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

This report presents a conceptual scheme of networking of centres for high-quality research in Slavic lexicography and their language digital resources. The report, produced by the MONDILEX project Consortium, summarizes the studies and analysis, presented at the five open workshops, organised in the framework of the project. The experts' recommendations on standardisation and integration of Slavic languages resources, on establishment of a virtual organisation supporting research infrastructure for Slavic lexicography, on rights and metadata management to work within grid supported network, etc. are also discussed and summarized. The results of the project MONDILEX will be very useful for a future project proposal for implementation of such research infrastructure for the coming years.

The Consortium of the GA no. 211938 project MONDILEX is composed of six members appointed as centres for high-quality research in area of NLP of Slavic languages. The partners are research organisations from six European countries, whose national languages belong to the Slavic group – Bulgaria, Poland, Russia, Slovakia, Slovenia and Ukraine, four member states and two from ICP Countries:

1. Institute of Mathematics and Informatics, Bulgarian Academy of Sciences (IMI-BAS) (coordinator)
2. Institute of Slavic Studies, Polish Academy of Sciences (ISS-PAS)
3. Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences (LSIL)
4. Jožef Stefan Institute, Ljubljana, Slovenia (JSI)
5. Institute for Information Transmission Problems, Russian Academy of Sciences (IITP-RAS)
6. Ukrainian Lingua-Information Fund of the National Academy of Sciences of Ukraine (ULIF-NASU)

Contact information:

Dr. Ludmila Dimitrova, Assoc. Prof., IMI-BAS, Sofia, Bulgaria, *Coordinator of the MONDILEX project*, ludmila@cc.bas.bg
Prof. Violetta Koseska, DSc, ISS-PAS, Warsaw, Poland, amaz@inetia.pl
Dr. Radovan Garabík, LSIL, Bratislava, Slovakia, garabik@kassiopeia.juls.savba.sk
Dr. Tomaž Erjavec, Assoc. Prof., JSI, Ljubljana, Slovenia, tomaz.erjavec@ijs.si
Dr. Leonid Iomdin, IITP-RAS, Moscow, Russia, iomdin@iitp.ru
Prof. Volodymyr Shyrokov, ULIF-NASU, Kiev, Ukraine, shirokov@ulif.org.ua

Project web-site: www.mondilex.org

Main objectives

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.

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

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

The MONDILEX project examined strategies for the coordination, unification, integration and extension of existing digital lexical resources and the creation of new ones, in accordance with recent advances in the field and international standards. These resources include those developed along the lines of best practices and recommendations: corpora (mono- and multilingual, parallel, comparable, and annotated), dictionaries (mono- and bilingual, electronic and online), lexicons, thesauri, wordnets, ontologies etc.

The series of five MONDILEX open workshops investigated the following problems. The first workshop analysed the needs of the partners for a common infrastructure supporting scientific and applied activities in digital lexicography. The second workshop studied the state of the art in digital lexical resources and requirements for their integration. The third workshop tackled innovative solutions for lexical entry design in digital Slavic lexicography. The representation of semantics, phraseology, etymology and related matters were discussed in the fourth workshop. The last workshop concentrated on the research infrastructure for Slavic lexicography.

1. Conceptual scheme of a research infrastructure supporting the networking of centres for high-quality research in Slavic lexicography

MONDILEX presented a conceptual scheme of a research infrastructure supporting the networking of centres for high-quality research in Slavic lexicography. During the project, **all types of language resources** that should be included in a research infrastructure for Slavic lexicography were discussed in detail and evaluated. MODILEX surveyed also the fundamental concepts of traditional and digital lexicography.

1.2. Evaluation of Slavic language resources for digital lexicography

At first, a set of **lexical databases (**LDBs) for Slavic languages were analyzed and discussed, including a Slovak-Czech LDB, a Bulgarian-Polish LDB, a multilingual corpus linguistics terminology database, a Slovak morphology database, a paremiography database. The LDB analysis focused on the problems and difficulties of database support arising due to LDB's internal logical complexity, alignment of the structure and content tags of LDB's

structural units to international standards, as well as compatibility with language resources created in other projects and for other languages. In this context, some conceptual models of actual electronic database were described, among them, Slovak-Czech LDB and multilingual terminology database, both compiled with the MoinMoin wiki system, Bulgarian-Polish LDB based on CONCEDE[1] model for dictionary encoding. The proposed structure of LDB allows synchronized and unified representation of the information for languages.

