contains about 5 million wordforms, among them 3 million in parallel texts, that represent mostly modern Bulgarian and Polish literature (the second part of the XXth century).

The Bulgarian–Polish parallel corpus includes two parallel sub-corpora *a core and a translated*:

1) *A core Bulgarian–Polish parallel corpus* consists of original texts in Polish – literary works by Polish writers and their translation in Bulgarian, and original texts in Bulgarian - short stories by Bulgarian writers and their translation in Polish.

2) *A translated Bulgarian–Polish parallel corpus* consists of texts in Bulgarian and in Polish of brochures of the EC, documents of the EU and the EU-Parliament, published in Internet; Bulgarian and Polish translations of literary works in third language (mainly English).

Some texts in the ongoing version of the Bulgarian-Polish parallel corpus are annotated at “paragraph” level. This annotation allows texts in the two languages (Bulgarian/Polish and *vice versa*) to be aligned at paragraph level in order to produces aligned bilingual texts. The “paragraph” level allows drawing a broader context in the two languages. This means that we get the opportunity – thanks to the broader context – to study more precisely the meanings of word-forms in both languages. This approach is more correct – we are not comparing "word" with

"word", we compare word-forms in a broader context (“paragraph” level), which allows us to obtain the word's meaning.

The Bulgarian–Polish comparable corpus includes texts in Bulgarian and Polish: excerpts from newspapers and textual documents, shown in internet, excerpts from several original fiction, novels or short stories, with the text sizes being comparable across the two languages. Some of the Bulgarian texts are annotated at “paragraph” and “sentence” levels, according to CES (Ide 1998).

The advantage of processing a bilingual parallel corpus is to obtain context specific information about syntactic and semantic structures and usage of words in given language or languages. This bilingual corpus is useful to linguists-researchers for research purposes alike, for instance in contrastive studies of Bulgarian and Polish languages. Besides, the corpus can be used in education, in schools as well as universities in foreign-language instruction.

1.4 Grammars

For successful implementation of the tasks to be solved in the infrastructure of digital lexicographic resources, the availability of good and sustainable grammars of the languages involved is an advantage that can hardly be overestimated.

In this respect, advanced computerized grammars that could be used in NLP applications like machine translation, information retrieval and extraction etc. are especially valuable as they can be viewed as testing ground for the sustainability and quality of lexicographic resources developed. A typical example is the fully-fledged grammar of Russian created by IITP-RAS partner to the project and used within the ETAP-3 multipurpose linguistic processor (Apresjan et al. 2003).

The grammar covers the whole range of phenomena of the language style sometimes referred to as standard business prose (not accounting for specificities of highly colloqial speech, poetry, and sophisticated fiction), which is sufficient for analyzing texts belonging to scientific work, popular fiction, journalism, news etc.

In this case, the grammar is built on the principles of dependency syntax and consists of several hundreds rules called syntagms, each of which is designed to establish one particular binary syntactic link between two words of a sentence, labeled with names of syntactic relations (already outlined above in Section 1.3).

Ideally, every language concerned should be covered by at least one fully-fledged grammar, which should be used as a testing bed for lexicographic resources being developed. Considering the fact that, today, the two most advanced grammars of Slavic languages (ETAP-3 for Russian and Prague Dependency Treebank grammar for Czech) are based on dependency formalisms which provide adequate represent-

ation of the language structure; we believe that grammars for other Slavic languages should be developed on dependency syntax principles.

The first contrastive Polish–Bulgarian grammar is shortly presented in Koseska 2009a. This work is an extensive attempt at a semantic juxtaposition with a gradually developed semantic intermediate language, resulting from research on the structure of the grammatical rules of both languages and used for representing universal semantic categories, e.g., time, modality, definiteness and semantic case, which have not been described exhaustively in Bulgarian and Polish academic grammars.

The language form, its function, the value of a function and the meaning of a form are described (Koseska 2009b). Distinguishing between the form and its meaning in comparing the material of different languages (as is the case in the MONDILEX Project, which features six Slavic languages belonging to all three groups within the branch) will help avoid numerous substantial mistakes and erroneous conclusions. Hence dictionary entries should be verified and made uniform in this respect before they are "digitalized". A dictionary entry must by all means distinguish between a language form and its meaning.

The description of modality, in connection with a Petri net model, are presented in Times and Flow – the catalogue of descriptions of temporal and modal situations (see Koseska, Mazurkiewicz 2010). The catalogue aims at the creation of a language independent list of basic temporal situations. The list is a common framework for comparing language forms used for describing the listed situations. This monographic volume contains a collection of studies on temporal subjects, analyzed in accordance with the methodology of cognitive linguistics. A formal model of the grammatical structure of Bulgarian, Polish, and English are presented and illustrated with examples from the three languages. Thanks to the clarity and transparency of this type of formalization, achieved also through a spatial visualization of the developed models, conclusions from the analysis of temporal relations are available directly, which each time enables and facilitates their verification. The collection of problems discussed – temporal relations in their various variants – represents one of the key issues of linguistic semantics. Theoretical and methodological proposals contained in the volume constitute, with respect to their interpretation and mapping, an important contribution to the contemporary scientific discourse.

The presented concepts and views on the temporal forms and their meaning are based on Bulgarian, Polish and English linguistic data, next enhanced by Russian. The formal model, called **the net model of tenses,** is based on Petri's idea of representing processes by nets consisting of states, events, and flow relation occurring between them. Examples serve to show how Petri nets can be viewed as

a universal tool (an intermediate language) for analysing and comparing natural languages. The version of the analysis presented includes examples from Bulgarian, Polish, Russian and English. The model covers three basic groups of Slavic languages (Bulgarian, representing South Slavic group, Polish, representing West Slavic group, and Russian, representing East Slavic group). English language serves as a mean for confronting phenomena occurring in the three Slavic languages mentioned above.

The aspectual meaning of a verbal form is important for many Slavic languages, whereas in English aspect is a grammatical category. This requires taking into account the aspectual and temporal meanings which are formalized in Slavic languages only. These problems are dealt with in the network-based description of temporal meanings in Bulgarian, Polish and Russian compared to English.

The samples of situations related to present tense, past tenses, future tenses and modalities are given, together with examples of describing them sentences in four languages: English, and three Slavic languages.

The book sets out formalism for the representation of semantics of natural languages in a way that is designed for professional people – linguists, computational linguists, computer scientists and other specialists – who need to use it. The *catalogue* could be used (after transferring them into simple program procedure) in machine translation, electronic dictionaries, and other automated activities regarding time and aspect in Slavic languages in contrast to English.

Part 2. Standardisation of Slavic Lexicographic Resources

Lexicographic resources, in particular machine readable dictionaries, lexical databases, and monolingual 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.

This section proposes several mutually reinforcing recommendations and standards 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.

2.1. Morphosyntactic Annotation in Slavic Digital Lexicography

Slavic languages are well known for their complex inflectional morphology. In order for Slavic digital lexicography to be made operational in a unified framework, the languages must share a harmonised set of morphosyntactic features and morphosyntactic descriptions. 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.

MONDILEX discussed morphosyntactic annotations in Slavic digital lexicography. This section presents the MULTEXT-East (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 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 CONCEDES 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 (T. Erjavec 2010).

The specifications now cover 10 Slavic languages, providing a good basis for a unifying morphosyntactic framework for digital Slavic lexicography and future

developments (Dimitrova, Garabík, Majchráková, 2009, Dimitrova, Rashkov 2009). The resources are available at http://nl.ijs.si/ME.

MULTEXT-East Morphosyntactic Specifications:

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.

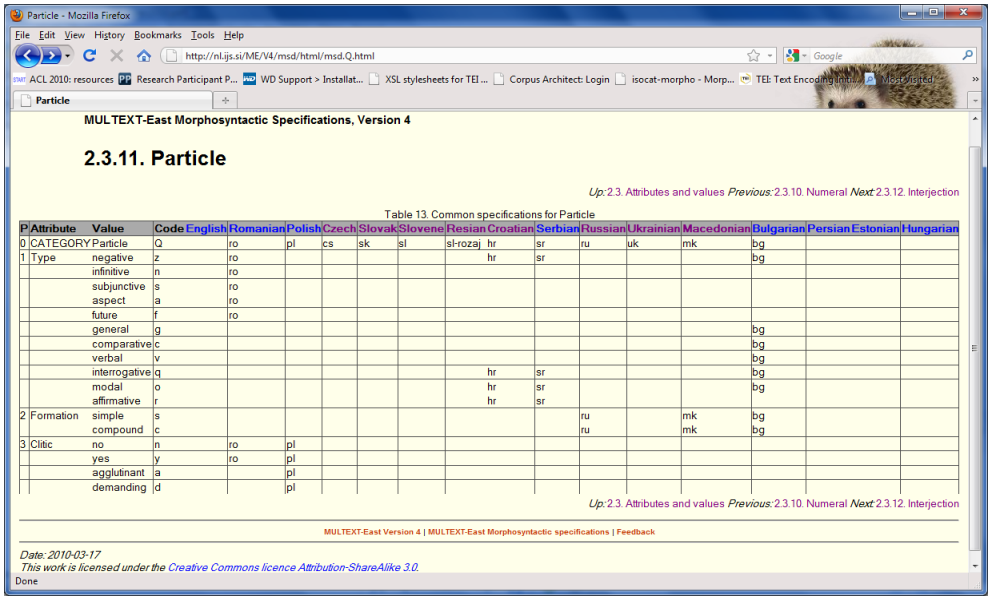
For example, they specify that the MSD Ncms is equivalent to the feature-structure consisting of the attribute-value pairs Category = Noun, Type = common, Gender = masculine, Number = singular.

These definitions are expressed in the so called common tables, which also specify for which languages each particular attribute-value pair is appropriate for. The following examples shows an attribute definition; it is formalised as a table (itself part of the category table) with the @role attribute giving the function of each row and cell.

```
<row role="attribute">
<cell
role="position">2</cell>
<cell
role="name">Formation</cell>
<cell>
<table>
<row role="value">
<cell
role="name">simple</cell>
<cell role="code">s</cell>
<cell role="lang">bg</cell>
<cell role="lang">mk</cell>
<cell role="lang">ru</cell>
</row>
```

```
<row role="value">
<cell role="name">com-
pound</cell>
<cell role="code">c</cell>
<cell
role="lang">bg</cell>
<cell
role="lang">mk</cell>
<cell
role="lang">ru</cell>
</row>
</table>
</cell>
</row>
```

The Figure below gives the full specification for Particle in the HTML rendering of the TEI P5 source. As can be seen, Particle has three attributes, each one assigned its position within the MSD string, each attribute then defines its values, and each values is given a code, and marked with the languages distinguishes this attribute-value combination i.e. feature.



Particle - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://nl.ijs.si/ME/V4/msd/html/msd.Q.html

Particle

MULTEXT-East Morphosyntactic Specifications, Version 4

2.3.11. Particle

Up: 2.3. Attributes and values *Previous:* 2.3.10. Numeral *Next:* 2.3.12. Interjection

Table 13. Common specifications for Particle

P	Attribute	Value	Code	English	Romanian	Polish	Czech	Slovak	Slovene	Resian	Croatian	Serbian	Russian	Ukrainian	Macedonian	Bulgarian	Persian	Estonian	Hungarian
0	CATEGORY	Particle	Q		ro	pl	cs	sk	sl	sl-rozaj	hr	sr	ru	uk	mk	bg			
1	Type	negative	z		ro						hr	sr				bg			
		infinitive	n		ro														
		subjunctive	s		ro														
		aspect	a		ro														
		future	f		ro														
		general	g													bg			
		comparative	c													bg			
		verbal	v													bg			
		interrogative	q								hr	sr				bg			
		modal	o								hr	sr				bg			
		affirmative	r								hr	sr							
2	Formation	simple	s										ru		mk	bg			
		compound	c										ru		mk	bg			
3	Clitic	no	n		ro	pl													
		yes	y		ro	pl													
		agglutinant	a			pl													
		demanding	d			pl													

Up: 2.3. Attributes and values *Previous:* 2.3.10. Numeral *Next:* 2.3.12. Interjection

MULTEXT-East Version 4 | MULTEXT-East Morphosyntactic specifications | Feedback

Date: 2010-03-17

This work is licensed under the Creative Commons licence Attribution-ShareAlike 3.0.

Done

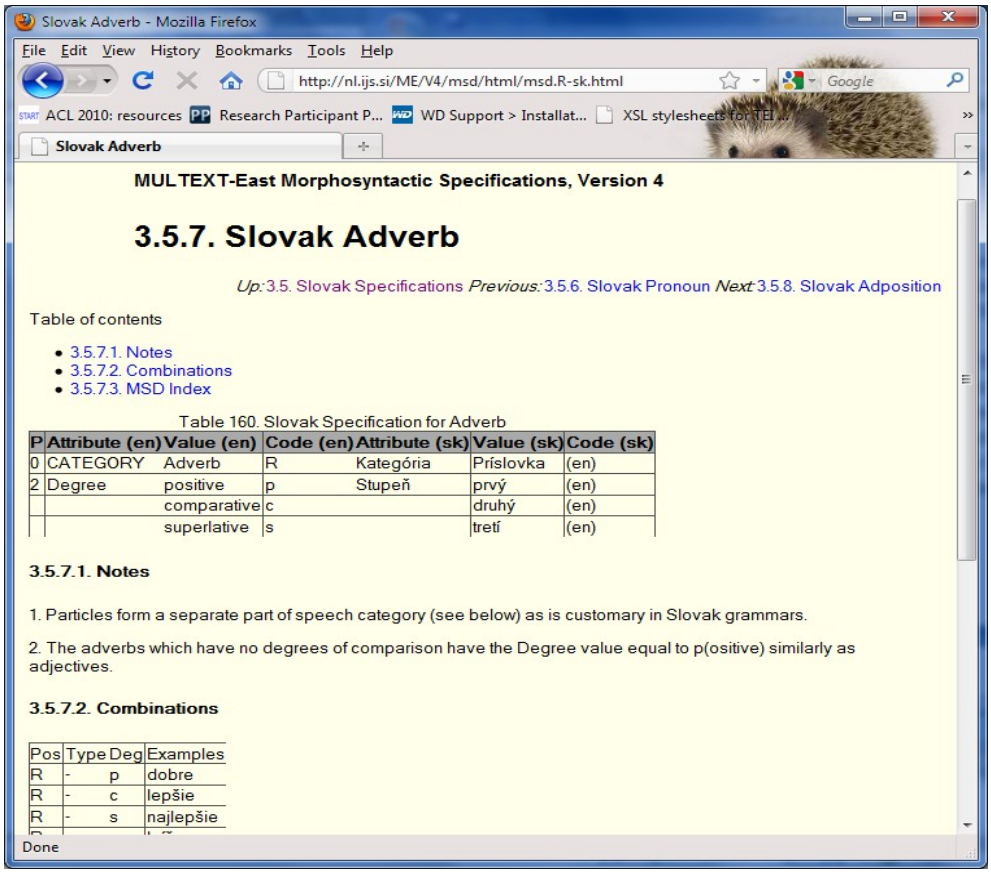
Example of a full specification for Particle in HTML

The second main part of the specifications is the language particular sections. These, in addition to the introductory matter, also contain sections for each category, with the table of attribute-value definitions appropriate for the language. These tables can be automatically derived from the corresponding common tables, but also modified from them, a novelty in Version 4. In particular, the position of the attribute in the MSD can be different from the common tables, leading to much shorter MSDs for particular languages. The tables can also contain localisation information, i.e. the names of the categories, attributes, their values and codes in the particular language, in addition to English. This enables expressing the feature-structures and MSDs either in English, or in the language in question. For example, they map the English MSD Ncmsn to the Slovene Somei i.e. samostalnik vrsta = občno_ime spol = moški število = ednina sklon = imenovalnik.

To illustrate, we give below the Slovak particular section for Adverb in HTML. The first part is similar to the common tables, except that only the features valid for Slovak are defined, together with their codes and positions. As mentioned only the

attribute and value names must be the same as in the common tables – the positions, however, can be different.

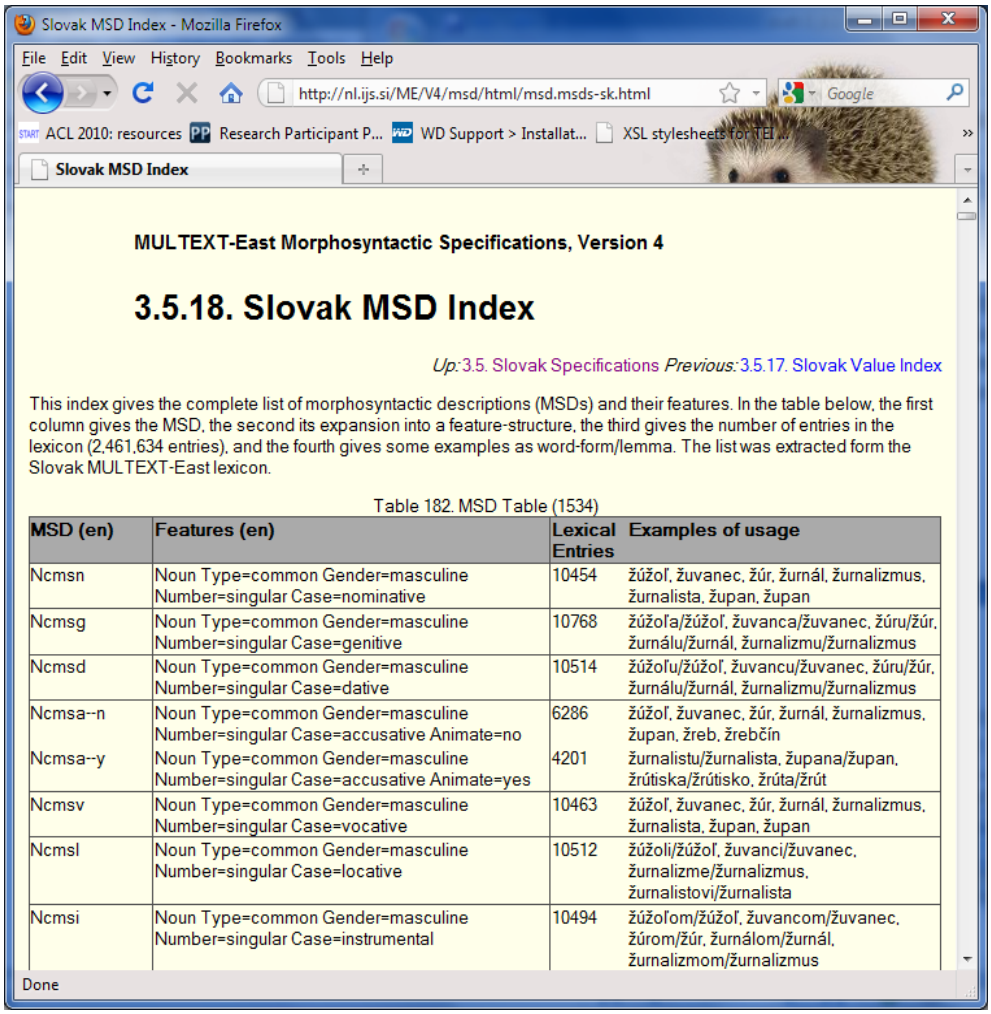
The table also gives the Slovak terms for the features; the code names (MSDs) are, however, not localised. The definitions for the language particular categories can also contain explanatory notes and combinations of allowed attribute-values.



Slovak terms for the features

Each language particular section furthermore contains an index containing all the valid MSDs for the language. Each MSD can be accompanied by explicative information, e.g. examples of usage. This index is the authority for the MSD tagset for the language.

In the Figure below the example of the start of the MSD tagset for Slovak is given. The MSDs are ordered according to the feature order (i.e. giving the paradigms in the conventional order for the language in question) and give, in addition to the required first column, also additional useful information about each MSDs, such as frequency of usage and examples.



Beginning of Slovak MSD Index

An important part of the specifications are the associated XSLT stylesheets, which allow for various transformations over the specifications. They take the specifications as input, usually together with certain command line arguments, and produce either XML, HTML or text output, depending on the style sheet. We provide three classes of transformations, the first ones to help in adding a new language to the specifications themselves, the second to transform the specifications into HTML, and the third to validate and transform a list of MSDs.

The specifications rendered in HTML largely follow the formatting of the original MULTEXT specifications, while various conversions of the MSD tagsets for each language are provided in a tabular format for easier use. So, for example, that

tables give for each MSD a canonical expansion into features, a sort-code for collating the MSDs in "linguist friendly" collation, or localisation equivalents.

As was seen, 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 presented here could sensibly be linked to other related attempts at standardisation of morphosyntactic features, in particular the ontology for descriptive linguistics GOLD[6] and the ISOcat Data Category Registry[7].

GOLD, the General Ontology for Linguistic Description (Farrar, Langendoen, 2003) is an effort to create a freely available domain-specific ontology for linguistic concepts, available at http://linguistics-ontology.org/. 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 form the MULTEXT-East specifications to GOLD, providing explication of their semantics.

The ISOcat Data Category Registry (Kemps-Snijders et al., 2008) is the Web service at http://www.isocat.org/ implementing the ISO standard 12620:2009 – Terminology and other content and language resources – Specification of data categories and management of a Data Category Registry for language resources. It provides an on-line registry, where, also terms from the domain of morphosyntax can be found. In the longer term it would be interesting to link up MULTEXT-East to isoCat (esp. as isoCat used the definitions of MULTEXT-East V3 in creating its initial registry) but the system and procedure is, for now, rather complex.

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

2.2 Corpus encoding

In particular, we discuss 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 proposed in the scope of CLARIN[8] initiative.

TEI is also suitable for encoding machine readable dictionaries, which is why these two resource types are discussed here and in the next section **2.3**.

TEI, however, 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". A proposal concentrating on the morphosyntactic level of description is proposed in the section **2.4**.

The TEI offers, inter alia, modules for modelling linguistically annotated corpora. However, more complex levels of annotation, such as syntactic and semantic annotation have several possible encoding in TEI, which aims to be more descriptive than prescriptive.

For common encoding of linguistic markup for Slavic digital lexicography we propose a particular encoding of three levels of linguistic annotation of corpora. Words are annotated by their MSD and lemma.

Syntactic annotation is stored in stand-off mark-up, with dependency labels marking pointers to the two connected tokens; the sentence id serves as the root.

Lexical semantic information concerns particular words of phrases, and connects them to an externally defined semantic lexicon, which can be expresses, say, in LMF.

We illustrate these particular points in the examples below, taken form the Slovene JOS corpus (Erjavec, Krek 2008), which is annotated by these three levels.

The semantic labels come from the Slovene wordnet lexicon (identical to the Princeton Word-Net synset ids) and are attached to the term element. Each term element is also marked for its head noun and possibly by a subtype indicating missing synsets (or specific enough hyponyms) in PWN.

The MSDs and dependency relations are given their Slovene label in the XML source – however, these can be interchanged with their English equivalents.

[8] www.clarin.eu

```
<s xml:id="F0020003.557.2">
  <w xml:id="F0020003.557.2.1" lemma="ta"
     ana="#Zk-sei">To</w><S/>
  <w xml:id="F0020003.557.2.2" lemma="biti"
     ana="#Gp-ste-n">je</w><S/>
  <term type="sloWNet" sortKey="kraj"
        subtype="missing_hyponym" key="ENG20-08114200-n">
    <w xml:id="F0020003.557.2.3" lemma="turističen"
       ana="#Ppnmein">turističen</w><S/>
    <w xml:id="F0020003.557.2.4" lemma="kraj"
       ana="#Somei">kraj</w>
  </term>
  <c xml:id="F0020003.557.2.5">.</c><S/>
</s>

<linkGrp type="syntax" targFunc="head argument"
corresp="#F0020003.557.2">
  <link type="ena" targets="#F0020003.557.2.2
#F0020003.557.2.1"/>
  <link type="modra" targets="#F0020003.557.2
#F0020003.557.2.2"/>
  <link type="dol" targets="#F0020003.557.2.4
#F0020003.557.2.3"/>
  <link type="dol" targets="#F0020003.557.2.2
#F0020003.557.2.4"/>
  <link type="modra" targets="#F0020003.557.2
#F0020003.557.2.5"/>
</linkGrp>
```

The proposed encoding is similar to the one used by most of XML encoded annotated corpora, except that unlike many, it uses TEI P5 as its basis. This has, inter ali, the advantage that other language resources can be modelled in the same scheme, from morphosyntactic specifications, to machine readable dictionaries.

2.3 Machine readable dictionaries

CONCEDE is an EC project whose aim was to harmonise the methodology, tools and resources for building Lexical Databases (LDBs) in a general-purpose document-interchange format, for six Central European languages: 2500-headword lexical databases for Bulgarian, Czech, Estonian, Hungarian, Romanian, and a 500-word one from the English-Slovene dictionary.

The project has produced lexical resources that respect the SGML (Standard Generalized Markup Language) guidelines for encoding linguistic corpora (Ide 1998) of the Text Encoding Initiative Dictionary Working Group (TEI-DWG), and so are compatible with other TEI-conformant resources.

The initial word lists for selection of headwords and word frequency were obtained from the MTE parallel corpus (section 1.3.2). The selection of headwords was made after word frequency and word class (POS) were taken into account, and the number of words there were in a given word-class and word-frequency band.

In order to achieve a harmonization of the LDBs according to the principal breakdown of lemmata to POS, the CONCEDE consortium decided on the following proportions: open parts of speech (nouns, verbs, adjectives, adverbs) - no more than 90 %, closed parts of speech (numerals, pronouns, conjunctions, prepositions, particles and interjections) – minimum 10% of the whole set of lemmata chosen. The LDBs were harmonized and a universal input formalism, the Document Type Definition (DTD), was created as a language-neutral, dictionary-neutral framework for the presentation of lexical information – CONCEDE DTD.

Under the CONCEDE project, an encoding scheme for lexicographic specifications was developed according to the standards for electronic dictionary encoding. The CONCEDE model for dictionaries encoding offers a standardized, understandable and intuitive structure and semantics of a dictionary entry (Erjavec et al. 2000, 2003).

In conformity with the CONCEDE model, all dictionaries use a common tagset, all were encoded according to the TEI. The hierarchical structure of the dictionary entry is a well-formalised tree-structure. The content of the CONCEDE entries is based on the information in published dictionaries for each of the six languages.

The first Bulgarian machine readable dictionary (Dimitrova 2008, Dimitrova 2009b) was created as a LDB of CONCEDE. The entries are equipped with lexicographic specifications for the Bulgarian language in TEI-conformant SGML. The electronic dictionary is based on the Bulgarian Explanatory Dictionary (BED) (Andreychin et al. 1994). The Bulgarian CONCEDE LDB developed in the project contains 2700 entries. The entries in the Bulgarian LDB retain the structure of the original paper dictionary as much as possible.

The entry with headword **стен|а** *//wall//* from the printed BED:

стен|а *ж.* 1. Отвесна, странична част на здание, помещение; зид. *Зидам стена. Външна стена*. 2. Висока каменна или тухлена ограда. *Фернандес лежи в полята пред стените на Мадрид*. Вапц. 3. Вертикална странична част или ограждаща, вътрешна повърхност на нещо кухо. *Казан с дебели стени. Стени на кръвоносен съд*. ◊ И стените имат уши. Китайска стена - нещо, зад което не може да се проникне. Притискам някого до стената - поставям го натясно, в безизходно положение.

The corresponding entry in the Bulgarian LDB follows:

```
<entry>
<hw>стен|а</hw>
<gen>ж.</gen>
<struc type="Sense" n="1">
<def>Отвесна, странична част на здание, помещение; зид.</def>
<eg><q>Зидам стена.</q></eg>
<eg><q>Външна стена.</q></eg></struc>
<struc type="Sense" n="2">
<def>Висока каменна или тухлена ограда.</def>
<eg><q>Фернандес лежи в полята пред стените на Мадрид.
</q><source>Вапц.</source></eg></struc>
<struc type="Sense" n="3">
<def>Вертикална странична част или ограждаща, вътрешна
повърхност на нещо  кухо.</def>
<eg><q>Казан с дебели стени.</q></eg>
<eg><q>Стени на кръвоносен съд.</q></eg></struc>
<struc type="Phrases">
<struc type="Phrase" n="1"><orth>И стените имат
уши.</orth></struc>
<struc type="Phrase" n="2"><orth>Китайска стена.</orth>
<def>нещо, зад което не може да се проникне.</def></struc>
<struc type="Phrase" n="3"><orth>Притискам някого до
стената.</orth>
<def>поставям го натясно, в безизходно
положение.</def></struc>
</struc>
</entry>
```

Finally, an examination was carried out – a validation process of the CONCEDE LDBs, which takes two forms, “formal validation” and “content validation”. The formal validation was a matter of ensuring that the databases were valid SGML documents and has been done by means of a validating SGML-parser. The content validation of the entries required human intervention and was therefore performed manually.

2.4 Lexical databases (on the example of Slovene)

This section presents a **proposal for lexical encoding concentrating on morphological properties of words**, with special emphasis given on the rich inflectional properties of Slavic languages. The encoding format is an application of the ISO standard LMF, while the core lexical structure and morphosyntactic annotation come from the MULTEXT-East proposal. On the example of Slovene, we detail the representation of inflectional paradigms, regular derivational relations, variant spellings, etc.

The proposed lexicon format is encoded in XML, with the schema being based on the ISO standard "Lexical Markup Framework",[9] which is the last in long tradition of HLT standardisation projects, starting with EAGLES.[10] LMF is the ISO International Organization for Standardization ISO/TC37 standard for natural language processing (NLP) and machine-readable dictionary (MRD) lexicons. The scope is standardization of principles and methods relating to language resources in the contexts of multilingual communication and cultural diversity. The goals of LMF are to provide a common model for the creation and use of lexical resources, to manage the exchange of data between and among these resources, and to enable the merging of large number of individual electronic resources to form extensive global electronic resources (Krek, Erjavec 2009).

Types of individual instantiations of LMF can include monolingual, bilingual or multilingual lexical resources. The same specifications are to be used for both small and large lexicons, for both simple and complex lexicons, for both written and spoken lexical representations. The descriptions range from morphology, syntax, and computational semantics to computer-assisted translation. The covered languages are not restricted to European languages but cover all natural languages. The range of targeted NLP applications is not restricted. LMF is able to represent most lexicons, including WordNet, EDR and PAROLE lexicons.

LMF is composed of the following components:

- The core package which is the structural skeleton which describes the basic hierarchy of information in a lexical entry.
- Extensions of the core package which are expressed in a framework that describes the reuse of the core components in conjunction with the additional components required for a specific lexical resource.

[9] In November 2008 LMF became the international standard ISO-24613:2008. The Web page or LMF is http://www.lexicalmarkupframework.org/

[10] EAGLES, Expert Advisory Group on Language Engineering Standards: http://www.ilc.cnr.it/EAGLES/home.html

The extensions are specifically dedicated to morphology, MRD, NLP syntax, NLP semantics, NLP multilingual notations, NLP morphological patterns, multiword expression patterns, and constraint expression patterns.

The normative part of LMF is a set of UML diagrams, however, the standard comes with an informative annex giving a DTD according to which LMF lexica can be expressed in XML. This DTD could be used in developing the lexicon format for future development.

2.4.1 Basic structure of a lexical entry

An LMF lexicon starts with some meta-information, which we do not discuss here, and is then composed of lexical entries. We give a simple example of a non-inflecting entry below:

```
- <LexicalEntry id="LE_itak">
    <feat att="besedna_vrsta" val="členek" />
  - <Lemma>
      <feat att="zapis_oblike" val="itak" />
    </Lemma>
  - <WordForm>
      <feat att="zapis_oblike" val="itak" />
    </WordForm>
  </LexicalEntry>
```

As can be seen, a lexical entry is assigned an ID, which uniquely identifies the entry; in case several entries have the same lemma, the ID is decorated with a number, to distinguish homonymous entries. The lexical entry then specifies which part of speech it belongs to. More generally, the top level features contain all the invariant features of the lemma, such as gender for nouns.

Next comes the lemma form, with a feature specifying how the lemma form is written. The lemma is still an abstract form, not meant as a particular word-form to be found in text. Finally, the lexical entry specifies the word-form or word-forms that constitute its paradigm.

2.4.2 Inflectional paradigms

For inflected words the complete inflectional paradigm becomes part of the lexical entry, with each word-form being specified to its form and distinguishing features, as shown on the start of the paradigm for the lemma *čakati*:

```
- <LexicalEntry id="LE_čakati">
    <!-- Inflected forms of the verb "čakati"  -->
    <feat att="besedna_vrsta" val="glagol" />
    <feat att="vrsta" val="glavni" />
    <feat att="vid" val="nedovršni" />
  - <Lemma>
      <feat att="zapis_oblike" val="čakati" />
    </Lemma>
  - <WordForm>
      <feat att="zapis_oblike" val="čakati" />
      <feat att="oblika" val="nedoločnik" />
    </WordForm>
  - <WordForm>
      <feat att="zapis_oblike" val="čakat" />
      <feat att="oblika" val="namenilnik" />
    </WordForm>
  - <WordForm>
      <feat att="zapis_oblike" val="čakal" />
      <feat att="oblika" val="deležnik" />
      <feat att="spol" val="moški" />
      <feat att="število" val="ednina" />
    </WordForm>
  - <WordForm>
      <feat att="zapis_oblike" val="čakala" />
      <feat att="oblika" val="deležnik" />
      <feat att="spol" val="ženski" />
      <feat att="število" val="ednina" />
    </WordForm>
```

It should be noted here that it is easy to move from the feature-based encoding present in the lexicon to the MSD encoding used in corpora: for each word-form we take the unification of the (disjoint set of) features on the lemma level with those on the word-form level, arriving at the complete feature-structure, which is then, via the specifications or derived tabular files converted to the MSD.

2.4.3 Derivational relations

Derivational relations connect two or more lexical entries of which one is a morphological derivation of the other. The connection always goes from the unmarked lexical entry to the derivationally marked one, and is encoded in the lexical-entry level as the related form, containing a pointer to the ID of the related entry, as shown in the example below:

```
<LexicalEntry  id="LE_česen">
  <feat att="besedna_vrsta" val="samostalnik"/>
  <feat att="vrsta" val="občni"/>
  <feat att="spol" val="moški"/>
  <Lemma>
   <feat att="zapis_oblike" val="česen"/>
  </Lemma>
```

```
  <WordForm> … </WordForm>
  <WordForm> … </WordForm>
  …
  <RelatedForm>
    <feat att="idref" val="LE_česnov"/>
  </RelatedForm>
</LexicalEntry>
```

2.4.4 Variant spellings

Lemmas can have word-forms with the same features, but different spellings, either due to register or regional variation, or possibly common mistakes. The guide to when a certain, possibly non-standard form is to be included in the lexicon is based on frequency of corpus occurrence.

In these cases the form representation element is used, which appears under the word-form. The word-form itself gives the morphological features, while form representations give the spelling of the variant, together with the status of the variants and the number of occurrences attested in the reference corpus, as shown in the example below:

```
<WordForm>
    <feat att="število" val="ednina"/>
    <feat att="sklon" val="rodilnik"/>
    <FormRepresentation>
    <feat att="zapis_oblike" val="gejzirja"/>
          <feat att="norma" val="variantno"/>
          <feat att="pogostnost" val="24"/>
    </FormRepresentation>
    <FormRepresentation>
    <feat att="zapis_oblike" val="gejzira"/>
          <feat att="norma" val="variantno"/>
          <feat att="pogostnost" val="6"/>
    </FormRepresentation>
</WordForm>
```

2.5 Universal networking language

The Universal Networking Language (UNL) is a tool for global information exchange in computer networks (http://www.undl.org). It was originally proposed by Hiroshi Uchida in 1990s. It is not a language for direct oral communication, but a semantic interlingua, offering a formal way to record the meaning of a natural language text. The important aspect of UNL is that the words of UNL are unambiguously defined elementary concepts. The inventory of UNL concepts is infinitely extensible. Theoretically, it is able to accommodate lexical meanings of all words of any language. UNL provides unique identifiers for individual concepts, called Universal Words (UW).

Due to this fact, UNL UWs can be used as a pivot to record the lexical meanings of words in the monolingual and multilingual dictionaries developed for Slavic Languages and relate the words of different languages to each other. A tentative experiment performed within the MONDILEX project showed that UNL can be successfully used to start the development of bilingual dictionaries for language pairs that had no such resources in the past (Boguslavsky, Dikonov 2009).

This approach can significantly reduce the cost of and facilitate the development of new bilingual dictionaries. The initial set of raw data needs to be prepared only once for each natural language. It can be done in collaboration by teams of lexicographers who need not speak any other languages but their native language. The latter is important because it can be difficult to find a large team of experts for rare language pairs. Such experts are only needed for post editing of already assembled raw dictionaries.

An additional benefit is that UNL is already linked with lexicons of several major world languages beyond the scope of MONDILEX, including English, Spanish, French, Hindi, etc. which simplifies creation of dictionaries for these languages. UNL is also linked with other semantic resources, including Princeton Wordnet and IEEE Suggested Upper Merged Ontology (SUMO). UNL is supported by several Natural Language Processor (NLP) systems developed by researchers taking part in the global UNL project in Spain (Universidad Politécnica de Madrid), France (GETA CLIPS), Russia (IITP RAS), India (Anna University) etc.

Part 3. Software Environments for Digital Lexicography

3.1 Conceptual Modelling of Services for the Bilingual Lexicographic Systems

Principles of Designing

The development of the theoretical principles of bilingual dictionary systems design is called forth by the need to enhance information systems with the linguistic functions of translation, comparison, synchronization, cross-language adaptation. The main trends in development of the bilingual systems are increasing the number of directions of translation, improving the formatting quality of the textual information presented to the user and integrating lexicographic information from various resources.

A special role is played by the lexicographic systems designed to build lexicographic resources. Therefore, this section is dedicated to the review of the conceptual foundations of the toolkit supporting the bilingual lexicographic systems that are designed and developed in ULIF-NASU. Implementation of the principles of conceptual modelling used in the development of all systems of this class leads to the need to use the L-systems structures in the ANSI/X3/SPARK or just ANSI/SPARK architecture. The main components of the ANSI/SPARK architecture are used in the following interpretation:

$$ARCH\_LS = \{CM, EXM, INM; \Phi, \Psi, \Xi\},$$

where *CM* means the conceptual model of the lexicographic system *LS*. $EXM \in \{exM\}$ identifies a set of its external models, conforming to the conceptual model CM, and $INM = \{inM\}$ – the corresponding set of its internal models. A set of *CM* to *EXM* mappings is denoted by CM:

$$\varphi : CM \rightarrow exM, \text{ where } exM \in EXM;$$

respectively, $\Psi = \{\psi\}$ is a set of mappings of the *CM* into *INM:*

$$\psi : CM \rightarrow inM, \text{ where } inM \in INM;$$

$\Xi = \{\xi\}$ is a set of mappings of INM into EXM:

$$\xi (inM) = exM.$$

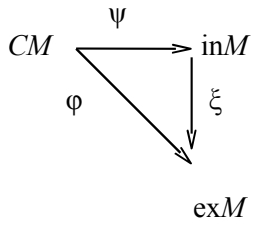
Let us interpret the architecture elements.

A conceptual model (conceptual level of presentation) of the subject area is a semantic model integrating notions of different experts in the subject field in an unambiguous, finite and consistent form.

The internal model (internal level of presentations) defines types, structures and formats of data presentation, storage and manipulation, an algorithmic base and the software environment in which the modules implementing the model must be integrated.

The external model (external level of presentation) reflects the views of the end users and, hence, application programmers, to the information system. It defines a set of tools enabling the authorised user to establish connection and manipulate the data provided by the internal level.

The mappings are constructed in such a way that the diagram:



is commutative: $\xi \circ \psi = \varphi$. The requirement of commutativity of the diagram is essential since it ensures consistency of all levels of the system architecture. In this case, it is assumed that all L-systems support tools, that require remote access, data synchronization and distributed user work, are designed and developed following the principles of the virtual lexicographic laboratories (VLL) (the concept of virtual lexicographic laboratory was first introduced by V.A. Shyrokov in his book "The Information Theory of the Lexicographic Systems", 1998).

The important features of the virtual lexicographic laboratories are:

- centralized storage and administration of the lexicographic data;
- interaction among all subjects and objects of VLL in the real time;
- isolation of some functionality from the end users. It allows the users to receive the most current information, but eliminates the possibility of unauthorized access and destructive actions.

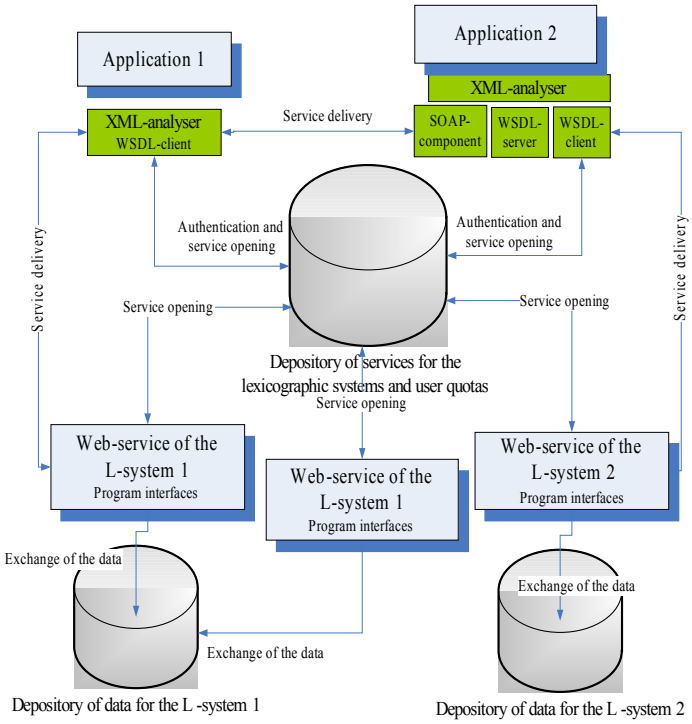
Such computer systems enable the linguists, who work in different organizations, different cities and even countries, to collaborate remotely in the framework of large scale linguistic projects. Moreover, the modern trends in the realm of computer communications and Internet give stimulus towards more interactive and dialogue-based lexicographic process.