Furthermore, MONDILEX designed and implemented a prototype of a multilingual corpus linguistics terminology database intended to facilitate collaboration among MONDILEX member institutes. The aim of this database is to minimize the hindrance of internal communication due to either missing or incompatible Slavic languages terminology of modern aspects of linguistics and to unify existing terminology. The content of database includes corpus linguistics entries from the Slovak Terminology Database, and contains terms in Bulgarian and Slovak, with relevant English equivalents. Future extensions shall proceed towards creating a database of all languages of the MONDILEX project, including English (added as a hub language, and also because most terminology originates there). Such a database can serve as a nucleus of a multilingual terminology database of lexicographic (or even general linguistic) terms.

Second, different kinds of applications were exemplified by the following **digital dictionaries of Slavic languages**: the dictionary of Slovak collocations, Bulgarian-Polish on-line dictionary, Ukrainian online dictionaries, and semantic concordance in Slovene. One well-known fact is that the dictionaries could be treated simultaneously as text and as databases, because 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.

The project concentrated also on **representation of semantics problems in Slavic digital lexicography** in the context of the very substantial growth of digital dictionaries of all types. For example, one problem is the development of innovative solutions for lexical entry content in Slavic lexicography. Determining the content of a lexical entry in bi- and multilingual digital dictionaries is a complex task having to do with the description of linguistic forms with various meanings in the languages concerned. The difficulty stems from the fact that so far, the form rather than the content has been the starting point for language description. A second problem here is distinguishing between the form and its meaning in the comparative material from the six languages of the MONDILEX project, in order to avoid numerous substantial mistakes and erroneous conclusions. To achieve this goal, MONDILEX concluded that a *semantic interlanguage* or a *dictionary/lexicon of concepts* be developed, on which multilingual dictionaries should be based. Such a universal dictionary of concepts is a language-independent intermediary lexical tool, developed as a part of the effort to create a semantic language for global information exchange. It can evolve further into an open and freely available language-neutral resource, a tool to uniformly record and link meanings of words of different languages and help the creation of bi- and multilingual dictionaries. A third

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

problem MONDILEX discussed is the representation of temporal situations and some issues of the description of modality, using Petri nets formalism. A *catalogue of descriptions of temporal and modal situations*, expressed in different languages, was published in Warsaw. The entries in this catalogue are parameterized names of temporal and modal situations, and the corresponding values precise formal descriptions of such situations. The catalogue contains a collection of studies on temporal subjects, analyzed in accordance with the methodology of cognitive linguistics. The catalogue can be used to create a language-independent list of basic temporal situations. *Semantic concordances* are another useful resource for a wide range of applications, such as automatic word sense disambiguation or corpus-based studies of sense frequency, distribution and co-occurrence, and are also invaluable as an aid for translation as well as for vocabulary acquisition in a foreign language. MONDILEX described the following set of **corpora** as resources for digital lexicography: *multilingual* parallel and annotated corpora – Bulgarian-Polish, Polish-Ukrainian, and *monolingual* – morphologically and syntactically tagged corpus of Russian SynTagRus, Slovene language corpus with semantic annotation.

Next, MONDILEX discussed **morphosyntactic annotations** in Slavic digital lexicography. MULTEXT-East[2] (MTE) morphosyntactic specifications, and especially standardisation of Slavic lexicographic resources and their metadata were discussed and described in full with an emphasis on the importance of the developed harmonised lexical specifications in CES format and of the language independence of the tools. MONDILEX discussed also the Text Encoding Initiative recommendations, an XML-based framework suitable for encoding a wide variety of text types, from those constituting digital libraries, to machine readable dictionaries, and annotated corpora; e.g. a TEI based encoding for linguistic annotation of corpora is now being proposed in the scope of CLARIN initiative. TEI is also suitable for encoding machine readable dictionaries, however, TEI does not have a module for lexical databases, but a model for those has been recently proposed as the ISO standard LMF, "Lexical Markup Framework". In addition, MONDILEX made a proposal for lexical encoding concentrating on morphological properties of words, esp. of the strongly inflecting Slavic languages. The format of this encoding is an application of the new ISO standard LMF; the core lexical structure and morphosyntactic annotation are from MTE project, with recent extensions for Slovene. MTE language resources represent standardised multilingual 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 previous editions of specifications covered Bulgarian, Croatian, Czech, Estonian, English, Hungarian, Romanian, Serbian, Slovene, and the Resian dialect of Slovene. The **fourth release of these resources was developed in the scope of the MONDILEX** project and introduces XML-encoded morphosyntactic specifications, using the latest version of the Text Encoding Initiative Guidelines, TEI P5. The MONDILEX edition adds Macedonian, Polish, Russian, Slovak, Ukrainian, and Persian. The specifications now cover 10 Slavic languages, providing a good basis for a unifying morphosyntactic framework for digital Slavic lexicography.