Organization of Services

The development of such virtual systems follows the principles of the so-called service-oriented approach (Shyrokov 2009b, 2009c, 2009d). The complexity of interaction with lexicographic systems is defined by two seemingly contradictory requirements. On the one hand, the program interfaces that represent the functionality of these systems must be able to achieve a high degree of independence from each other and from the runtime. On the other hand, the need for integration requires interaction between the interfaces while preserving their internal autonomy.

The L-system has an isolated depository of data. The service part represents the program interfaces required to manipulate this data, i.e perform any kind of processing, filtering, transformation, etc. It is possible to have multiple service interfaces to the same data depository. The depository is an add-on part of the service. Its basic functionality is to filter the requests to services, managing user rights and quotas depending on the user's role. There are client applications that provide graphical interface for the user. One client application can communicate with multiple services, integrating functionality of several lexicographic systems.

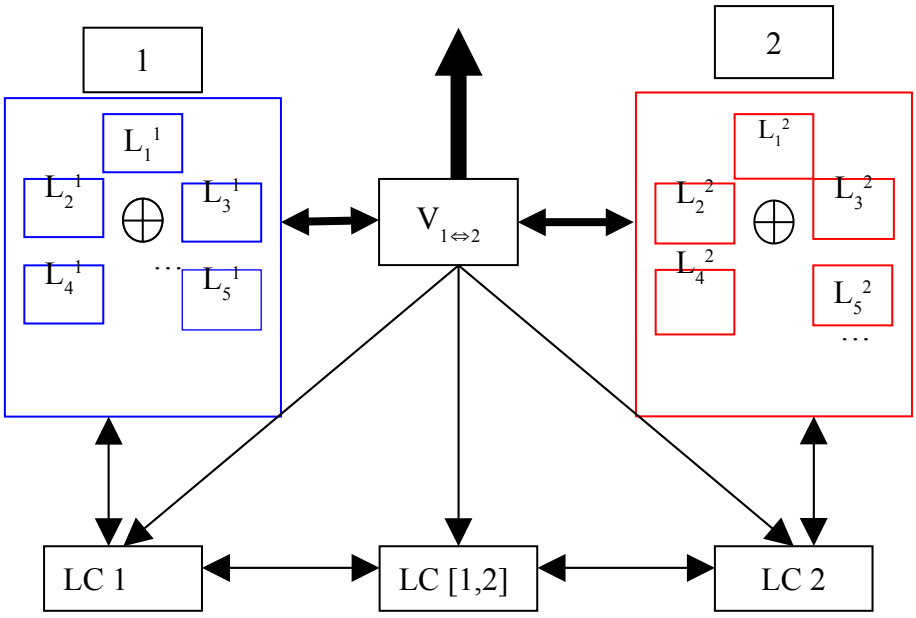
The following figure shows schematically the organization of interaction through the example of services of two lexicographic systems.



Service-oriented approach to designing VLL

A common approach towards designing bilingual systems

The overall conceptual scheme for designing an integrated virtual bilingual L-system is described in detail in (Shyrokov 2008). A simple example of the Bilingual L-system is shown in the following scheme:



where L_1^1, L_2^1, … are L-systems for the language 1, L_1^2, L_2^2, … – L-systems for the language 2, both systems are under the operation of L-system integration; $V_{1\leftrightarrow 2}$ – L-system – interface between 1 and 2; LC1, LC2, LC[1,2] – linguistic text corpora in the languages 1, 2 and the parallel one [1,2].

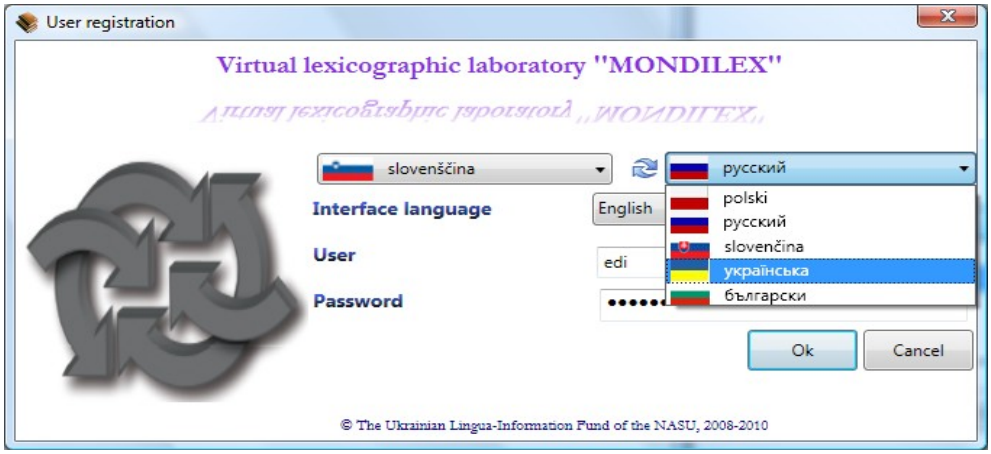
An optimal internal structure of the lexicographic data storage has been determined as a result of the analysis of a number of bilingual dictionaries.

The classes and objects of the bilingual L-system have been allocated, the procedure of unification of the basic concepts and abstractions have been carried out according to the milestones of microdesigning the system. The internal structure is shown in the next scheme.

The structure of the lexicographic system includes several optional elements that fully cover the content of the internal representation of the most bilingual L-systems. So this structure is used as a basis. Let us see the external interface of the bilingual system.

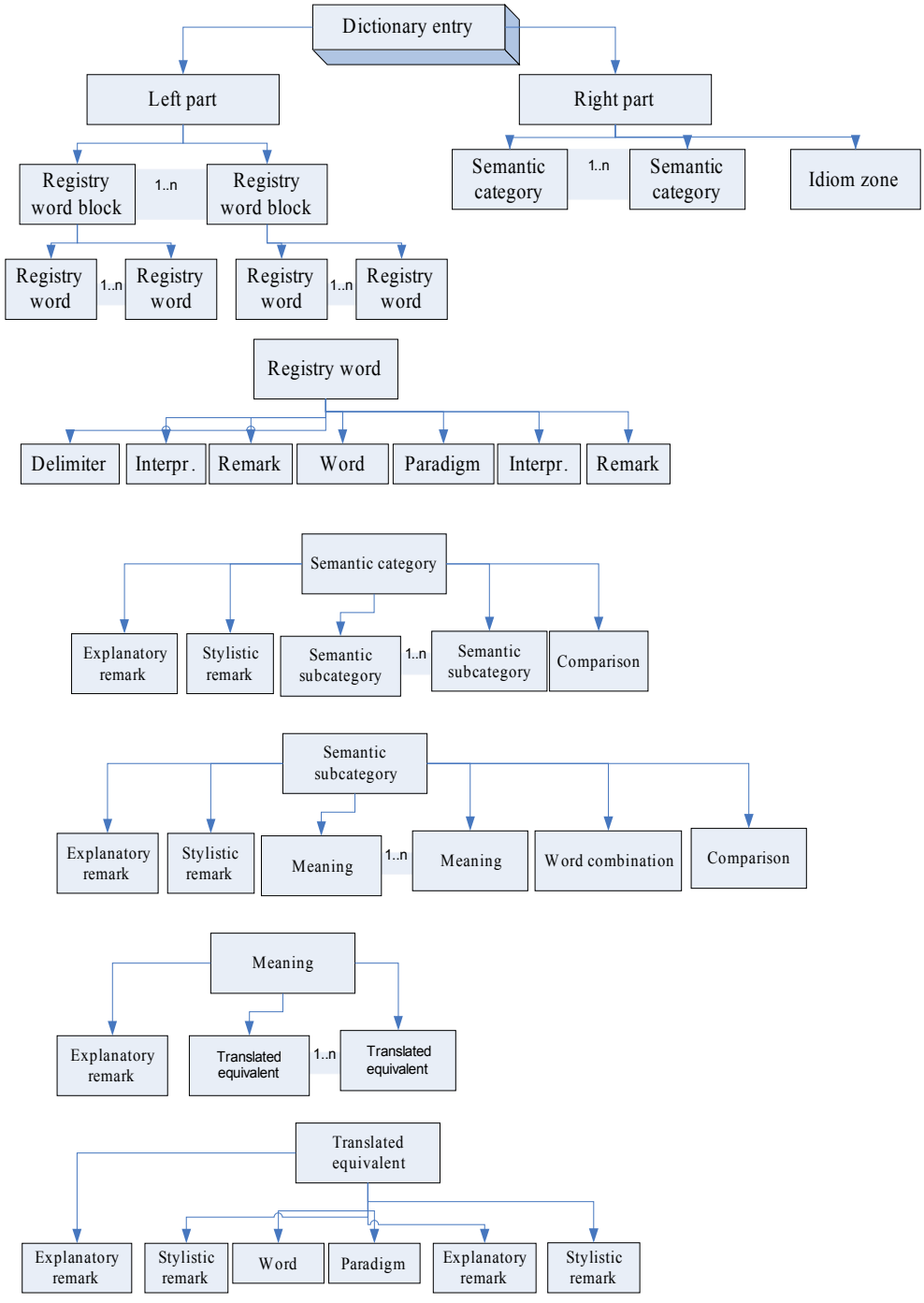
The interface language selection functionality becomes required for bilingual dictionary authoring software. It is designed to provide interface in at least two languages.

Authentication is a mandatory procedure, because the system is allows editing and deleting lexicographic data.

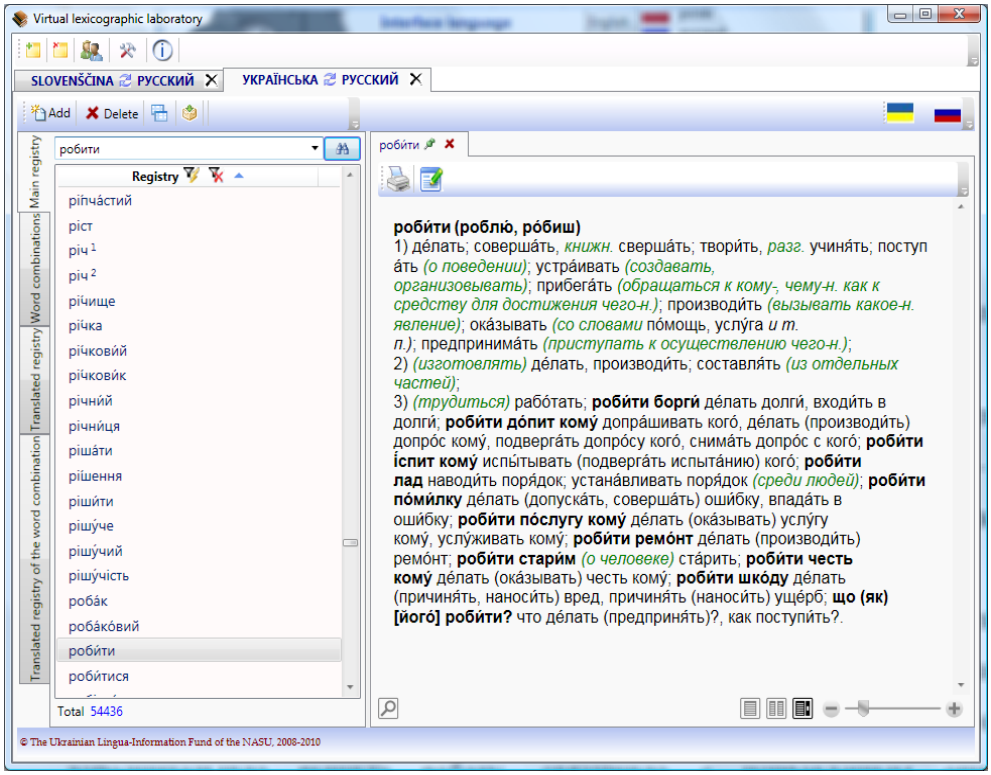


User's Authentication internal structure

The internal structure of the dictionary entry is shown at the next page.


Dictionary entry
Left part
Right part
Registry word block
1..n
Registry word block
Semantic category
1..n
Semantic category
Idiom zone
Registry word
1..n
Registry word
Registry word
1..n
Registry word
Registry word
Delimiter
Interpr.
Remark
Word
Paradigm
Interpr.
Remark
Semantic category
Explanatory remark
Stylistic remark
Semantic subcategory
1..n
Semantic subcategory
Comparison
Semantic subcategory
Explanatory remark
Stylistic remark
Meaning
1..n
Meaning
Word combination
Comparison
Meaning
Explanatory remark
Translated equivalent
1..n
Translated equivalent
Translated equivalent
Explanatory remark
Stylistic remark
Word
Paradigm
Explanatory remark
Stylistic remark

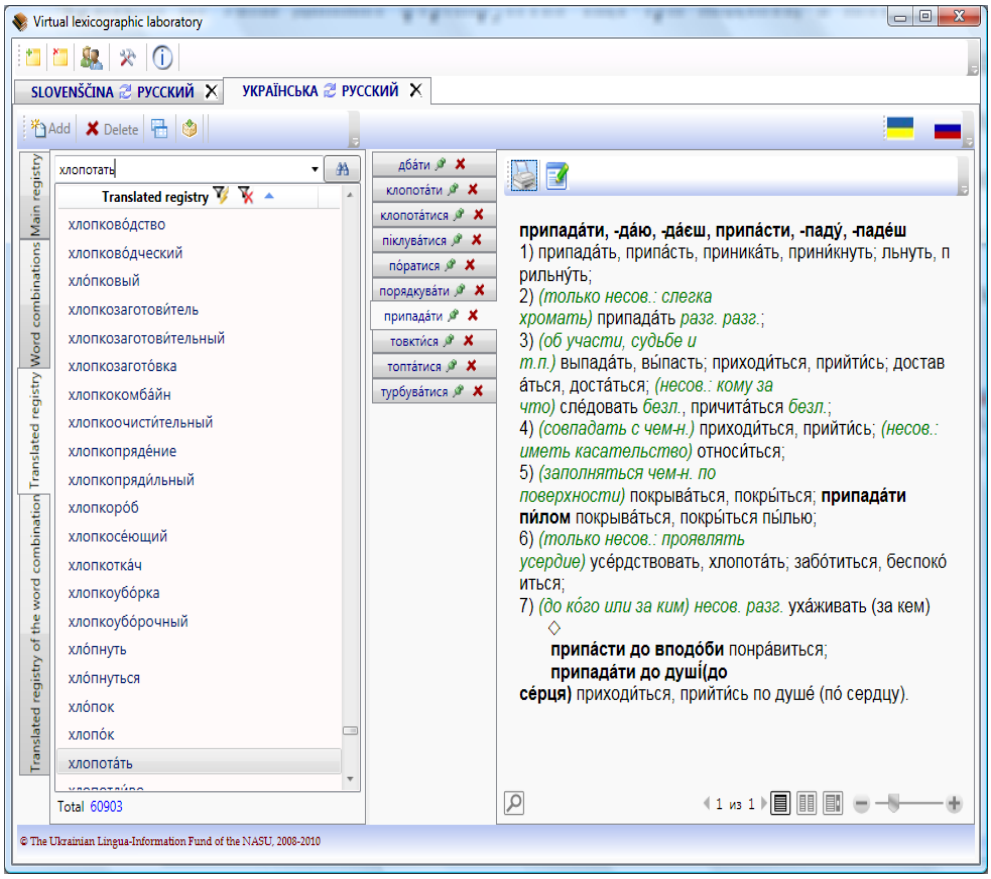
After successful authentication the main window of the program of the corresponding bilingual lexicographic system is opened.



External Interface of the System of Bilingual Dictionary

The user can load several bilingual systems in a single interface window. The registry of the system is presented in the left side of the main window. Selecting concrete registry unit allows viewing entries in the usual form, close to a printed book.

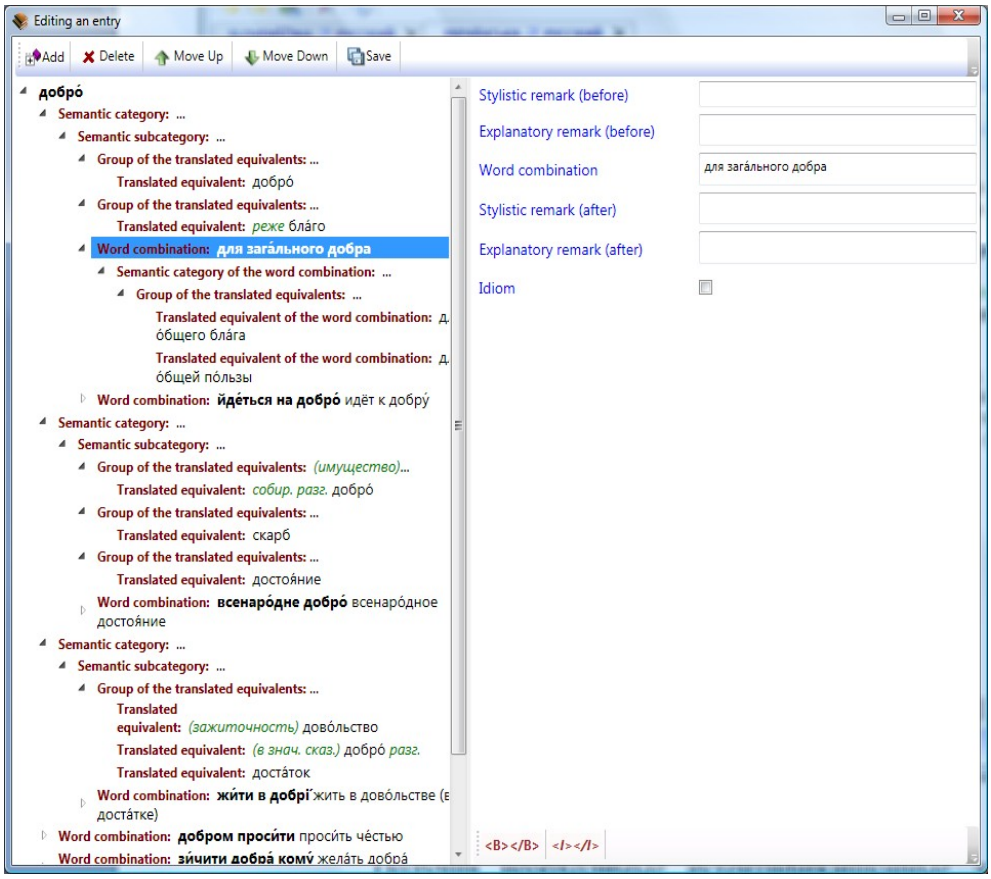
A feature of this bilingual system is that three additional indices are formed from the database in real time, namely: the registry of word combinations, the registry of translation equivalents and the translated registry of word combinations. This provides additional input to the dictionary entries and a step towards automatic reversal of the direction of translation for a dictionary. For example, the Russian word «*ячейка*» is mentioned in 3 Ukrainian entries «комірка», «осередок», «чарунка», Russian word «хлопотать» is mentioned in 10 Ukrainian entries:



Index of the Translated Equivalents

Accordingly, when the user selects a word combination, he or she receives one or more entries, which include the specified word combination.

The search functions, direct and reverse order sorting function as well as the filtering function are available for the registry. The filter function allows to allocate a part of the registry, which "begins with", "contains", "ends with" or "does not contain" some text. The entries are displayed as tabs, which change when selecting another entry. To view multiple entries at the same time, the user can “fixate” a tab and it will remain on the screen regardless of the current entry for as long as the user does not explicitly close it. Naturally, there are functions to adjust the font size, search within a chosen entry, change the type of entry display, print and edit the entry. In the edit mode the entry is presented as a tree providing direct access to any structural element:



Dialogue of Editing the Entry

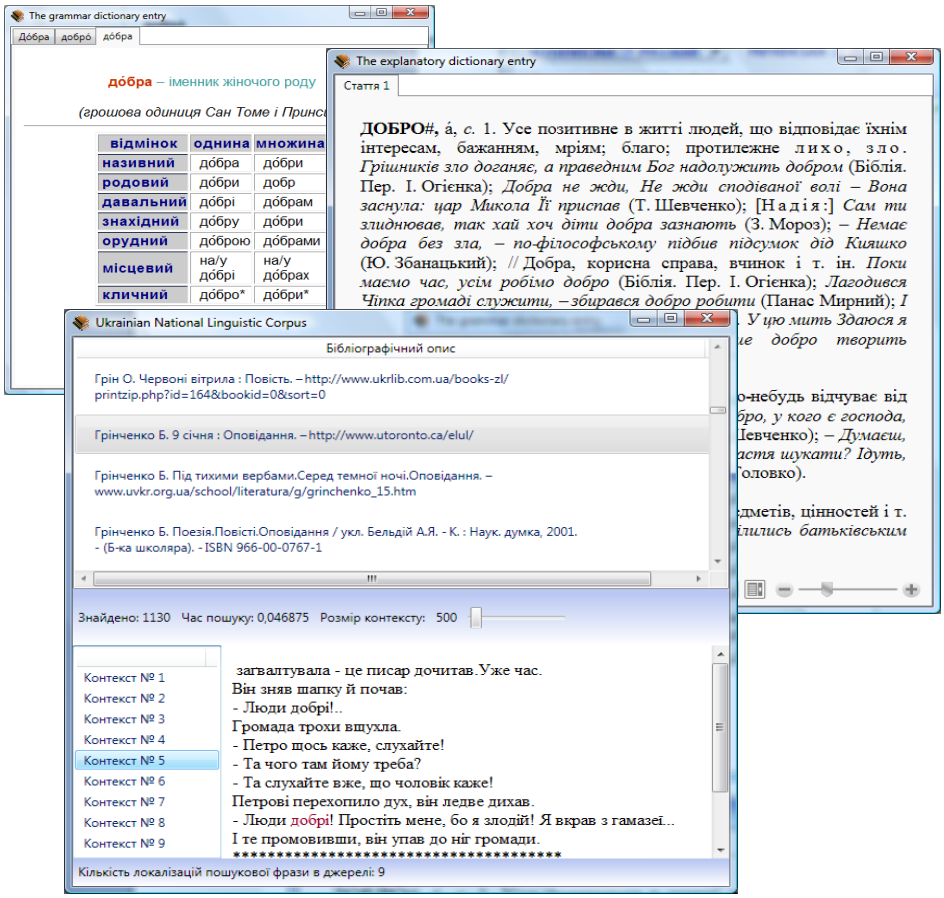
This approach to editing enables the user to monitor incoming data, prevent the structure violation and see the entry from another position.

The virtual lexicographic laboratory is deployed through a browser using the so-called ClickOnce technology. After authentication, the user works with VLL in the remote mode with a full range of functions that are implemented in the local version. Moreover, the client program version is monitored and is automatically updated when needed. All user actions are documented on the server. Therefore, the real picture of the lexicographic product development, the volume of the work done and the change tracking of the lexicographic data are available at any time. A pilot version of the Polish-Ukrainian Virtual Laboratory has been put into operation and runs between the Ukrainian Lingua-Information Fund and the Institute of Slavic Studies, Polish Academy of Sciences.

3. 2 Integration with Other Services of the Lexicographic Systems

Service-oriented approach allows integration with other lexicographic services. This connection with the following services is implemented in this software product: the explanatory dictionary (Ukrainian and Russian), the grammar dictionary (Ukrainian), the Ukrainian National Linguistic Corpus.

The context menu of the registry units gives access to a grammar dictionary, which provides word forms, an explanatory dictionary, and word contexts from the UNLC (for example the word "good" in the next figure). Note that the explanatory dictionary, grammar dictionary and corpora lookup services are not required for the bilingual dictionary authoring tool to function. The client access programs, which provide extended functionality, exist for each of these services. This feature allows implementing the interface schemes with arbitrary set of service functions.



Integration of Services of Different Lexicographic Systems

3.3. Software Environments for Creating Digital Dictionaries

Problems of the computer realization of dictionaries

The main problems related to the computer realization of dictionaries arise from the fact that they are simultaneously treated as text and as databases. They obviously look like text and have common points with other types of text. However, users do not normally read dictionaries, from A to Z, as they do with the majority of texts, but rather use them to obtain specific information through a given key (in this case a headword). The information associated with this key can include: pronunciation, grammar information, definitions, etymology, etc. Electronic dictionaries are capable of fulfilling users' requests many times faster than paper dictionaries, as well as of providing the possibility to find all entries whose headwords satisfy the user-defined criteria. Despite the fact that dictionary entries resemble a text on the screen, the internal representation of electronic dictionaries is a database.

Dictionaries are among the most complex text types because of the high level of structuring and information content. A dictionary entry – in terms of structure and content – is a complex unit and a structured object which uses different abbreviations and structural units in order to present the whole information succinctly. The structure of dictionary entries varies a lot within the dictionary as well as between different dictionaries. The external structure (text formatting and presentation) does not completely determine the internal structure (information content in the database). There is a great diversity of hierarchical structures: in some dictionary entries the hierarchy organization of their structure may be deeply embedded (i.e. it allows many levels), whereas in other cases some structural elements from this hierarchy may be missing. In spite of these variations some strict and constant structural rules exist so that the dictionaries can be understood by their readers. All these specific features make the database supporting the dictionary logically complex and difficult to create.

The build-up of electronic dictionaries is a complex and strenuous process, associated with several difficulties: (1) Lack of a sufficient number of formal models that allow words to be divided into formal language classes and a given word to be automatically included in one or another class. Electronic dictionaries can be created by ways of manual input of the dictionary articles – a process through which paper dictionaries are converted into a digital form (also possible with a scanner) or new dictionaries are prepared for printing. Such dictionaries, known as "machine-readable dictionaries" are different from their paper counterparts mostly in that they exist on magnetic carriers as files and can be processed as files. They follow a certain order and the articles have a concrete structure. As they are meant to be

used by a human, their disadvantage from a computer point of view is that they are not sufficiently formalized (formal structures are missing from their descriptions) and the extraction of knowledge from them requires the development of special computer modules. (2) A great variety of structures and content, which presupposes a conflict between universality and detail. The conflict between universality and detail is particularly strong in the case of dictionaries due to the large diversity of their structures and content, which turns the creation of a standard for dictionary encoding into a major challenge. In order to avoid this conflict the TEI workgroup created a universal standard for coding different types of dictionaries which encompasses fundamental principles of high degree of structure and diversity of dictionary entries (Ide, Sperberg-McQueen 1995).

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 wiki engine is based on the concept of "pages" – each page keeps separate information, is uniquely identified by its name and can optionally belong to one or more categories.

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 diff-like 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. Two of them are described in detail in this document – the reason is that they 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.

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

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. MoinMoin contains a built-in full text search engine, or it can use the Xapian libraries (http://www.xapian.org).

MoinMoin can be extended by writing macros or plugins – in particular, it could be extended by different parsers to accommodate specialised lexicography markup language, or to display terse, compact information in human readable form. MoinMoin also supports XML-RPC access to the data, a feature that can be potentially interesting in view of eventual integration of the database into external linguistic resources.

MediaWiki

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.

Automated database processing

There are several options for automated data modification. First and most obvious is to access the SQL backend directly, reading and modifying the tables. However, this method requires detailed knowledge of internal MediaWiki database structure, and modifying would have to be done with a great care, in order not to disrupt the database and introduce structural inconsistencies.

Much better way is to use a MediaWiki API, designed for a remote access. As the MediaWiki is probably the most widely used Wiki framework, there is a plethora of tools available for automated processing in various programming languages. However, there is even simple approach possible – WikipediaFS (http://wikipediafs.sourceforge.net/), a fuse-based (http://fuse.sourceforge.net/) file system that presents remote WikiMedia installation as a fake file system, so that the pages can

be read and written as simple text files, either for automated scripted processing or to be edited with an ordinary text editor. The advantage of WikipediaFS over using MediaWiki API is the availability of plain text, file system like view of the data, which makes it easy to use standard UNIX command line tools for text processing (sed, awk, grep, etc..

Recommendations for using the wiki-based system

Storing rich morphology information on the level of tens of thousands of words into a MoinMoin wiki-based system is viable, as long as special care is taken not to use features that scale badly with increasing the volume of data (Garabík 2008). The wiki is used as a source of data for various morphology-related automatized tasks, as well as a source for a human-readable morphological dictionary. Storing data in plain text format is perfectly suitable for information without a complicated structure. However, for richly structured data other options (XML) should be evaluated, together with the possibility of specialized modules providing easier user editing, while keeping all the advantages of a standard wiki system.

3.4 Software Environments for Creating Digital Corpora

In creating annotated text corpora, it is important that robust, sustainable and user-friendly software environment be available for the developers of such corpora. A good example is a language-independent system, called Structure Editor (StrEd: Iomdin, Sizov 2009), or structure editor, that is used for the preparation of the Syn-TagRus corpus, briefly outlined in Section 1.3 above (Iomdin, Sizov 2008).

This is a complex software environment aimed at 1) automatic generation of morpho-syntactic and lexical functional annotation of texts, 2) manual editing of annotation results, and 3) fully manual annotation. Automatic generation is only possible for texts in natural languages that are supported by ETAP-3 linguistic processor (see Section 1.4 above). At the moment, these include Russian and English, but can be extrapolated to other languages provided that grammars for these languages are developed to reach a sufficient level of coverage. StrEd is oriented to languages with rich morphology, so it may be used for creating corpora of all Slavic languages.

StrEd has a number of different viewing options and dialogue interfaces that can be chosen by the annotator depending on the particular task he or she is performing at the moment. In particular, the annotator may view:

1) the whole text of the corpora;

2) a sentence as a table in which every line corresponds to a particular word of the sentence;

3) the syntactic dependency tree for a sentence;

4) dictionary information on a particular word of the sentence;

5) the discrepancies within the results of automatic tagging and manual tagging of a sentence – a very important feature enabling the annotator to correct the errors in the annotation but at the same time use them as feedback for the grammar underlying the automatic parser.

In order to diagnose non-trivial annotation errors, a powerful instrument, Intellectual Debugger (IntelDeb), was specially created as a feature of StrEd. It enables the human editor to verify, in one quick step, whether the current syntactic annotation of a sentence (probably the result of several human interventions) is compatible with at least one of the parsing in principle achievable through the automatic parser. As a matter of fact, IntelDeb can be considered as a specific parser which, unlike the regular parser, does not produce multiple parses of a sentence. Instead, if the IntelDeb finds that the structure being subject to verification is inadmissible, its goal is to diagnose the cause, or causes, of the situation as precisely as possible.

The underlying idea is to run the parser consecutively on all binary subtrees as presented by the annotation and see whether the existing syntactic rules and dictionaries permit the construction of such subtrees. The Intellectual Debugger algorithm checks all rules with regard to a specific syntactic link (there may be dozens of such rules) and all possible lemmas for the given pair of words, starting with the rules and lemmas cited in the annotation but gradually loosening the grip and resorting to other rules and lemmas if the current choice cannot be confirmed.

Roughly, the algorithm of IntelDeb operation consists of the following stages:

- Loading the structure to be verified and extracting the text of the sentence.
- Morphological analysis of this text.
- Checking whether a morphological parse exists for all words of the sentence. For missing parses, a diagnostic message is generated and a substitute word is chosen.
- Generating hypothetical syntactic links.
- Checking whether the required links exist for every word of the sentence. In case of a missing link, a diagnostic message is generated and a substitute link is formed. Links whose names do not coincide with the required ones are deleted.
- Launching the procedure of tree generation, checking for the required links and words at every step. If these are missing, diagnostic messages are generated and substitutes are formed.
- Launching the tracer for syntactic rules responsible for the production of the required links. If IntelDeb cannot confirm the correct structure, viewing the tracer operation step by step helps the annotator understand the causes of errors: in most cases, they are connected with errors in syntactic rules or dictionary entries.

As the result of IntelDeb processing of a tagged sentence, either the parse in confirmed, or diagnostic messages are produced which show unconfirmed morphological parsed or syntactic links. Another outcome of this processing is tracing of syntactic rules.

Tools for Language Technologies

The experience that the developers of the poly-functional multilingual processor ETAP-3 (Laboratory of Computational Linguistics of IITP-RAS) have gained using the Lexicographer's Companion shows that the system increases the lexicographer's output and precision. This is especially important when specialized entries are produced on a mass scale. Therefore the set of specialized lexicographic types should be extended in the nearest future. Also, the choice of correct parses of

pairs of translation equivalents should be improved, taking into account those cases where the lexemes as elements of those pairs are in citation form. It is to be expected that this software system may be of much help if applied to a multilingual lexicographic resource (Iomdin, Sizov, 2008).

Recommendations on Corpus Storage and Processing

As regards 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 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 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 corpus data also has to be available in a standard format. 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.

Part 4. Technological Platform for Research Infrastructure for Digital Language Resources and Research for Slavic Lexicography

4.1 Research Infrastructure for Digital Lexicography

On one hand, research infrastructure is a combination of research activity, specialized education, training and innovation that advances the knowledge and understanding across all scientific domains. On the other hand, research infrastructure is a set of large-scale or singular facilities, scientific instruments, distributed facilities and interconnected network, which are shared widely within and between scientific research communities. The process of identifying, designing, developing, constructing, managing and sharing such infrastructure is complex and costly. The term **e-infrastructure** describes the comprehensive infrastructure that is needed to address the complex, multi-disciplinary and cross-border needs of modern science. Such kind of infrastructure should address the tasks of storing, analyzing and processing enormous amounts of data and information, of enabling world-scale scientific collaborations and the access to and sharing of scientific resources and information regardless of their type and location in the world.

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 optimally the headword; acquisition of digitally presented information for the creation of a new digital dictionary (e.g. a multilingual one), etc. 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.

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 (HLT) 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 (Erjavec, Javoršek 2008). In view of the above, it is quite natural that the grid paradigm has started to be applied, albeit slowly and with some time lag as compared to other areas, to the area of HLT, especially to subareas that deals with the processing of large amounts of data, i.e. corpora.

Grid technologies give possibilities to transfer and exchange of tools and data with enormous volume (such as digital corpora and dictionaries); and to process unified data in different Slavic languages in parallel by the same tools. A network based on the philosophy and structure of the grid could provide a research infrastructure for effective exchange of multilingual resources and tools for their creation, support, and processing (Dimitrova, Pavlov 2008).

The relationships between some features of grids and lexicographic activities include the following:

- Typical objects of the grid and the language technologies (for example, electronic dictionaries and corpora) share some specifications, including the structural complexity of monolingual, bilingual and multilingual dictionaries, the large volume of the dictionaries, the internal structure of the dictionaries as a sequence of well-defined tagged-tree lexical entries, etc.
- The grid provides appropriate services that digital dictionaries require for the coordination and unification of existing digital linguistic resources and for their further cooperative development and enrichment in accordance with recent advances in the field and with international standards, while ensuring their reusability, interoperability and openness.
- The grid allows for the creation of an operational structure for the effective communication between the partners and with potential stakeholders, and will support the partners' cooperative efforts to attain the principal objective of the project.

The possibilities of the grid technology could provide for the creation of a general lexical data base with a rich system of links between forms and meanings of words, with the possibility of searching in any language provided with a digital dictionary.

The problems of the usage of new technological platforms like Grid, as a high-performance universal system for supporting language technologies, are connected with the problems of the compatibility and unification of data (in different languages and produced by different tools).

MONDILEX concluded that the compatibility of digital resources in Slavic languages (corpora, lexical databases, monolingual, bilingual and multilingual dictionaries) can be achieved through carrying out two major tasks: (1) development of standardised and unified lexical descriptions for Slavic languages to

annotate texts and word-forms in corpora; lexicons lines; dictionaries entries, head-words, etc., (2) use of language-independent programming tools for processing of language resources annotated in such manner.

The synchronisation and interoperability of tools require: (1) defining of common & domain-specific & repository "services", (2) common format & organisation of input files; (3) uniform way for presentation of the specific morpho-syntactic information for each language.

Distributed Tasks of Language Processing

The modern period of the society development has generated two scientific-technical revolutions: communicative and digital. The Internet has become its world incarnation. The World Wide Web now provides a global digital communication worldwide. However, it should be noted, that it decides mainly the information retrieval tasks. Apart from the entertainment and recreation functions of the Internet (films, video, Web-museums, e-libraries, music...), the main direction of the Internet is searching for information online.

The Internet search tools are based on the mechanisms of natural language. Therefore, even this direct function of the network has necessitated the development of the effective natural language tools of the Internet. This is how the Semantic Web has appeared. The main task of it is knowledge mining. This has intensified research and development in the field of cognitive linguistics and its technological applications.

The processing function of the network was developed in parallel with the information retrieval function. This is how the Grid has appeared, which was originally specialized in solving the super computational problem. Gradually the ability of Grid to real-time processing of the super large volumes of information has clarified.

The interaction of the standard Internet and Grid now is becoming more definite. We can confidently predict that the computational component of the Grid will increase, and the information retrieval and processing functions of the network will become more integrated. Undoubtedly, this integration will sooner or later lead to the emergence of a new quality.

In what ways will the integration of information retrieval and processing problems most likely be expected? And what will the role of language be?

It is expected that gradually the integration of the knowledge domain languages and the natural language will take place in a general conceptual representation. Currently the knowledge domain ontology language, in particular, the construction of linguistic ontologies, seems to be such a language.

The next point is connected with modeling of the linguistic communication structure. In particular, a deeper psycho-and neurolinguistic study of the language core and the periphery of the language communication (verbal and written), as well as a construction of the formal models and technological tools, are expected. In this connection, the tasks of studying the mental-linguistic communication and system connections in the triad of 'Information – Language – Intelligence' become actual.

Thus, the problems of creating knowledge Grid have a completely distinct set of linguistic tasks that follow from the above concept of knowledge and the role of linguistic structures in its definition. These problems are the following:

1. Statistical processing of the large text arrays (written and oral).
2. Understanding the natural language.
3. Modelling the images, metaphors, and metonymy.
4. Automated construction of the classifications, ontologies, thesauri.
5. Finding logical-linguistic defects in the texts and their solutions.
6. Conceptual scheme construction, lexicographic, conceptographic and ontographic processing of the multilingual texts.
7. Cross-language adaptation and natural language translation.

This complex of problems requires quite large network computing resources for its solution.

4.2 Virtual lexicographic system – technological platform for research e-infrastructure for digital lexicography

The process of virtualization of the lexicographic systems can take place in the process of functioning of the aggregated lexicographic systems in real socio-technical environments. This happens when the 'subject area' (X) – a carrier of the super system of lexicographic systems – has distributed system characteristics and is parameterized with a structured set of system (network) addresses, like those that are adopted in the Internet. Then every element $x \in X$ becomes a function from a tuple of addresses: $x = f(a_1 a_2 \ldots a_n)$, i.e. a lexicographic system $ELSx[\mathbf{L}]$ becomes a virtual object, distributed in the physical space that is represented with a point $(a_1 a_2 \ldots a_n)$ in the space of network addresses. Moreover the agreement of the related data models on the conceptual and internal levels is not required (although this agreement could be very useful). The agreement at the level of external models is only necessary, particularly at the level of network protocols, ensuring the minimum integrity of the virtual lexicographic system and the possibility of its identification as a single object.

Such virtual lexicographic system can be used as a virtual lexicographic laboratory (VLL for short: Rabulets 2009), which provides facilities for performing joint lexicographic project by various institutions distributed geographically and even by different countries. The existing communication infrastructure of the Internet is fully capable to provide the necessary bandwidth, the normal work of the VLL within its functions planned. This can be achieved by applying the systems engineering of the so-called Service-Oriented Architecture (SOA).

The main issues that are faced when creating the distributed systems like VLL are:

- Heterogeneity of modern information systems;
- Metadata exchange between the systems of different manufacturers;
- Data exchange between the systems that differ significantly using different data formats;
- Large volumes of data transmitted between the systems;
- Guarantee of message delivery;
- Routing the messages and addressing the 'end points';
- Process coordination;
- Service interaction security.

Let us consider the approaches used to solve these issues, and the principles for creating VLL. It was decided to develop VLL on the basis of Web services (one of the implementations of the SOA-applications).

The Web services (WS) are positioned as a universal technology for binding the significantly heterogeneous systems. It is based on several standards: XML to describe data, SOAP to transfer information from one system to other, WSDL to describe services (including the tasks of types of the input and output data) and UDDI to store and provide WSDL-descriptions on request.

These standards are enough for creating a relatively simple system. But any non-trivial solutions (as a rule, they are necessary in a corporate environment) require the use of such things as guaranteed asynchronous message delivery, transaction management, data encryption forwarded between the systems, and provision of their authenticity. All these areas are somehow close to WS. Some add-in of various specifications is actively created, allowing entering these technologies to the world of WS.

SOAP (Simple Object Access Protocol) is a transport protocol, a remote call of the functional. This protocol is designed for organizing interaction of the distributed systems using asynchronous exchange of the XML-formatted documents (XML Infoset is applied). Such documents have three parts: an envelope (wrapper), title and body, the general purpose of which is clear from their names.

Such distribution is caused with the fact that SOAP creates its virtual transport environment. SOAP-message is able to follow the route that includes several units, each of which can make changes to it or process it somehow. The status of these changes is reflected in the message header blocks. The title is an expansion mechanism, which allows sending data in the SOAP-message that is not actually the main workload (for example: directives and/or context information needed for message processing). This allows expanding the messages with a method specific to a particular application. Another large required section is ‘body’. It contains the XML-block with the information that should be delivered to the end recipient. Both these sections are contained within the envelope.

SOAP is a simple "bridge" that provides application interaction. It has a paradigm of the unidirectional, not supporting the integrity of this messaging state. Therefore, additional means providing the crossings of the firewall border, multiunit routing, guaranteed delivery, are required to create systems with complex sequences of information exchange. However, SOAP defines the infrastructure within which an infrastructure private for each application can be described in a relatively unified form. In addition, the general principles, by which the binding of SOAP-messages to an abstract transport protocol can be performed, are set out in the standard. The general scheme of creating the SOAP-shells for RPC-oriented interfaces (Remote Procedure Call) is described; the particular mechanisms are given; and the particular realization of the method for processing SOAP-messages is set as content of GET and POST commands for the HTTP protocol. The binding

to a particular transport protocol allows reducing the amount of programming, needed for writing a SOAP-based application, and reducing the traffic amount. In other words, some information is removed from the original message and placed in its packages by means of the transport protocol at the point of departure. And it is reconstructed in its original form at the point of receiving a message.