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

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

---

[2] The EC project MULTEXT-East *Multilingual Text Tools and Corpora for Central and Eastern European Languages*, a continuation of the MULTEXT project. See http://nl.ijs.si/ME/.

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

The other is MoinMoin, very successful software written in the Python programming language, and as such particularly interesting because of the ease of installation, usage and customisation. MoinMoin is a wiki written completely in the Python programming language, using flat text files as a storage backend, rather than a database. This makes it particularly attractive for the needs of digital lexicography, because of the programming language involved and the ease of making various data modifications and extraction, using just common text processing tools. MoinMoin is also fully Unicode aware, and all the stored data, output and input is invariably in UTF-8 encoding. 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.

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

1.2. Standardisation of Slavic Lexicographic Resources

Lexicographic resources, in particular machine readable dictionaries, lexical databases, and mono- or multilingual annotated text corpora are developed and stored in a variety of formats, which makes them difficult to process on a common platform and to achieve interchange between programs and applications. The effectiveness of language technologies ultimately depends on the quantitative and qualitative parameters of the lexicographic description of units, relations and levels of language on which these technologies are based. If we take into account the importance of finding a solution to the problem of multilingualism in the global information medium, then we quite reasonably formulate the task of integration of lexicographical descriptions for all languages, i.e., the compiling of a dictionary for the entire mankind, a unique *Summa lexicographiae*.

This section proposes several mutually reinforcing recommendations which can serve to overcome this obstacle. All the proposed frameworks have already been extensively tested in practice and, in certain cases, further developed in the scope of the MONDILEX project. The work of the project demonstrates the potential for developing useful lexicographic reference works (both digital and hardcopy) by using the format of the lexical data base and an adequate mathematical foundation. Various parameters of classification of the lexicon are likely to emerge in the process of developing the lexical data base, possibly through

distributed effort, which highlights the importance of the interface to the lexicographic system. The lexical data bases forming the foundation of the dictionaries should be brought in line with one another by sharing theoretical concepts and platforms.

Design of a common encoding scheme for Slavic multilingual dictionaries:

1. The use of modern database technologies for fast access to dictionaries requires careful design and implementation of an underlying data structure and storage.
2. 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.

The previous points can be partly met by using advanced wiki-based collaboration editing systems. The following recommendations are made:

3. Unification of the classifiers of the headword in the dictionary entry. The headwords in the dictionary entries of the digital dictionary must be indexed according to the number of meanings, and each meaning must be related unambiguously to the form. In this manner most meanings of the form can be encompassed. Such a description might require more classifiers, but also provide a more adequate correspondence.
4. Unification of the systems of categories and tags used for annotation in the various systems.
5. A uniform presentation of the lexical entry content.
6. Creation of a corpus linguistics terminology database 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.
7. Creation of a special digital lexicographic environment adapted to the LDBs and digital dictionary entry structures and oriented to the creation of a multilanguage index in the automatic mode is necessary.

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 should also 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.
2. Knowledge Grid – a technological platform for a future implementation
MONDILEX investigated the features of Knowledge 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. 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.

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.

2.1. Virtual research infrastructure

The Grid computing technology, as a form of distributed computing where a "virtual supercomputer" is composed of a cluster of networked, has been applied to computationally-intensive problems, requiring the storing and sharing of large amounts of data, in many areas of science. Some domains where it has been used (e.g., processing data from medical records) demand a high level of data protection and controlled access. User authentication and digital rights management is part of the grid infrastructure. Because of this overlap of requirements, this paradigm has started being applied to the area of Human Language Technologies, esp. to areas which deal with large amounts of data, i.e., with corpora. While virtual organizations in modern grids are self-contained infrastructure elements, they must be included in the common infrastructure of all sites supporting the virtual organization. The key points of *Grid infrastructure requirements* needed for supporting research activities in digital lexicography that could be mentioned here, are: virtualization techniques, specific legal issues (the data to be processed are in most cases copyrighted, and the research institutions either have very

strict legal agreements governing the use of the data, or are operating entirely on copyright law sections allowing scientific and research use of the data), security measures used in the Grid infrastructure (public key infrastructure, virtual organizations, proxy certificates, and data protection).