For example, the HTTP protocol has already the means for providing message correlation (i.e., the means for logical binding of request and reply), and developers do not need to be anxious of the correlation request-reply. The binding to HTTP also allows making Web services more relevant to the general style of WWW and passing error messages more clear. The service of the class 'read only' can be identified with some address URI in the Web and give the SOAP-formatted information at the command GET, which has no parameters. But this binding is valid only between two neighboring nodes that support the transport protocol.

Typical SOAP-message

```
<s:Envelope xmlns:s="http://www.w3.org/2003/05/soap-envel-
ope" xmlns:r="http://schemas.xmlsoap.org/ws/2005/02/rm"
xmlns:a="http://www.w3.org/2005/08/addressing">
<s:Header>
<r:Sequence s:mustUnderstand="1">
<r:Identifier>urn:uuid:238b448e-3c97-47ec-bf9f-478333000ff2
</r:Identifier>
<r:MessageNumber>4</r:MessageNumber>
</r:Sequence>
<r:SequenceAcknowledgement>
<r:Identifier>urn:uuid:870db09b-33df-47d9-abb0-
33d3c422328d</r:Identifier><r:AcknowledgementRange Lower="1"
Upper="4"/>
<netrm:BufferRemaining xmlns:netrm="http://schemas.mi-
crosoft.com/ws/2006/05/rm">8</netrm:BufferRemaining>
</r:SequenceAcknowledgement>
<a:Action
s:mustUnderstand="1">http://ulif.org.ua/services/expl/IDict-
Connect/GetServerNameResponse
</a:Action>
<a:RelatesTo>urn:uuid:6072ec60-56c0-47ce-895a-
96fe39333c19</a:RelatesTo>
</s:Header>
<s:Body xmlns:xsi="http://www.w3.org/2001/XMLSchema-in-
stance" xmlns:xsd="http://www.w3.org/2001/XMLSchema">
<GetServerNameResponse
```

```
xmlns="http://ulif.org.ua/services/expl">
<GetServerNameResult><servername><n>mainulifserver</n><d>осн
овний сервер УМІФа</d></servername>
</GetServerNameResult>
</GetServerNameResponse>
</s:Body>
</s:Envelope>
```

WSDL (Web Services Description Language) is a description of service properties. WSDL describes the services as some abstract resources that can accept the documents of certain types at the input and initiate sending the documents of other types. WSDL defines service from two viewpoints: the abstract and concrete. At the abstract level the service is defined in terms of messages sent and accepted, which are described by means of XML Schema in the form irrespective of the concrete transport protocol. At the concrete level the bindings to the transport formats and points of physical placement are defined.

According to this standard the WSDL-description of the service consists of five parts:

1. The data types used by the service are described using XML Schema notations (section <wsdl:types>).
2. The description of the input WSDL-messages (<wsdl:message>) is set consisting of the elements that have types described in <wsdl:types>.
3. The ports are described (<wsdl:portType>) – their names, the names and specifications of operations allowable to them (<wsdl:operation>). Each such operation is characterized with a triple of messages – input, output and failure. Four types of operations are set in the standard: unidirectional, request-reply, confirmation-reply and messages (the latter two are the inversion of the first two). Respectively, the WSDL-port can be unidirectional and bidirectional. The information about failures is a feature of bidirectional ports.
4. The binding (<wsdl:binding>) to the transport protocol is set. There is a transition from a logical data model to an actual physical model. To describe the transition, the so-called SOAP-extensions of WSDL are used (the bindings of WSDL to HTTP and MIME are set). Using these extensions we can simply specify to the server that to form a real SOAP-document, the bodies

of the WSDL-messages described should be copied to its body. The service address in WWW is also set here.

5. The service descriptions are grouped (<wsdl:service>) – the service name, port data, bindings and comment are combined in the form suitable for human perception. Using this section the service can be bound to several alternative mirrors.

UDDI (Universal Description Discovery & integration) is a standard for features and structure of the database of service descriptions. UDDI, SOAP and WSDL create three basic Web service standards. UDDI is a standard for internal device and external interfaces of the database (repository) that stores service description. It sets the data model and standardizes API, including Web service API. All descriptions in the database are stored as XML-records.

The latest version provides the replication of repositories with complex models of their subordination to each other, the creation of a repository of multiple nodes (and replication of data between them), the global uniqueness of results and keys, API of publications for descriptions and subscriptions to changes, means of ensuring the data integrity, internationalization of records, content encryption.

While UDDI 2.0 version was designed to support e-business catalogues, version 3.0 is focused on the internal use – to build enterprise systems within the ideology of Service-Oriented Architecture. Therefore it admits creating the registry of several types (public, private and with shared access).

To facilitate searching UDDi-registry offers a standard mechanism for classification, cataloguing, searching and managing Web services:

1. It allows setting different taxonomies (classifications) in one registry, i.e. an element can simultaneously be classified in different ways within different logical models;

2. UDDI allows expanding the number of ways for classifying any item to information publishers. It is possible to verify the compliance of element data to the classifier's requirements;

3. UDDI Inquiry API allows specifying a classifier and classification attributes in the search parameters, as well as connecting data of various search queries.

UDDI is based on WSDL and XML Schema.

Optimization of basic specifications

The standard form of SOAP is very inefficient technology in terms of consumption of the computational resources. For example, the message EDI (Electronic Data Interchange) has a length of 80 bytes, while a similar XML-message is 1.5 KB.

SOAP will give it the title and markup tags, and this will increase its size. If the SOAP-message body has multimedia, the situation becomes quite catastrophic. Here are some of the emerging problems:

1. The inclusion of binary data into the message body requires additional operations to encode it to the Base64 format and decode back. This leads to excessive consumption of CPU resources, and excessive widening of the message size;
2. The inclusion of other XML-documents and their fragments into a SOAP-message - extremely complex operation, especially if the XML-fragments use a different character encoding;
3. Although SOAP-messages are self-marked, specific data block can be detected only after viewing the entire message. This means a significant growth of capacity on the computing resources.

SOAP 1.2 Attachment Feature describes the abstract model of forming SOAP-messages with attachments. It solves the first two problems listed above, entering a model of forming complex SOAP-messages (SOAP envelope plus attachments). The specification describes the abstract complex structure consisting of the main part with SOAP-messages and related secondary parts – attachments with multimedia data. Each such structure is characterized by one or more URI-identifier used for referring to it from other parts. The names SOAPMessage and SecondaryPartBag are assigned to the main and secondary parts on some basic URI.

The complex structure is neither a generalization of SOAP structural model, nor a generalization of SOAP envelope and does not define the main message processing model. This is just an abstract model, the basic "rules" that must guide the further implementation of SOAP bindings to specific transport protocol. In fact, the specification tends to bindings to the HTTP protocol. Here are the examples of possible use of SOAP Attachment Feature:

1. The main part and a JPEG-image can be encapsulated in one DiME-message (see WS-Attachments) and transmitted via TCP or HTTP;
2. The main part and a JPEG-image can be encapsulated in MIME Multipart/Related message and transmitted via HTTP;
3. The main part can be sent via HTTP without encapsulation, and a JPEG-image can be obtained on a separate request by the additional command HTTP GET.

The specification will also postulate some requirements following the questions of the data safety in the attachments. The processing of secondary units is determined

not with it, but with SOAP semantic structures specific to those programs for which the attachments are intended.

WS-Attachments and DIME is an optimization of the binary attachments in SOAP-documents and the format of their transmission.

XOP (XML-binary Optimized Packaging Mechanism) is an optimization of the XML-document volume and presenting the data (encoding) in it.

SOAP Message Transmission Optimization Mechanism is an optimization of the SOAP-traffic.

SOAP Resource Representation Header is a SOAP extension for traffic optimization.

Transport Layer of the WS-Architecture are specifications that define the rules of the guaranteed delivery for the SOAP messages.

Web Services Reliable Messaging Protocol (WS-ReliableMessaging). The specification describes a protocol that enables delivering WS-messages between the components of the distributed applications even in the case of software, hardware and network failures.

Routing the messages and addressing the 'end points.' The SOAP-message can pass through many nodes before it gets to the final recipient. Each node can not only perform the transport function, but also the processing – logging, auditing, verification. A protocol of the transport level is used between any pair of these nodes, but in general the protocols in the chain can be different. This means that the virtual transport infrastructure must be constructed on the level of SOAP. The protocols of routing the messages solve this problem. The routing enables virtualization of network resources when the user should not know what subject it communicates with in the internal network (for example, for security or load balancing).

WS-Addressing – allows to resolve logical service model and its physical implementation more (compared to the WSDL), specifies the routing rules.

Protocols of coordination for businesses-processes using the context. The distributed applications that solve the problems connected with business-processes support rely on Web-services more often. The complexity of these applications leads to the necessity of their structuring as separate units that perform complete pieces of work. The process that flows through such groups can be very long and require mechanisms to maintain its state. In other words, some general information (context) is required for the coordination of work within certain groups. At least this context must include the business-process ID that allows distinguishing it from another one of the same process.

WS-Coordination – the main specification, which describes the mechanisms of coordination for the Web-service operations. Other specifications including WS Transaction, WA Atomic Transaction and WS Business Activity Framework are based on it and extend it spreading to narrower fields of managing the atomic and business transactions.

Web-services and transactional systems. The protocols of action coordination are framework. They do not describe the order of calling the participants of coordination and do not impose any special restrictions on these calls. To perform the work more significant than the context transfer, they need, 'plug-in modules' as transactional protocols.

WS-Transaction (WS-Tx) is the most famous specification in the field of transactions. It has preceded WS Atomic Transaction and WS BA Framework for the environment described with the WS-Coordination.

WS Atomic Transaction is a subset of WS-Transaction specification that was marked out into independent specification and remade a little. It defines the protocols for short-lived atomic transactions in the environment described in the WS-Coordination.

Web Services Security (WS-Security) describes a basic layer for many other technologies in the field of Web-services security, namely how to ensure integrity, confidentiality and authenticity of an individual SOAP-message transmitted within the established sessions, context and security policy. The specification generalizes a number of early developments of IBM and Microsoft in this area including SOAP-SEC, WS-Security and WS-License, etc.

The Security Tokens represent a set of assertions made by the sender. The content of these assertions in the WS-Security is not specified because it depends on the specific implementation. The assertions can be username, key, permission for the operation, etc. The token can be certified (but not necessarily) by a digital signature. Verifying it, the recipient is ascertained that the sender knows the necessary key and so is credible.

Signing and Encrypting. To ensure the integrity of the message, WS-Security is based on the digital signature standard XML Signature. All signatures are stored in a unit. The specification allows attaching several signatures (even of different types) to a message relating to its various parts, including overlapping ones.

Web Services Trust Language (WS-Trust). To establish a secure connection between two parties, they must explicitly or implicitly exchange some mandates of confidence. And each party should have a mechanism to determine whether it could trust the mandate sent to it from its counterpart. WS-Trust defines the

means for this, namely: extensions of WS-Security that provide delivery, recovery and verification of the security tokens, and establishment of the trust relationship between domains, including the use of services of the agents.

Web Services Secure Conversation Language (WS-SecureConversation). This specification defines the extension of WS-Security and WS-Trust necessary to establish a secure channel, by which you can send many messages.

WS-Policy and Web Services Policy Assertions Language (WS-PolicyAssertions). WS-Policy defines the XML-grammar for describing the capabilities and characteristics of WS-system and requirements to its clients. Sets of similar descriptions (assertions) are reduced to the documents called policies. The assertions in WS-Policy are formed from the expressions and can be as simple declarations of availability for any properties in the service, as complex parameterized verifications of the incoming data for compliance with some criteria.

Web Services Policy Attachment (WS-PolicyAttachment). Typically, policies are not stored by themselves; they must be adapted to existing infrastructure. WS-PolicyAttachment defines a common mechanism for binding policy descriptions to the service descriptions, as well as its three specific variations: binding at the level of WSDL-types, binding to the elements of UDDI catalogs and binding to specific implementations of the services through WSDL-descriptions.

Web Services Metadata Exchange (WSMetadataExchange). This specification is designed to simplify getting metadata about the service connected to the remote 'end point'.

Authentication in the federal environment. WS-Trust and WS-Policy dictate that a resource should verify a set of assertions encoded in the security token of the request applicant, according to the policy adopted.

Attributes and Pseudonyms. The second important set of scenarios described by WS-Federation, concerns the use of service of attributes and pseudonyms (CAP, Attribute / Pseudonym services). CAP not only performs the client authentication, but it is able to expand the security tokens with some additional information about the client. The binding to UDDI as attribute repository is described in detail in this specification. A special model tModel is introduced for storing the attributes.

The virtual lexicographic laboratory based on the mentioned technological tools was created in development environment Microsoft Visual C # 2005 Professional Editions. It works in the operating system Microsoft Windows XP/2003 or Vista running Microsoft.NET Framework version 3.5. The complex has a layered architecture: the database server is responsible for communication with the

lexicographic database (LBD), functions of data acquisition and storage; the session server establishes sessions for individual users, manages privileges and installs access levels; the client software provides a user interface that allows users to edit, view entries and perform several other functions.

Thus, the software is designed to work in the network (both local and global, as the use of technology for creating the distributed service-oriented systems Windows Communication Foundation (WCF) for interaction between the specified levels of the complex allows its effective functioning in the Internet environment), where multiple users access LDB VLL simultaneously. Thus, depending on privileges the users can access the entire database or its part, can edit entries or only view them.

4.3 Grid Infrastructure Requirements for Supporting Research Activities in Digital Lexicography

Contemporary NLP tasks are rather varied; some of them require a lot of "pure" computing power, but many tasks, especially in the area of corpus linguistics, merely process large data files. From the software point of view, the tools used cannot be more diverse – they are often programmed in typical computer languages, like C or C++, but a lot of data processing is done in scripting languages, such as Perl or Python, and Java is increasingly popular, and more often than not, one specific task uses several different tools bound by short programs written in a shell script. The use of (high level) scripting languages even for the computing intensive tasks means that the analysis is less effective than it could be, but the ease of creating and maintaining the tools more than outweighs this particular disadvantage. From this follows than the tools are often fragile and require a specific environment, which sometimes means that even using a different GNU/Linux distribution that the one the software has been developed on can be a major problem.

The Grid environment, due to its initial connection with the use in High Energy Physics, predominately uses Scientific Linux CERN distribution (SLC) version 4 for the job computing environment (with a changeover to version 5 currently in progress). The ideal solution would be of course to put all the necessary NLP software into the execution environment (which is available at each of the computing nodes) and use the standard distribution. It is, however, sometimes much more convenient to use an operating system environment more suitable for the users and their tools. There are two possible solutions: to run under a chroot environment or to use virtualization. Both options are discussed below.

Virtualization

Chroot is a UNIX system that changes the effective root of the filesystem for the process and its children. The basic usage for chroot is twofold: it can be used to restrict untrusted (or potentially dangerous) processes from accessing the rest of the filesystem, or it can be used to run processes in a different filesystem environment (different filesystem layouts with different system executables and dynamic libraries). It should be noted that chroot does not offer true virtualization since isolation from the host system is not complete – in particular, system kernel, networking subsystem and process management are shared with the host system, so that the processes in the chroot environment cannot bind to sockets that are used on the host system (and vice versa), and if process management is to be possible in a chroot environment, the proc filesystem has to be mounted inside chroot environment, enabling the guest to access the information about host processes.

On the other side of the spectrum, there are complete virtualization solutions, emulating the guest system. These can emulate the CPU completely in software (approach commonly used in emulating vintage computers on modern operating systems, or when a computer platform switches the architecture), or run the guest machine natively, trapping and emulating only privileged or unimplemented instructions. Modern computer architectures usually offer dedicated hardware features to facilitate the implementation of virtual machines.

Then there are several different approaches that lie somewhere in between those two extremes, ranging from paravirtualization, which requires cooperation from the guest operating system kernel (in order to achieve negligible performance loss due to the virtualization), used e.g. by the XEN virtualization solution; to compartmentalization (i.e Linux virtual servers and OpenVZ), which divides the host operating system into different compartments with completely separated processes, network access and filesystems but sharing the same kernel; to vanilla kernel namespace support, which only separates user and process management (slightly extending chroot separation).

The virtualization techniques mentioned differ on performance impact (Padala et al. 2007) – ranging from none at all in case of a simple chroot or chroot with namespaces, over very little for OpenVZ-like compartmentalization to a more significant one for full virtualization. The specific areas of impact vary, too – while the raw CPU performance rarely decreases by more than a few percent (with the exception of complete software emulation of the guest architecture), I/O penalties are sometimes severe.

The best way to use the specific software is to install it inside a runtime environment which is made available to the jobs when submitted to the Grid. This is directly supported by the Grid infrastructure and requires no additional steps or privileges. However, at this time this requires a significant effort, since all the tools and their dependencies have to be compiled (or installed in a non-standard location inside the runtime environment) on the standard SLC distribution, which can be problematic if the software has many external dependencies.

Installing a chroot environment, on the other hand, enables us to avoid porting the software to the SLC distribution – inside the chroot, any reasonably standard GNU/Linux distribution and any necessary software packages can be installed. In addition, many of the commonly used distributions already have support for (at least partial) installation inside a chroot environment built in. But in the context of Grid infrastructure this solution has a significant disadvantage, since it requires support from the cluster administrator since chroot environments are not a standard feature of the Grid environment.

Using a complete virtual machine allows us to run a complete GNU/Linux distribution, with completely separate networking support and user management, including the ability to run processes with superuser privileges, and the ability to use filesystems otherwise not supported by the host system. But the main advantage is the possibility to run completely different operating system (therefore it is possible to use e.g. the tools available only for Microsoft® Windows® family of operating system, if one can get around their mostly point-and-click nature and run them noninteractively.). However, installing and using virtual machines requires not just administrator cooperation, but often also nonstandard host operating system extensions (such as special kernel modules). One of the more interesting virtualization systems in this context is User Mode Linux, which does not require any special host support, runs as an ordinary user process and provides a complete guest Linux kernel environment. Unfortunately, guest environment in this case suffers from a big I/O performance degradation, which can be a noticeable problem when dealing with very large corpus data.

While there is significant research in the use of different kinds of virtualization in the context of Grid technologies, this is not a wide spread feature at this time. While it is possible to use clusters with full support for chroot environments, for quick adoption and widespread use of Grid computing in NLP, porting of tools to the most often supported environment, i.e. SLC, will be necessary.

Legal Issues

The actual deployment of Grid computing in the natural language processing area (especially relevant for corpus linguistics) faces specific legal issues – the data being processed are in majority of 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 (fair use in the U.S.A. jurisdiction, citation and educational use in many of the EU countries' copyright laws). The situation is somewhat similar to the problems the users of Grid computing in health care systems – though in that case, metadata are the most sensitive and protected part of the data-set, while in corpus linguistics the data (i.e. texts in the corpora) are sensitive, but the metadata is usually freely accessible (Santos, Koblitz 2008).

In any case, the research institution using the data for research most likely does not have the right to distribute the data at all. If the contractual obligations prevent the institution from physically copying the data beyond the premises of the institution, it might be still advantageous to use the Grid infrastructure for computing clusters of the institution itself, and use middleware functions to restrict data-replication to those processing nodes and data storage elements physically located in the organization. This way, the whole Grid can still be used for less sensitive tasks, or for post-

processing the results of operations on sensitive data (when the post-processing does not include access to sensitive data), while at the same time the computing nodes will be available as part of the whole Grid computing pool when they would be left idle otherwise.

While the actual uploading of the data to Grid-enabled storage is not to be considered a form of "distribution" as long as no other person or organization is allowed to get the data, it is nevertheless desirable to protect the data from casual snooping. For one thing, an administrator of the Grid node where the data physically reside can get access rather trivially; and while he or she is legally obliged not to misuse his access (usually by rather strict agreements, in the case of European Grid infrastructure), a measure of additional protection seems to be necessary – to avoid data leaking in case the computer hosting the Grid node is compromised, unbeknown to the administrators.

Computing grids had to be very security-conscious from the very beginning, since the very premise of a Grid network is, from the point of view of the site administrator, to give external users access to the local computing infrastructure and, from the point of view of Grid users, to entrust data and applications to untrusted, foreign sites.

Moreover, the basic requirement for a viable, scalable and sustainable security infrastructure in the context of large Grid networks has to be a robust solution with as few single points of failure as possible to avoid failures of security services that could effect negatively the availability of the whole infrastructure (Laccetti, Schmid 2007).

Grid security has several components:

- Authentication, a method of confirming the identity of the user or organization behind an operation, is implemented on the basis of the Public Key Infrastructure (PKI) and standard x509 digital certificates (with a number of extensions to facilitate the use of PKI in the context of Grids).
- Authorization is provided in the framework of virtual organizations (VOs), a mechanism enabling Grid users all over the world to organize themselves according to research topics and computing requirements, regardless of geographic constraints, and permitting sites to regulate the use of their resources according to user, discipline, software requirements etc.
- Monitoring and ticketing permits users and administrators to keep track of infrastructure availability and to react to technical and security matters in a timely fashion.
- Accounting reports on the use of the infrastructure and enables the community to regulate and enforce the use of the infrastructure.

Public Key Infrastructure

Public Key Infrastructure, first introduced to the general public in the context of securing the web and enabling on-line shopping and banking, has become the standard authentication model in many application domains. Defined by a number of Internet Drafts, RFCs and standards, PKI is a widely deployed and evolving system (http://www.ietf.org/dyn/wg/charter/pkix-charter.html).

PKI is based on the property of asymmetric ciphers, where a different key is used for encryption and decryption. This property allows the encryption key to be always kept private and secret and the decryption key to be public, usually published with some information about the owner of secret key in the form of a x509 digital certificate.

In PKI, such a digital certificate is used as the token of identification: it is issued by a certification agency (CA) on the basis of an identification process (i.e. checking legally acceptable personal ID documents in person). But the certificate is coupled with a secret key that has been generated by the user requesting the certificate and is never exposed to the CA. To issue a certificate, the CA now sets up information about the entity (user, host or service) to be certified in accordance to the identification data provided in a standard form called a Distinguished Name (DN, following a LDAP-like name scheme: CN = Joe User, OU = My Department, DO = Institute of Dispersive Linguistics, DC = San Marino, and signs it with their own secret key from the CA certificate.

This scheme ensures that nobody, not even the CA, can use the certificate (since only the owner of the certificate possesses the secret key) and protects the information in the certificate with the signature, produced with the CA's own secret key.

To make the system work, CA certificates with public keys are published in a well advertised manner (or shipped with software, such as. web browsers, Grid middleware packages and GNU/Linux distributions). Recipient of a document or a connection that uses a client certificate and is encrypted or signed with such a certificate can therefore verify that the document or connection really was encrypted or signed by the said certificate by decrypting it with the public key included in the certificate, and it can verify the information in the certificate by checking the certificate with the CA public key in the same manner.

A number of additional security measures are used in the Grid: CA secret keys are kept in off-line systems or in dedicated certified hardware modules (hardware security modules or HSM) while end-entity certificates are re-issued with new keys yearly or kept in hardware security tokens. In addition, actual user certificates are never entrusted to non-trusted entities: for almost all operations in the Grid, short-lived proxy certificates are used instead.

Virtual organizations

While PKI provides authentication, a different system is needed to provide authorization, i.e. to help decide if a given user, host or service is to be allowed to carry out a specific task: use a specific resource or access specific data. In the context of Grid computing infrastructure, this role is implemented in the framework of virtual organizations (VOs).

A Virtual Organization serves two purposes:

- As an organizational form, a VO permits a number of researches from different organizations, usually geographically dispersed, to collaborate and share tools, data and resources.
- In the Grid security infrastructure, a VO provides means of regulating access to resources, i.e., a VO provides authorization after authentication is provided by PKI.

With this combination of roles, Virtual Organizations have proven themselves to be most efficient in enabling a higher level of international collaboration and have permitted the European Grid network to foster new, faster development in many disciplines by providing an unprecedented framework for international collaboration.

In practice, members of a research project or a discipline can set up a VO and decide on its modes of operations and access to resources quite independently. They have to decide what kind of tools the VO members will be using in the Grid, define the data formats, prepare data repositories, develop execution environments with the tools installed and set up a Virtual Organization Membership Service server (VOMS server) to store authorization credentials.

Then some resources have to be made available to the community of VO members. In practice, that means obtaining support of a number of Grid sites (organizations owning computing clusters partaking in the Grid) that have to configure their Grid middleware installations to include the new VOMS server in its authorization procedures and to either install the execution environment (or, more realistically, environments) for the VO or give access to some members of the VO so that they can perform the installation and maintenance if the execution environment on the site themselves. Additionally, a number of Grid storage elements (SE) has to be configured to allow the VO members to access and store the data on their disk space.

Proxy certificates

With the VO and VO supporting Grid sites, a VO member can submit Grid jobs and access VO-owned data using his certificate. This is implemented in an indirect

manner by means of Grid proxy certificates, as mentioned previously in the discussion of PKI infrastructure.

Grid proxy certificates are primarily used to permit a job to authenticate in the name of the user spawning the job, without the requirement of direct user interactions during the course of the job. This means that the proxy certificate must have the same DN as the users' certificate, but it has a different secret key which is not protected with a pass-phrase that would require user interaction on the keyboard. Proxy certificates are generated with a tool that uses the users' certificate to sign the proxy (as if it were a CA), thus confirming that the proxy was indeed generated by the user. In addition, gird proxy certificates are protected with file permissions and are always short-lived (from several hours to a few weeks) to mitigate the risk of the unprotected secret key.

To interact with the VO authorization system, the user generates a VOMS Grid proxy certificate that obtains special certificate extensions from the VOMS server and incorporates them in the proxy certificate. These extensions encode VO group and role attributes of the user and are themselves signed by the VOMS server with its service certificate, using the PKI infrastructure's authentication facilities to implement an authorization layer.

In this manner, a job can obtain authorization to use computing resources and data simply by providing a suitable VOMS proxy certificate. Its attributes are recognized by the Grid manager servers that provide it with to data storage (storage resource managers, SRM) and other resources.

As and additional level of security, Grid managers assign each job a temporarily unique user ID in the underlying operating system mapped from its active VO role in such a way that no jobs with different roles (and therefore potentially different access permissions) can share access on the underlying implementation.

In this way the system implements fine-grained control over the use of Grid resources and data without any reliance on the availability of authentication and authorization servers, thus avoiding a single point of failure that would have a significant impact on the scalability of the system.

Data Protection

Using these security components, additional measures of data protection can be implemented when necessary (Garabík et al. 2009). In the context of NLP, such a measure is of critical importance, since most of the data-sets in corpus linguistics contain copyrighted texts that need to be protected.

To solve this problem, the corpus data has to be suitably protected where it is permanently stored. Therefore the data should be stored in encrypted form in a

dedicated storage element and the access authorization should be set up in such a way that access is restricted to VO users who belong in a VO group of users who signed the necessary legal agreements to access the data. Furthermore, the data should be transferred to the untrusted environment of Grid worker nodes, where jobs perform their computations, in the encrypted form and that the decryption keys are issued to the jobs protected with asymmetric encryption decryptable only by the job's Grid proxy keys so that only the jobs can access the keys and decrypt the data.

In this manner, access and decryption is regulated with the authorization of embedded VOMS attributes in the proxy certificate without any additional authorization steps, while the data is never shipped or stored in unencrypted form.

If the tools used by the job have to store temporary files on disk, these are protected from other processes (with the exception of system administrators, who are already bound by strong agreements pertaining to data security on the Grid) and are in addition of short-lived nature.

There exist different implementations of the system described. The simplest form involves the use of a decryption filter in the job script and is rather simple to deploy. A more flexible solution, based on CryptoSRM (cryptographic storage resource manager) and Hydra Key Storage (a distributed fragmented encryption key storage system) is described in (Santos, Koblitz 2008).

From Grid to Web Services

Currently, the efforts have been concentrated on the minutae of job and task management and grid resource allocation. While such an approach could be acceptable for researchers that want to develop new tools, researchers that want to merely use them will require more flexible and easy to use interfaces, usually in the form of web services. As ToTaLe already has a web interface (http://nl2.ijs.si/analyze/), including a facility allowing a user to upload a small corpus as a compressed archive), it has been relatively easy to adapt the web application to use the grid backend to perform the annotation and to enable the service to process much larger data-sets in a reasonable time. Similarly the task for the term extractor was straightforward. Providing a web interface for a generic n-gram processing service seems less likely at this time, since the work to perform depends heavily on a number of factors, such as the structure of the corpus, the kind of n-gram analysis required etc. For such task, it is possible to add some web-based interfaces to grid resources, possibly structured around the meta-data catalogue. This interface should enable a user to quickly set up a number of generally useful but computationally expensive tasks, where the system should take care of factors such as the management of individual jobs, necessary conversions of corpus data and allocation of suitable grid storage for end results.

4.4. Case studies

Installation and usage

The case studies described further have been carried on the Squeeze (testing) Debian distribution, which is a “moving target”' distribution, meant for users that want newer version of the distribution and included packages, but do not want to deal with (potentially) broken bleeding edge packages from the unstable Debian repositories. To summarize, a package will get into testing if it has no release-critical bugs, has spent several days in the unstable repository and its inclusion in testing will not break other packages. Testing distribution has been used deliberately, because it is advantageous to use new versions of the required packages which will not become obsolete in near future, even if the packages in testing repositories will be rather quickly replaced by still newer versions (Javoršek, Erjavec 2009).

Debian has a standard method for installing the base system into a chroot environment, implemented by a tool, called debootstrap. Installation of a particular disctribution using debbootstrap is straightforward, after the dicstribution is installed, desired software packages can be installed inside the chroot in their usual way.

Morphosyntactic Annotation (Tagging) with ToTaLe

Automated annotation is a time consuming and computing intensive task, so it has been considered for the experiment. The tagging has been based on ToTaLe, an automated multilingual annotator (Erjavec et al. 2005). Since ToTale has recently had a new tag-set added for Slovenian, an experimental re-tagging of the fidaPLUS corpus of modern Slovenian (621 million words), seemed a natural task to do on the grid. fidaPLUS is stored in the form of 44 000 files encoded in the Text Encoding Initiative format and contains full morpho-syntactic annotation (lemma, MSD tag) and marks for punctuation and sentence boundaries. To perform the annotation, a new execution environment has been created on the experimental setup for the future HLT VO, and ToTaLe with its dependencies and language models has been installed. In splitting up the task of annotation into a suitable number of jobs, the maximum amount of available computing cores is targeted, and for that reason job description files containing approximately 70 files (with minor differences due to differences of file sizes) have been used, which gives 630 jobs. The actual job consisted of the job description file (specifying the input and output data files, execution environment, hardware requirements, start-up script etc.), a small control script and filter that extracted the plain texts from the compressed annotated corpus files in TEI XML form and passed them to ToTaLe in sequence, compressing the

results on the fly. The actual run has shown the mean time of execution per job to be around 10 hours, 2 hours of which have been spent queuing (waiting for computing resources) and in file upload or download. The task has been completed in under 12 hours, while consuming on the order of 6500 hours of computing time and processing and regenerating over 70 GB of corpus data—automatically annotating a 621-million words corpus in less than a day. Practical applications of this service, particularly having in mind that ToTaLe supports several MULTEX-East languages and tag-sets and will, hopefully, some day support all of them, are obvious to most linguistic users.

Morphosyntactic annotation with morče

Morphosyntactic tagging of the Slovak National Corpus consists of two steps. The first performs morphosyntactic analysis, where each word in the input texts is assigned a set of possible morphosyntactic tags. This step essentially consists of 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. The software is implemented in the Python programming language and is actually quite fast, since the core of the task consists simply of a look-up in the possibilities in the tables, and most of the CPU work is spent on I/O operations, parsing the input file and assembling the output. On a reasonably recent hardware (Intel Xeon 2.33 GHz CPU) it is able to process over 10 000 words per second. It can also parallelize easily, since the words can be analyzed independently of each other.

The second step is disambiguation, where each word is assigned a unique lemma and a morphosyntactic tag out of the possibilities assigned in the first step. For disambiguation, morče (Spoustová et al. 2009), an averaged perceptron model (originally used for the Czech language tagging) has been used, re-trained on the Slovak manually annotated corpus. Disambiguation is much slower that the morphology analysis, its average speed reaches only about 300 words per second. Parallelization at the application level is also 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 morče instantiations in parallel.

Given the speed differences between morphology analysis and disambiguation, the morphology analysis execution time can be considered negligible and it is possible to design the whole tagging to be done in one step, without the need to parallelize the morphology analysis process while the disambiguation is to be run in parallel.

n-gram Processing

Another task was an example of n-gram statistics, namely frequencies for 1-grams and 2-grams for the whole Slovene fidaPLUS corpus separately for words, lemmas and MSDs, totaling a corpus of 1863 million words. Due to n-gram counting being a much simpler task compared to automatic annotation, it is possible to ship the counting program and control script directly with the jobs (no installation in the execution environment necessary; Ted Pedersen's n-gram statistics package for Perl has been used) and could also process more files (500) per job. This resulted in 90 submitted jobs which finished in under 4 hours and consumed under 80 hours of computing time. Again, the source files of the corpus had to be downloaded, uncompressed, processed so that relevant data was extracted from TEI XML form in a plain text file and then processed. Since these jobs have been much shorter, clearly more time (but not computing resources) was spent queuing or downloading and uploading data than in actual processing, although it has to be noted that this occurred only in some cases (where due to faults in network transfers, files had to be downloaded several times) and most jobs finished around the second hour mark. Similar experiment was based on a term extractor described in (Vintar 2004) and its web-based interface. The web interface takes a text file, performs the necessary conversions (text, PDF and different offce formats are accepted), uses the ToTaLe web service to lemmatize and annotate it and runs an n-gram statistical analysis on the lemmatized text. Using a combination of statistical scores based on lexical statistics and linguistic extraction (based on MSD patterns), a list of possible candidates for terminologically relevant terms in the text is generated.

TectoMT

TectoMT is a software framework aimed at machine translation at the tectogrammatical level of analysis (Žabokrtský et al. 2008). The system is modular – the framework itself consists of many independent modules (blocks in TectoMT terminology), each implementing one specific, independent NLP-related task. Each of the blocks is a Perl module that interacts with the system using a single, uniform interface. However, sometimes the module serves only as a wrapper for the underlying implementation in another programming language. The tectogrammatical annotation and consequently the TectoMT framework primarily stores linguistic data in its own format, called TMT. TMT is an XML-based format, designed as a schema of the Prague Markup Language (PML)[11]. Nevertheless, its blocks are by no means obliged to use this format (Pajas, Štěpánek 2006).

TectoMT has been developed with modern Linux systems in mind, and as such its installation requirements are easily met by any contemporary Linux distribution. It

[11] Not to be confused with the Physical Markup Language

should be noted that TectoMT, being written mostly in Perl, depends on many external Perl modules and its installation scripts are intelligent enough to automatically download and install any missing dependencies; this, however, circumvents standard distribution packaging systems, therefore it is better to install all the necessary packages with the packaging system tools before attempting to install TectoMT. There are also some C language modules that are not compiled by default, but have to be compiled separately inside the TectoMT installation source tree.

TectoMT also has some built-in capabilities for parallelization of its tasks, using the Sun Grid Engine – it is possible to adapt the Sun Grid Engine batch software to various Grid middlewares (Borges et al. 2007), but TectoMT can be run on the Grid system directly without relying on its internal parallelization possibilities, if the user takes care of splitting the input data into appropriate chunks for parallel processing.

4.5 Recommendations

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 (Erjavec, Javoršek 2008). 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. Here we describe a number of steps that are should be taken to provide this service to the community.

Creation of Core Services

To support the HLT VO, a Virtual Organization Membership Service (VOMS) server to provide VO user and service access control has been set up. 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 VO, 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. At the time of this writing, HLT VO VOMS is supported by the SiGNET cluster and it is included as a supported service in the Slovenian National Grid Initiative project. 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.

Registration of the VO

While the HLT VO could be registered as a supported VO in the European grid infrastructure (i.e. with the EGEE and NorduGrid projects), it has not been yet done so as at the time of this writing, no organizations from other nations support the VO and so it lacks international membership.

As soon as HLT VO is registered, it will be discoverable using the central services of both above mentioned infrastructures. It is also expected to become one of the supported VOs in the future European Grid Initiative (which starts its operations in 2010).

After the VO is registered, as members 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.

Data and Metadata

Due to many restrictions that are often applied to the use of corpus data according to contracts regulating the use of copyrighted and other non-free materials, it is essential to provide a managed distributed data access with a central metadata server and full support for VO-based access control and authorization. While no such a solution has been implemented, it is an essential element to allow international collaboration. A number of existing solutions for grid infrastructure has been tested and we recommend a metadata service on the base of AMGA, the Arda Metadata Catalogue Project 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.

VO Execution environments

For testing purposes, a set of command-line tools for typical linguistic grid jobs have been developed and execution environments with all the necessary software packages pre-installed are prepared. These tools already provide a way to perform resource-intensive tasks using distributed corpus data and distributed computing resources in the HLT VO. This tool set should be expanded and developed into a viable basis for the future use in the new VO and into more advanced tools. A set of web services and web grid interfaces should be built, to enable linguists to use the new tool-set with ease. The final form of the HTL VO execution environment is not yet decided as it will be shaped according to the needs and requirements of future member organizations.

Web interfaces and central services

A dedicated web site for information, documentation and user management of HTL VO is being set up at JSI as part of Slovenian National Grid Initiative effort. It will provide the central grid services for the VO, 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 a 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.

5. Concluding Remarks

In conclusion we want to discuss in brief the question ***"What are the impacts of research infrastructures supporting Slavic languages resources"***? The impacts of research infrastructures relate to the impacts of the research and innovation that they facilitate. These can be classified as:

- *Direct scientific impacts*, relating to the new knowledge creation (scientific outputs) and the theoretical advancement of science achieved via the research they facilitate, training and capacity building;
- *Technological impacts,* relating to the innovations in the production of data and services that arise as effects from the development of research infrastructures;
- *Social impacts* - the contribution to general welfare arising from progress made in science, which stems from the research process and its contribution to improving the quality of language communication of EU citizens.

From a scientific perspective, social and technological impacts may seem irrelevant – the value of a research infrastructure to the process of scientific discovery may be regarded as the single most important aspect of its potential impact.

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 (Dimitrova et al. 2010a). The project also emphasized the important role of scientific collaboration in the development of digital language resources, online accessibility and digital preservation of Europe's cultural heritage and collective memory. The project organised a series of five open MONDILEX workshops for discussing a conceptual scheme of networking of centres for high-quality research in Slavic lexicography and their language digital resources. The Proceedings of these events (Iomdin, Dimitrova (Editors 2008), Shyrokov, Dimitrova (Editors 2009), Garabík (Editor 2009), Koseska, Dimitrova, Roszko (Editors 2009), and Erjavec (Editor 2009)) 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.

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 this area – especially where it crosses disciplinary borders – is still not sufficient. 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.

Open access to research infrastructure via Web

Given the exponential growth of information, managing and providing an efficient access to data represent a major challenge. Tools and processes for managing data, promoting interoperability, integrating databases and ensuring access rights require significant development effort in order to provide sustainable services. Pan-European collaboration in this area – especially where it crosses EU borders – is still not sufficient. Some issues arise in this respect. The virtual environments will provide facilities for e-Research via open access and exploration of language resources and tools necessary for the creation of dictionaries such as corpora (including parallel and comparable), concordances, word sketches, morphosyntactic taggers, parsers, semantic annotation. It will ensure user-oriented access to digitalized monolingual, bilingual and multilingual dictionaries of Slavic languages for research, educational, and cultural purposes. A crucial element here is a clear policy of access. For applied research and innovation access conditions should be clearly defined. Management of appropriate usage needs to include the development of clear access control policies, and, wherever possible, promote wider collaboration between different groups of users. Access to specific databases and repositories for research and development purposes and innovative aims should be considered attentively. The policies for access to the research infrastructure could be regulated by dedicated public documents where issues concerning data protection, software development and other similar topics are indicated. Common regulations should define various type of access for regular partners, associates, third parties as well as casual users. The specific provisions for all types of partners, external users, as well as differentiations of services would be a subject of additional agreement.

Preservation of Web content

The Web-content is harvested and deposited somewhere, either in the country of production or abroad, so the order for permission of the use of such deposited material should be regulated in accordance with the copyright. The rights holders stress that digitization and on-line accessibility need to be achieved in full respect of the current copyright rules. The general rule-of-thumb is that works in the public domain should remain in the public domain also in the digital environment. Public domain content in the analogue world should remain in the public domain in the digital environment. In particular, one can recommend that public domain material that has been digitized with public money by public institutions be not locked up, and it should continue to play its essential role as a source for creativity and innovation.

Acknowledgements

Firstly, we would like to acknowledge the 7th FP of EC, since the financial support by the Commission under the Grant agreement 211938 ensured the successful fulfilment of this project. We would like to thank to Maria Theofilatou (European Commission, Research Infrastructures), MONDILEX project officer.

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

Our special thanks to all of the colleagues from the six MONDILEX participants' teams that contributed to the project. Six organisations participated in the MONDILEX project: Institute of Mathematics and Informatics, Bulgarian Academy of Sciences (IMI-BAS), Sofia, Bulgaria, which coordinated the project; Institute of Slavic Studies, Polish Academy of Sciences (ISS-PAS), Warsaw, Poland; Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences (ĽŠIL-SAS), Bratislava, Slovakia; Jožef Stefan Institute (JSI), Ljubljana, Slovenia; Institute for Information Transmission Problems, Russian Academy of Sciences (IITP-RAS); and the Ukrainian Lingua-Information Fund of the National Academy of Sciences of Ukraine (ULIF-NASU).

The partners are research institutions from European countries whose national languages belong to the Slavic group: four EU member states Bulgaria, Poland, Slovakia, Slovenia, as well as two international cooperation partner countries – Russia and Ukraine. All partners are national centres for research in natural language processing, computational and traditional linguistics and lexicography.

Short profiles of the MONDILEX participants appear below.

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences

The Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences (IMI-BAS) is a leading Bulgarian centre for scientific research and applications in mathematics, informatics and information technologies.

The Department of Mathematical Linguistics at the IMI-BAS, founded in 1977, (www.math.bas.bg/ml/) pursues research in theoretical, computational and mathematical linguistics, natural language processing, human-language technologies, and knowledge technologies. In the latest 15 years the staff of the Department has developed TEI-compliant digital language resources, among them: morpho-syntactic specifications for Bulgarian (for encoding and annotating digital corpora and

lexica), MULTEXT-East Bulgarian-English parallel and aligned corpora, MULTEXT-East Bulgarian annotated comparable corpus and lexica, lexical databases (LDBs) for integrated multilingual resources – CONCEDE LDB, LDB supporting a Bulgarian-Polish online dictionary, Bulgarian-Polish parallel and comparable corpora and bilingual digital dictionaries – a Bulgarian-Polish electronic dictionary and an experimental Bulgarian-Polish online dictionary.

The first Bulgarian-Polish bilingual digital resources are being developed in the framework of a bilateral collaboration between IMI-BAS and ISS-PAS. The Bulgarian-Polish parallel corpus contains more than 3 million words, mostly from works of fiction. Some of the parallel texts are aligned at the paragraph and sentence level. An experimental version of the Bulgarian-Polish electronic dictionary consists of approximately 20 000 dictionary entries. Trilingual Bulgarian-Polish-Lithuanian parallel (1 million words, mainly literary work) and aligned corpora are also in preparation. For the first time, a small Slovak-Bulgarian parallel corpus (approx. 1.2 million words) and sentence-aligned corpus (approx. 177 000 words) are currently being developed in the framework of the joint collaborative project between IMI-BAS and ĽŠIL-SAS. A small parallel corpus with texts in Bulgarian, Polish, Slovak, Slovene, and English as a hub language, of official documents of the European Commission available through the Internet is also currently collected.

Institute of Slavic Studies, Polish Academy of Sciences

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

Starting from 2004, the department extended its activities to the field of corpus linguistics, NLP, bilingual electronic dictionaries, and started the projects on design and development of Polish-Ukrainian digital resources (in cooperation with ULIF-NASU), Bulgarian-Polish digital resources (in cooperation with IMI-BAS). The Bulgarian and Polish teams are developing (currently for research purposes) the first Bulgaria-Polish-Lithuanian experimental parallel corpus. The parallel corpus contains over one million words.

Ľudovit Štúr Institute of Linguistics, Slovak Academy of Sciences

Ľ. Štúr Institute of Linguistics is the central linguistic institution in the Slovak Republic. In the lexicography field, ĽŠIL is active in compiling traditional dictionaries. There are also other dictionary projects currently carried on, for example tradi-

tional Czech-Slovak dictionary and a wiki-based Slovak-Czech dictionary, produced in collaboration with the Czech Language Institute of the Czech Academy of Sciences.

Slovak National Corpus is a representative corpus of contemporary written texts, containing about 780 million words with automatic lemmatisation and morphological tagging. A smaller, balanced subcorpus consists of one third of journalistic texts, one third of specialised texts and one third of fiction, amounting to 200 million words. Another subcorpus contains manually lemmatised and annotated texts of about 1.2 million words. A manually syntactically annotated corpus contains about 50 000 sentences, and a corpus of spoken Slovak contains about 1 200 000 words. The Russian-Slovak, French-Slovak, Bulgarian-Slovak and Czech-Slovak sentence-aligned parallel corpora are intended for linguistic research, teaching, translation, cross-linguistic studies and applications in natural language processing, primarily for machine translation, as well as dictionary compilation. ĽŠIL designed and implemented a multilingual terminology database of corpus linguistics terms, with the goal of describing terminology of all the MONDILEX languages. The database has been tested with several Slavic language entries.

Jožef Stefan Institute, Ljubljana, Slovenia

The Department of Knowledge Technologies at the Jožef Stefan Institute (http://kt.ijs.si/), is the major Slovenian AI research group with 25 years tradition in R&D in artificial intelligence, intelligent systems, information systems, machine learning, and natural language processing. The Department has long-standing experience in the development of language resources, including research in automatic annotation techniques and encoding standardisation. The department has coordinated the FP5 R&D project SolEuNet, was involved in thirteen FP6 projects; those partially or wholly dealing with human language technologies include IP SEKT "Semantically Enabled Knowledge Technologies", NoE PASCAL "Pattern Analysis, Statistical Modelling and Computational Learning" (with JSI a core partner in both), STREP ALVIS (Superpeer Semantic Search Engine), and the project SMART, "Statistical Multilingual Analysis for Retrieval and Translation". The department was involved in the recently completed FIDA+ corpus, the continuation of the first reference corpus of Slovene language, FIDA. FIDA+ contains 600 million words of contemporary Slovene language, with the corpus composition carefully selected to be balanced and representative. Other monolingual corpora include the DSI corpus (1 million words, conference papers in informatics) and, as a test bed for syntactic annotation, the Slovene Dependency Treebank. The multilingual corpora are mostly parallel English-Slovene ones, with a total volume of 13 million words (EU legal text, technical writing, medical abstracts, mixed genres). The department is also involved in producing the 20-way parallel corpus JRC-ACQUIS, a freely available aligned corpus of EU legal texts, developed at the

European joint Research Centre, in Ispra, sloWNet, the Slovene WordNet, and other lexical resources, such as the Japanese-Slovene learner's dictionary.

In MONDILEX, the JSI partner has concentrated on two connected issues, the establishment of a Grid infrastructure, primarily for lexicography oriented corpus processing, and standards of encoding digital resources, with a focus on describing the morphosyntactic properties of words in lexica and annotated corpora.

Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences

The Laboratory of Computational Linguistics of the Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, has developed a multipurpose linguistic processor, ETAP-3, which includes, among other things, a machine translation system operating between Russian and English, with small prototypes for other language pairs (French-Russian, Russian-German, Russian-Korean, Russian-Spanish, and Arabic-English); a system of synonymous and quasi-synonymous paraphrasing of natural language utterances (in English and Russian), a module that enables computer-assisted translation of texts from UNL (Universal Networking Language, a semantic interlingua specially designed to facilitate multilingual communication in Internet) to natural languages and vice versa.

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

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

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

The ULIF-NASU is a repository of the National Dictionary Base of Ukraine. The institution's efforts are focused on computer technologies for creating the monolingual, bilingual and multilingual dictionaries and natural language processing systems. ULIF publishes series of academic dictionaries "Dictionaries of Ukraine", which now numbers more than 70 volumes. As a member of TEI, ULIF develops

national standards for electronic text processing. The Fund also develops the Ukrainian National Linguistic Corpus.

To create a unified language for dictionary structure description, a theory of lexicographic systems (L-systems) was developed. The theory, which combines the features of several formal structures for data description (data models, logical-linguistic calculi), was used to create an integrated L-system that tackles the phenomena of inflection, orthoepy, synonymy, antonymy, and phraseology of the Ukrainian language. An electronic dictionary based on this system can be accessed at the Ukrainian Linguistic Portal (http://ulif.org.ua).

A considerable part of ULIF's activity is devoted virtual systems of professional interaction in linguistics that enable the development of common lexicographic projects by researchers from different organizations or countries.

BIBLIOGRAPHY

Andreychin, L. et al. 1994: Bulgarian Explanatory Dictionary. /Dictionary of the Bulgarian Language. 4th revised edition, prepared by Dimitar G. Popov/ Nauka i Izkuvstvo Publishing House, Sofia, 1994 (in Bulgarian).

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

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. 1378-1381.

Boguslavsky et al. 2000: Boguslavsky, I., S. Grigorieva, N. Grigoriev, L. Kreidlin, and N. Frid. Dependency treebank for Russian: Concept, tools, types of information. In: Proceedings of COLING'2000, 987–991.

Boguslavsky et al. 2002: Boguslavsky, I., I. Chardin, S. Grigorieva, N. Grigoriev, L. Iomdin, L. Kreidlin, and N. Frid. Development of a dependency treebank for Russian and its possible applications in NLP. In: Proceedings of LREC'2002, 852-856.

Boguslavsky et al. 2008: Igor Boguslavsky, Leonid Iomdin, Denis Valeev, Victor Sizov. Синтаксический анализатор системы ЭТАП и его оценка с помощью глубоко размеченного корпуса русских текстов. In: Труды Международной конференции «Корпусная лингвистика – 2008». СПб.: Санкт-Петербургский государственный университет, 2008. 56-74. ISBN 978-5-288-04769-5. (In Russian)

Boguslavsky, Dikonov 2008: Boguslavsky I., Dikonov V. Universal Dictionary of Concepts. In: Iomdin, Dimitrova (Eds.), Lexicographic Tools and Techniques. Proceedings of the MONDILEX First Open Workshop, 3-4 October 2008, Moscow, Russia. 31-41. ISBN 978-5-9900813-6-9.

Borges et al. 2007: Borges, G., David, M., Gomes, J., Fernandez, C., Lopez Cacheiro, J., Rey Mayo, P., Simon Garcia, A., Kant, D., and Sephton, K. Sun Grid Engine, a new scheduler for EGEE middleware. In: IBERGRID – Iberian Grid Infrastructure Conference, 2007.

Brants, T. 2000: TnT — A Statistical Part-of-Speech Tagger. In Proceedings of the 6th Applied Natural Language Processing Conference ANLP-2000, (224-231). Seattle, WA.

Bueti et al 2004: Bueti G., Congiusta A., Talia D. Developing Distributed Data Mining Applications in the KNOWLEDGE GRID Framework. In: High Performance Computing for Computational Science - VECPAR'04, Valencia, Spain, LNCS, vol. 3402, 156-169, Springer-Verlag, 2004. ISBN 978-3-540-25424-9.

Buryachok 2002: Бурячок А.А. Орфографічний словник української мови: 4-те вид. Kiev, Наук. думка, 2002. 464 pages. (In Ukrainian)

Buryachok A. A. (Ed. 1999): Dictionary of Synonyms of the Ukrainian language. У 2 т. / Под ред. А.А. Бурячка; НАН України; Інститут мовознавства ім. О.О. Потебні. – Kiev, Наук. думка, 1999. v. 1: А - Н. 1040 p.; v. 2: О - Я. 960 pages. (In Ukrainian)

Calzolari, N. 1996: In: Monachini, M. (Ed. 1996), Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora: A common proposal and applications to European languages. EAGLES Report EAG—CLWG—MORPHSYN/R. Pisa: ILC.

Cannataro 2003: Cannataro M., Talia D. The Knowledge Grid. Communications of the ACM, vol. 46, n. 1, 89-93.

Cannataro 2004a: Cannataro M. and Talia D. Semantics and Knowledge Grids: Building the Next-Generation Grid. IEEE intelligent systems & their applications, vol. 19, n. 1, 56-63, January 2004.

Cannataro et al. 2004b: Cannataro M., Congiusta A., Mastroianni C., Pugliese A., Talia D., Trunfio P. Grid-Based data mining and knowledge discovery. In: N. Zhong, J. Liu (Eds.), Intelligent Technologies for Information Analysis, Springer, chapt. 2, 19-45, 2004. ISBN 3-540-40677-8.

Chirst, O. 1994: A modular and flexible architecture for an integrated corpus query system. COMPLEX'94, Budapest.

Congiusta et al. 2003: Congiusta A., Pugliese A., Talia D., Trunfio P. Designing Grid Services for Distributed Knowledge Discovery. In: Web Intelligence and Agent Systems, vol. 1, n. 2, 91 – 104, IOS Press, 2003.

Congiusta et al. 2006: Congiusta A., Talia D., Trunfio P. Service-Oriented Knowledge Discovery in Grids. In: Proceedings of the Workshop on Grid Technologies for Knowledge-Based Industries and Businesses, co-located with IST 2006, Helsinki, Finland, November 2006.

Congiusta et al. 2007a: Congiusta A., Talia D., Trunfio P. Using Grids for Distributed Knowledge Discovery. In: G. Felici, C. Vercellis (Eds.), Mathematical Methods for Knowledge Discovery and Data Mining. IGI Global, Hershey, USA, chapt. XVII, 284 – 298, 2007. ISBN 978-1-59904-528-3.

Congiusta et al. 2007b: Congiusta A., Talia D., Trunfio P. WSRF-Based Services for Distributed Data Mining. In: D. Talia, A. Bilas, M. Dikaiakos (Eds.), Knowledge and Data Management in Grids. Springer. USA, 203-220, 2007. ISBN 0-387-37830-8.

Demchenko et al., 2006: Demchenko, Y., Gommans, L., Tokmakoff, A., van Buuren, R. Policy Based Access Control in Dynamic Grid-based Collaborative Environment. In: Proceedings of the International Symposium on Collaborative Technologies and Systems, CTS 2006, 14-17 May 2006. 64-73.

Dimitrova, Ludmila 2008: Bulgarian Digital Resources as a Base for Automatic Disambiguation. In: Études Cognitives. Vol. 8. SOW, Warsaw, 2008. 255-271. ISSN 1641-9758.

Dimitrova, Ludmila 2009: From Electronic Corpora to Online Dictionaries (on the example of Bulgarian Language Resources). In: Levická, Garabík (Eds.), Proceedings of the Fifth Intrnational Conference NLP, Corpus Linguistics, Corpus Based Grammar Research, Smolenice, Slovakia, 25-27 November 2009. 78-92. ISBN 978-80-7399-875-2.

Dimitrova, Ludmila 2010: Multilingual Digital Resources with Bulgarian language. In: Cognitive Studies/Études Cognitives. Vol. 10, SOW, Warsaw, 241-252. 2010. ISSN 2080-7147.

Dimitrova et al. 1998: Dimitrova, Ludmila, Tomaz Erjavec, Nancy Ide, Heiki-Jan Kaalep, Vladimír Petkevic, and Dan Tufis. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. Proceedings of COLING-ACL '98. Montréal, Québec, Canada, 315-319.

Dimitrova et al. 2002: Dimitrova, Ludmila, Radoslav Pavlov, and Kiril Simov. The Bulgarian Dictionary in Multilungual Data Bases. Cybernetics and Information Technologies. Vol. 2, num. 2, 12-15.

Dimitrova et al. 2005: L. Dimitrova, R. Pavlov, K. Simov, L. Synapova. Bulgarian MULTEXT-East Corpus – Structure and Content. Cybernetics and Information Technologies. Vol. 5, num. 1, 67-73.

Dimitrova et al. 2009a: Dimitrova, L., Koseska–Toszewa, V., Derzhanski, I., Roszko, R. Annotation of Parallel Corpora (on the Example of the Bulgarian–Polish Parallel Corpus). In: Shyrokov, Dimitrova (Eds.), Organization and Development of Digital Lexical Resources. Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009. Dovira Publishing House. Kiev, 2009, 47-54. ISBN 978-966-507-252-2.

Dimitrova et al. 2009b: Dimitrova, L., Panova, R., Dutsova, R. Lexical Database of the Experimental Bulgarian-Polish online Dictionary. In: Garabík (Ed.), Metalanguage and Encoding scheme Design for Digital Lexicography. Proceedings

of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009. 36-47. ISBN 978-5-9900813-6-9.

Dimitrova et al. 2009c: Dimitrova, L., V. Koseska, Satoła-Staśkowiak. Towards a Unification of the Classifiers in Dictionary Entry. In: Garabík (Ed.), Metalanguage and and Encoding Scheme Design for Digital Lexicography. Proceedings of the MONDILEX Third Open Workshop, 15-16 April 2009, Bratislava. 48-58. ISBN 978-80-7399-745-8.

Dimitrova et al. 2009d: Dimitrova Ludmila, Violetta Koseska, Ralitsa Dutsova, Rumyana Panova. Bulgarian-Polish online Dictionary – Design and Development. In: Koseska, Dimitrova, Roszko (Eds.), Representing Semantics in Digital Lexicography. SOW, Warsaw, 2009, 76-88. ISBN 978-83-89191-87-8

Dimitrova et al. 2009e: Dimitrova, L., Koseska, V., Roszko, D., Roszko, R. 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 RANPL'2009. Borovec, Bulgaria, 17 September 2009. 1-8.

Dimitrova et al. 2009f: Dimitrova, L., Koseska, V., Roszko, D., Roszko, R. Bulgarian-Polish-Lithuanian Corpus – Problems of Development and Annotation. In: Erjavec (Ed.), Research Infrastructure for Digital Lexicography. Ljubljana, 2009, 72-86. ISSN 1581-9973/ISBN 978-961-264-012-5.

Dimitrova et al. 2010a: Dimitrova, L., Koseska, V., Garabík, R., Erjavec, T., Iomdin, L., Shyrokov, V. MONDILEX – Towards the Research Infrastructure for Digital Resources in Slavic Lexicography. In: Cognitive Studies/Études Cognitives. Vol. 10, SOW, Warsaw, 147-162. 2010. ISSN 2080-7147.

Dimitrova et al. 2010b: Dimitrova, L., Koseska, V., Roszko, D., Roszko, R. Application of Multilingual Corpus in Contrastive Studies (on the example of the Bulgarian-Polish-Lithuanian Parallel Corpus. In: Cognitive Studies/Études Cognitives. Vol. 10, SOW, Warsaw, 217-240. 2010. ISSN 2080-7147.

Dimitrova, L., Garabík, R., Majchráková, D. 2009: Comparing Bulgarian and Slovak Multext-East morphology tagset. In: Organization and Development of Digital Lexical Resources. Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009. 38-46. Dovira Publishing House.ISBN 978-966-507-252-2.

Dimitrova, Koseska 2007: Dimitrova, L., V. Koseska–Toszewa. Digital Dictionaries – Problems and Features. In: Proceedings of the Jubilee International Conference Mathematical and Computational Linguistics, 6 July 2007. Sofia, Bulgaria. 25-34. ISBN 978-954-8986-28-1.

Dimitrova, Koseska 2008a: Dimitrova, L., V. Koseska–Toszewa. Some Problems in

Multilingual Digital Dictionaries. In: Études Cognitives. Vol. 8. SOW, Warsaw, 2008. 237-254. ISSN 1641-9758.

Dimitrova, Koseska 2008b: Dimitrova, Ludmila and Violetta Koseska-Toszewa. The Significance of Entry Classifiers in Digital Dictionaries. In: Iomdin, Dimitrova (Eds.), Lexicographic Tools and Techniques. Proceedings of the MONDILEX First Open Workshop, Moscow, 3–4 October 2008. 89-97. ISBN 978-5-9900813-6-9.

Dimitrova, Koseska 2009a: Dimitrova, L., V. Koseska. Classifiers and Digital Dictionaries. In: Cognitive Studies/Études Cognitives. Vol. 9, SOW, Warsaw, 2009. 117-131. ISSN 2080-7147.

Dimitrova, Koseska 2009b: Dimitrova, L., V. Koseska. Bulgarian-Polish Corpus. In: Cognitive Studies/Études Cognitives. Vol. 9, SOW, Warsaw, 2009. 133-141. ISSN 2080-7147.

Dimitrova, L., Pavlov, R. 2008: On Compatibility of Slavic Language Resources. In: Iomdin, Dimitrova (Eds.), Lexicographic Tools and Techniques. Proceedings of the MONDILEX First Open Workshop, Moscow, 3–4 October 2008. 15–22. ISBN 978-5-9900813-6-9.

Dimitrova, L., Rashkov, P. 2009. A New Version for Bulgarian MULTEXT-East Morphosyntactic Specifications for Some Verbal Forms. In: Shyrokov, Dimitrova (Eds.), Organization and Development of Digital Lexical Resources. Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009. 30-37. ISBN 978-966-507-252-2.

Divjak. 2008. Designing and evaluating a Russian tagset. In LREC'08, Paris. ELRA.

Ďurčo, Peter 2007: Zásady spracovania slovníka kolokácií slovenského jazyka. Online documentation: 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 (Eds.), Representing Semantics in Digital Lexicography. Proceedings of the MONDILEX Fourth Open International Workshop, Warsaw, Poland, 29 June–1 July 2009. SOW, Warsaw, 128-137. ISBN 978-83-89191-87-8

Ďurčo, Peter, Garabík, Radovan 2009: Slovak Paremiography Database. In: Erjavec (Ed.), Research Infrastructure for Digital Lexicography. Proceedings of the MONDILEX Fifth Open Workshop, 14-15 October 2009, Ljubljana. Informacijska družba Publ. House, Ljubljana. 20-26. ISSN 1581-9973/ISBN 978-961-264-012-5.

Erjavec, Tomaž 2004: MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In Fourth International Conference on Language Resources and Evaluation, LREC'04, 1535 – 1538, Paris. ELRA.

Erjavec, Tomaž 2006: MULTEX-East morphosyntactic specifications and XML.

In: SLAVCHEVA, M., SIMOV, K., ANGELOVA, G. (Eds). Readings in multilinguality: selected papers for young researchers. Sofia: Institute for Parallel Processing, Bulgarian Academy of Science, 2006, 41-48.

Erjavec, Tomaž 2007: An Architecture for Editing Complex Digital Documents. In Proc. of the 1st Intl. Conference "Digital information and heritage". Zagreb, 2007, pp. 105–114.

Erjavec, Tomaž 2009 (Editor): Research Infrastructure for Digital Lexicography. Proceedings of the MONDILEX Fifth Open Workshop, 14-15 October 2009, Ljubljana. 125 pages. Informacijska družba, ISSN 1581-9973/ISBN 978-961-264-012-5.

Erjavec, Tomaž 2009: MULTEXT-East Morphosyntactic Specifications: Towards Version 4. In: Garabík (Ed.), Metalanguage and Encoding Scheme Design for Digital Lexicography. Proceedings of the MONDILEX Third Open Workshop, 15-16 April 2009, Bratislava. 59–70. ISBN 978-80-7399-745-8.

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

Erjavec et al. 2000: Erjavec, Tomaž, Roger Evans, Nancy Ide, and Adam Kilgarriff. The Concede model for lexical databases. In Second International Conference on Language Resources and Evaluation, LREC'00, Athens, 2000. ELRA.

Erjavec et al. 2003: Erjavec, Tomaž, Roger Evans, Nancy Ide, and Adam Kilgarriff. From Machine Readable Dictionaries to Lexical Databases: the Concede Experience. In Proceedings of the 7th International Conference on Computational Lexicography, COMPLEX'03, Budapest, Hungary, 2003.

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

Erjavec, Džeroski 2004: Erjavec, T. and Džeroski, S. Machine Learning of Language Structure: Lemmatising Unknown Slovene Words. Applied Artificial Intelligence, 18(1):17–41. Taylor & Francis.

Erjavec, T., Javoršek, J. J. 2008: Grid Infrastructure Requirements for Supporting Research Activities in Digital Lexicography. In: Iomdin, Dimitrova (Eds.) Lexicographic Tools and Techniques. Proceedings of the MONDILEX First Open Workshop, Moscow, 3–4 October 2008. 5–14. ISBN 978-5-9900813-6-9.

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

Farrar, S., Langendoen, D. T. 2003: A linguistic ontology for the Semantic Web. GLOT International, 7(3):97-100. http://linguistics-ontology.org/

Fox G. C., Sun X. 2007: Special Issue: Progress of the Knowledge Grid, Concurrency and Computation: Practice and Experience, 19 (15), 2007.

Garabík, Radovan 2008: Storing morphology information in a wiki. In: Iomdin, Dimitrova (Eds.), Lexicographic Tools and Techniques. Proceedings of the MONDILEX First Open Workshop, Moscow, 3-4 October 2008. 55-59. ISBN 978-5-9900813-6-9.

Garabík, Radovan 2009: (Editor), Metalanguage and Encoding scheme Design for Digital Lexicography. Bratislava, 2009, 192 pages. ISBN 978-80-7399-745-8.

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

Garabík, Radovan, Špirudová, Jana 2009: Design of a New Slovak-Czech Lexical Database. In: Garabík (Ed.), Metalanguage and Encoding scheme Design for Digital Lexicography. Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009. 71–76. ISBN 978-80-7399-745-8.

Genči, J. 2009: Experience with Building Slovak Electronic Lexical Database. In: Garabík (Ed.), Metalanguage and Encoding scheme Design for Digital Lexicography. Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009. 77-82. ISBN 978-80-7399-745-8.

Hajič et al. 2006: Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, Magda Ševčíková-Razímová. Prague Dependency Treebank 2.0, Linguistic Data Consortium, Cat. No. LDC2006T01.

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

Ide N., C. M. Sperberg-McQueen 1995: The TEI: History, Goals, and Feature. In: Computers and the Humanities, 29, 5-15, 1995.

Ide, N., Véronis, J. 1994: Multext (multilingual tools and corpora). In: Proceedings of the 15th International Conference on Computational Linguistics COLING'1994, 90-96, Kyoto, Japan. ACL.

Iomdin, L., Dimitrova, L. (Editors 2008). Lexicographic Tools and Techniques. Proceedings of the MONDILEX First Open Workshop, Moscow, 3-4 October 2008. IITP Publ. House, Moscow, Russia, 109 pages. ISBN 978-5-9900813-6-9.

Iomdin L., Sizov V. 2009: Structure Editor: a Powerful Environment for Tagged Corpora. In: Erjavec (Ed.), Research Infrastructure for Digital Lexicography. Proceedings of the MONDILEX Fifth Open Workshop, 14-15 October 2009, Ljubljana. Informacijska družba Publ. House, Ljubljana, 2009, 1-12. ISSN 1581-9973/ISBN 978-961-264-012-5.

Iomdin L., Sizov V. 2008: A Specialized Software Systems for Managing Electronic Bilingual Dictionaries. In: Iomdin, Dimitrova (Eds.), Lexicographic Tools and Techniques. Proceedings of the MONDILEX First Open Workshop, Moscow, 3-4 October 2008. ISBN 978-5-9900813-6-9.

Iomdin et al. 2009: L. Iomdin, S. Timoshenko, I. Boguslavsky, T. Frolova. Development of the Russian Tagged Corpus with Lexical and Functional Annotation. In: Garabík (Ed.), Metalanguage and Encoding Scheme Design for Digital Lexicography. Proceedings of MONDILEX Third Open Workshop 15-16 April, 2009, Bratislava, Slovakia. 83-90. ISBN 978-80-7399-745-8.

Ivanovska et al. 2006: Ivanovska, Aneta, Katerina Zdravkova, Tomaž Erjavec, and Sašo Džeroski. Learning Rules for Morphological Analysis and Synthesis of Macedonian Nouns, Adjectives and Verbs. In: Proceedings of 5th Slovenian and 1st international Language Technologies Conference, Jožef Stefan Institute, Ljubljana.

Javoršek, Jan Jona, Erjavec, Tomaž (2009). Empowering Human Language Technologies with Grid. In: Erjavec (Ed.), Research Infrastructure for Digital Lexicography. Proceedings of the MONDILEX Fifth Open Workshop, 14-15 October 2009, Ljubljana. Informacijska družba Publ. House, Ljubljana, 2009, 13-19. ISSN 1581-9973/ISBN 978-961-264-012-5.

Kemps-Snijders et al. 2008: Marz Kemps-Snijders, Menzo Windhouwer, Peter Wittenburg, and Sue-Ellen Wright. ISOcat: Corralling Data Categories in the Wild. In LREC'08. Paris.

Kilgarriff et al., 2004: Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. The Sketch Engine. Proceedings of EURALEX 2004, Lorient, France.

Koseska–Toszewa, V. 2009a: Many-volume Contrastive Grammar of Bulgarian and Polish. In: Shyrokov, Dimitrova (Eds.), Organization and Development of Digital Lexical Resources. Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009. 87-97. ISBN 978-966-507-252-2.

Koseska-Toszewa, V. 2009b: Form, Its Meaning, and Dictionary Entries. In: Garabík (Ed.), Metalanguage and Encoding scheme Design for Digital

Lexicography. Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009. 105–111. ISBN 978-80-7399-745-8.

Koseska, V., Dimitrova, L., Roszko R. (Editors 2009). Representing Semantics in Digital Lexicography. SOW, Warsaw, 2009. 224 pages. ISBN 978-83-89191-87-8

Koseska, Korytkowska, Roszko 2007: Koseska–Toszewa V., Korytkowska M., Roszko R. Polsko-bułgarska gramatyka konfrontatywna. Warszawa: Wydawnictwo Akademickie Dialog. (In Polish)

Koseska, Mazurkiewicz 2009a: Koseska–Toszewa, V., Mazurkiewicz, A. Net-Based Description of Modality in Natural Language (on the Example of Conditional Modality). In: Shyrokov, Dimitrova (Eds.), Organization and Development of Digital Lexical Resources. Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009. 98-105. ISBN 978-966-507-252-2.

Koseska, Mazurkiewicz 2009b: Koseska–Toszewa, V., Mazurkiewicz, A. On the Meaning of Verbal Forms and Its Net representation. In: Garabík (Ed.), Metalanguage and Encoding scheme Design for Digital Lexicography. Proceedings of the MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009. 112-118. ISBN 978-80-7399-745-8.

Koseska, Mazurkiewicz, 2010: V. Koseska, A. Mazurkiewicz. Time Flow and Tenses. SOW, Warsaw. 223 pages. ISBN 978-83-89191-94-6.

Koseska, Roszko 2008: Koseska-Toszewa, Violetta and Roman Roszko. Remarks on Classification of Parts of Speech and Classifiers in an Electronic Dictionary. In: Iomdin, Dimitrova (Eds.), Lexicographic Tools and Techniques. Proceedings of the MONDILEX First Open Workshop, 3-4 October 2008, Moscow. 80-88. ISBN 978-5-9900813-6-9.

Krek, S., Erjavec, T. 2009: Standardised Encoding of Morphological Lexica for Slavic Languages. In: Shyrokov, Dimitrova (Eds.), Organization and Development of Digital Lexical Resources. Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009. 24-29. ISBN 978-966-507-252-2.

Krygin, M. 2009: Statistical Methods Used for Comparison and Analysis of Texts. In: Shyrokov, Dimitrova (Eds.), Organization and Development of Digital Lexical Resources. Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009. 106-112. ISBN 978-966-507-252-2.

Laccetti, G. and Schmid, G. 2007: A framework model for grid security. Future Generation Computer Systems, 23(5), 702 – 713.

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

Levická 2008: Jana Levická. Analysis of “classical” and legislative definitions for the term record s of the Slovak terminology database. Proceedings of the Third Conference on Translation, Interpreting and Comparative Legi-Linguistics. Poznań, Poland.

Luchick, V. 2009: Problems of creation etymological dictionary of suffixes of Ukrainian language. In: Shyrokov, Dimitrova (Eds.), Organization and Development of Digital Lexical Resources. Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009. ISBN 978-966-507-252-2.

Lyubchenko, T. 2009: Modelling of a Grammar Dictionary of Russian. In: Shyrokov, Dimitrova (Eds.), Organization and Development of Digital Lexical Resources. Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009. 73-84. ISBN 978-966-507-252-2.

Manandhar S., Džeroski S. and Erjavec T. (1998). Learning Multilingual Morphology with CLOG. In: Proceedings of Inductive Logic Programming; 8th International Workshop ILP-98 (Lecture Notes in Artificial Intelligence 1446), 135–144. Springer-Verlag, Berlin.

Mazurkiewicz A. 2008: A formal description of temporality (Petri net approach). In: Iomdin, Dimitrova (Eds.), Lexicographic Tools and Techniques. Proceedings of the MONDILEX First Open Workshop, 3-4 October 2008, Moscow. 98-108. ISBN 978-5-9900813-6-9.

Miko, F. et al. 1989: Frazeológia v škole. Bratislava: Slovenské pedagogické nakladateľstvo.

Mlacek, Profantová 1996: J. Mlacek, Z. Profantová. Slovenské príslovia a porekadlá, zv. 1–2. Výber zo zbierky A. P. Zátureckého. Bratislava: Nestor.

Nivre et al. 2008: Nivre, Joakim, Igor Boguslavsky, Leonid Iomdin. Parsing the SYNTAGRUS Treebank of Russian. Proceedings of the COLING'2008 - 22nd International Conference on Computational Linguistics. Vol. 2, 641-648. ISBN 978-1-905593-47-7.

Ostapova, I. 2009: Digital Etymology (Illustrated by the example of the Etymological Dictionary of Ukrainian language): In: Shyrokov, Dimitrova (Eds.), Organization and Development of Digital Lexical Resources. Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009. 68-72. ISBN 978-966-507-252-2.

Padala et al. 2007: Padala, P., Zhu, X., Wang, Z., Singhal, S., and Shin, K. G. Performance evaluation of virtualization technologies for server consolidation. Technical report, HP Laboratories.

Pajas, P. and Štěpánek, J. 2006: XML-based representation of multi-layered

annotation in the PDT 2.0. In: Hinrichs, R. E., Ide, N., Palmer, M., Pustejovsky, J. (Eds.), Proceedings of the LREC Workshop on Merging and Layering Linguistic Information (LREC 2006), 40-47, Genova, Italy.

Pala et al. 2009: Pala, K., A. Rambousek, M. Khokhlova, and V. Zakharov. Russian Dictionary Base – First Steps. In: Garabík (Ed.), Metalanguage and Encoding scheme Design for Digital Lexicography. MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15-16 April 2009. ISBN 978-80-7399-745-8.

Parizoska 2009: Parizoska, Jelena. Idiom variability in Croatian: the case of the container schema. In: Koseska, Dimitrova, Roszko (Eds.), Representing Semantics in Digital Lexicography. Proceedings of the MONDILEX Fourth Open International Workshop, Warsaw, Poland, 29 June–1 July 2009. 123-127. ISBN 978-83-89191-87-8.

Polyuga 2001: Polyuga, L.M. Dictionary of Antonyms of the Ukrainian language: 2-ге вид., доп. і випр. / Под ред. Л.С. Паламарчук; НАН України; Інститут українознавства ім. І. Крип'якевича; Український мовно-інформаційний фонд – Kiev, Довіра, 2001. 276 p. (In Ukrainian)

Pugliese 2004: Pugliese A., Talia D. Application-Oriented Scheduling in the Knowledge Grid: A Model and Architecture. In: Proceedings of the International Conference on Computational Science and its Applications (ICCSA), Assisi, Italy, vol. 2, 55-65, Springer-Verlag, April 2004.

Rabulets, A. 2009: System Engineering Principles of Virtual Linguistics Laboratories. In: Shyrokov, Dimitrova (Eds.), Organization and Development of Digital Lexical Resources. Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009. 18-23. ISBN 978-966-507-252-2.

Roszko, R. 2009: Morphosyntactic Specifications for Polish. Theoretical foundations. Description of morphosyntactic markers for Polish nouns within MULTEXT-East Morphosyntactic Specifications (Version 3.0 May 10th, 2004). In: Garabík (Ed.), Metalanguage and Encoding Scheme Design for Digital Lexicography. Proceedings of the MONDILEX Third Open Workshop, 15-16 April 2009, Bratislava. 140-150. ISBN 978-80-7399-745-8.

Rychlý and Smrž, 2004: Rychlý, P. and Smrž, P. Manatee, Bonito and Word Sketches for Czech. In Proceedings of the Second International Conference on Corpus Linguistics, Saint-Petersburg. Saint-Petersburg State University Press, 2004. 124–132. ISBN 5.-288.-03531.-8

Santos, N., and Koblitz, B. 2008: Security in distributed metadata catalogues. Concurrency and Computation: Practice and Experience, 20(17), 1995-2007.

Scott, M., 2004: WordSmith Tools version 4, Oxford: Oxford University Press.

Shevchenko, I. 2009: Towards Creation of the Polish Grammatical Dictionary. In:

Shyrokov, Dimitrova (Eds.), Organization and Development of Digital Lexical Resources. Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009. 61-65. ISBN 978-966-507-252-2.

Shi 2006: Shi Xiaoqing, Zhao Jingzhu, Zhiyun Ouyang. Assessment of eco-security in the Knowledge Grid e-science environment. In: The Journal of Systems and Software, 79 (2006), 246-252.

Shi 2006: Shi Xiaoqing, Zhao Jingzhu, Zhiyun Ouyang. Assessment of eco-security in the Knowledge Grid e-science environment. In: The Journal of Systems and Software, 79 (2006), 246-252.

Shyrokov 1998: The Information Theory of the Lexicographic Systems. Kiev, 1998. (In Ukrainian)

Shyrokov 2004: Широков В. А. Феноменологія лексикографічних систем. Київ: Наукова думка, 2004, 326 с. (In Ukrainian)

Shyrokov 2005: Широков В.А. Семантичні стани мовних одиниць та їх застосування в когнітивній лексикографії. Мовознавство, 2005, №№ 3-4. (In Ukrainian)

Shyrokov 2008: V.A. Shyrokov. Integral Slavic Lexicography in the Linguotechnological Context. In: Iomdin, Dimitrova (Eds.), Lexicographic Tools and Techniques. Proceedings of the MONDILEX First Open Workshop, Moscow, 3–4 October 2008. 23-30. ISBN 978-5-9900813-6-9.

Shyrokov, V. 2009a: The National Dictionary Base of Ukraine. In: Shyrokov, Dimitrova (Eds.), Organization and Development of Digital Lexical Resources. Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009. 5-8. ISBN 978-966-507-252-2.

Shyrokov, V. 2009b: Theory of Lexicographic Systems. Part 1. In: Garabík (Ed.), Metalanguage and Encoding scheme Design for Digital Lexicography. MONDILEX Third Open Workshop, Bratislava, Slovak Republic, 15–16 April 2009. 140-150. ISBN 978-80-7399-745-8.

Shyrokov, V. 2009c: Theory of Lexicographic Systems. Part 2. In: Koseska, Dimitrova, Roszko (Eds.), Representing Semantics in Digital Lexicography. Proceedings of the MONDILEX Fourth Open International Workshop, Warsaw, Poland, 29 June–1 July 2009. 89-105. ISBN 978-83-89191-87-8.

Shyrokov, V. 2009d: Theory of Lexicographic Systems. Part 3. In: Erjavec (Ed.), Research Infrastructure for Digital Lexicography. Proceedings of the MONDILEX Fifth Open Workshop, 14-15 October 2009, Ljubljana. Informacijska družba Publ. House, Ljubljana, 2009. 98-119. ISSN 1581-9973/ISBN 978-961-264-012-5.

Shyrokov et al. 2005: Shyrokov V., Bugakov O., Griaznukhina T., etc. Corpus Linguistics. Kiev, 2005. 471 pages. ISBN 966-507-189-0. (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. //В.А. Широков, О.Г. Рабулець, І.В. Шевченко, К.М. Якименко. Інтегрована лексикографічна система "Словники України". Електронне видання на лазерному диску.// Kiev, 2009. ISBN 966-507-149-1.

Shyrokov, V., Dimitrova, L. (Editors 2009). Organization and Development of Digital Lexical Resources. Proceedings of the MONDILEX Second Open Workshop, Kiev, Ukraine, 2–4 February 2009. Dovira Publ. House, Kiev, Ukraine. 127 pages. ISBN 978-966-507-252-2.

Šimková et al. 2009: Šimková, M., Garabík, R., Dimitrova, L. Design of a multilingual terminology database prototype. In: Koseska, Dimitrova, Roszko (Eds.), Representing Semantics in Digital Lexicography. Proceedings of the MONDILEX Fourth Open International Workshop, Warsaw, Poland, 29 June–1 July 2009. SOW, Warsaw, 2009, 123-127. ISBN 978-83-89191-87-8

Smiešková, E. 1988: Malý frazeologický slovník. Bratislava: Slovenské pedagogické nakladateľstvo. Bratislava.

Sperberg-McQueen, and Burnard, 2002: Sperberg-McQueen, C. M. and Burnard, L., (Editors). Guidelines for Electronic Text Encoding and Interchange, The XML Version of the TEI Guidelines. The TEI Consortium.

Spoustová et al. 2009: Spoustová, D., Hajič, J., Raab, J., and Spousta, M. Semi-supervised training for the averaged perceptron POS tagger. In: EACL ’09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, 763–771, Morristown, NJ, USA. Association for Computational Linguistics.

Stanojević, Kryżan-Stanojević 2009: Stanojević, M., Kryżan-Stanojević, B. Croatian and Polish - Confluence of the Dative and Middle Voice. In: Koseska, Dimitrova, Roszko (Eds.), Representing Semantics in Digital Lexicography. Proceedings of the MONDILEX Fourth Open International Workshop, Warsaw, Poland, 29 June–1 July 2009. SOW, Warsaw. 123-127. ISBN 978-83-89191-87-8

Stork 2001: Hans-Georg Stork. The GRID, the WEB and KNOWLEDGE. http://www.cikon.de/Text_EN/gwk.html

Tamburini, F., 2004. Building distributed language resources by grid computing. In Proc. of the 4th International Language Resources and Evaluation Conference. pp. 1217–1220.

TEI (2007). TEI P5: Guidelines for Electronic Text Encoding and Interchange.

Tu et al. 2007: Tu W., Muppala J. K. and Zhuge H., Distributed End-Host Multicast

Algorithms for the Knowledge Grid, Concurrency and Computation: Practice and Experience, 19 (15), 2007.

Tvrdý, P. 1931. Slovenský frazeologický slovník. Trnava: Spolok sv. Vojtecha.

Tvrdý, P. 1933. Slovenský frazeologický slovník. Druhé doplnené vydanie. Praha and Prešov: Nákladom Ceskoslovenskej grafickej unie, úc. Spol.

Ukrainian Phraseology 2003: Dictionary of Phraseologisms of the Ukrainian language. Kiev. Наук. думка, 2003. 1104 p. (In Ukrainian)

Vintar 2004: Vintar, Špela. Comparative Evaluation of C-value in the Treatment of Nested Terms. Memura 2004 - Methodologies and Evaluation of Multiword Units in Real-World Applications (LREC 2004), 54-57.

Záturecký, A. P. 1896: Slovenská přísloví, pořekadla a úsloví. Praha: Česká akademie věd.

Záturecký, A. P. 2006: Slovenské príslovia, porekadlá, úslovia a hádanky. Bratislava: Slovenský Tatran.

Žabokrtský et al. 2008: Žabokrtský Z., Ptáček, J., and Pajas, P. TectoMT: Highly modular MT system with tectogrammatics used as transfer layer. In: ACL 2008 WMT: Proceedings of the Third Workshop on Statistical Machine Translation, 167 – 170, Columbus, OH, USA. Association for Computational Linguistics.

http://nl.ijs.si/ME

http://nl2.ijs.si/analyze/

http://ulif.org.ua

Conceptual scheme for a research infrastructure supporting development of digital resources and research in Slavic lexicography

Sofia, Institute of Mathematics and Informatics 2010

Published by:
Demetra Ltd Publishers, 2010
Acad. G. Bonchev St. bl. 8
1113 Sofia
Bulgaria

ISBN 978-954-8986-33-5