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

In order to provide the power of grid computing to researchers in the domains of digital lexicography, corpus processing and human language technologies in general, the technology needs to be accessible as a part of dedicated grid infrastructure. 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 a number of steps are described 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. This is the central server for the Virtual Organization user and server access control, including accreditation, authentication and authorization. To use the server, a user (organization or person) has to get a grid digital certificate for authentication and use the server to apply for accreditation. To support the 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. HLT VO VOMS is supported by the SiGNET cluster. Any organization wanting to participate in the HLT VO can enroll with the VOMS to use the infrastructure and include its configuration in the local set-up to support the infrastructure locally.

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

Registration of the VO

While Virtual Organizations in modern grids are self-contained infrastructure elements, they have to be included in the common infrastructure of all sites supporting the Virtual Organization (VO). Two different grid middleware solutions shall be supported: NorduGrid and gLite. NorduGrid ARC is simpler and very efficient, and is a good match for applications that, in grid terms, are not very resource intensive. It is also easier for setting up new sites due to much simpler installation and integration procedures. gLite from the EGEE project is, on the other hand, the most widely used and supported middleware and therefore has to be supported by the HLT VO.

As soon as HLT VO is registered, it will be discoverable using the central services of EU Grid infrastructure (i.e. with the EGEE and NorduGrid projects). It is also expected to become one of the supported VOs in the future European Grid Initiative (which is to start its operations in 2010). After the VO is registered, as 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 should be set up. 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 an EU Grid computing environment. One of the major attractions of the new system, next to the flexibility, compatibility of tools and the sheer computing and storage power, will be to provide a single method (and programming API) to many resources in different languages, and to resolve the difficulties inherent in different legal, technical and practical restrictions that make any multilingual research rather difficult today.

Some web-based interfaces to the resources incorporated in the grid shall also be added. The first of such planned services will be a grid-aware concordancer, accessible both as a web service and from grid jobs. The service will enable the user to access the available grid-based corpora according to user's authorization.

Initial processing pipeline

For testing purposes, a set of command-line tools for submitting typical linguistic grid jobs will be developed, based on a basic set of tools that will be prepared for the use on the grid (gridified) for testing purposes. These tools will have either the form of dedicated scripts or specialized makefiles and will be able to perform a resource-intensive task using distributed corpus data and distributed computing resources in the HLT VO.

In the first stage, the lemmatization and tagger tool **totale** shall be gridified. The tool has been extensively tested with TEI P5 encoded corpora and MTE tag sets. In addition a small test suite of generic *n*-gram processing tools is being prepared for statistical analyses on available distributed data. These prototype processing pipelines will serve as test cases and foundations

for developing more complex pipelines, user frameworks and web interfaces for advanced grid-based linguistic processing in the future.

2.3. Rights and Metadata Management

Obtaining digital certificates for users and sites: in order to work with the grid, users have to establish their digital identity using the international public key infrastructure for grids, in this case the International Grid Trust Federation established by regional grid certificate policy management authorities (EU Grid PMA for Europe). In practice, all members of the MONDILEX project can contact their national grid certificate issuers and national grid initiative organizations or start-up projects in their countries to receive help, training and instructions on how to obtain necessary digital certificates. In order to create a testing framework, all users and future grid site managers will have to submit requests for personal digital certificates that will be used for authorization with the HTL VO and grid services. Administrators will also have to submit requests for service and hosts certificates for services and machines that will be using the grid directly (local grid site servers, job managers and data pools). Grid certificates are used for identifying users to VO web services and to the VOMS service (Virtual Organization Membership Service).

The list of the **National Grid Contact Points** of MONDILEX project participants' countries follows:

Bulgaria: Institute for Parallel Processing, BAS CA: BG.ACAD CA http://www.ca.acad.bg/

Poland: PL-Grid, CYFRONET CA: PolishGrid http://www.man.poznan.pl/plgrid-ca

Slovakia: Institute of Informatics, SAS CA: SlovakGrid CA http://ups.savba.sk/ca

Slovenia: Jozef Stefan Institute CA: SiGNET CA http://signet-ca.ijs.si

Russia: Russian Data Intensive Grid, Skobeltsyn Institute of Nuclear Physics CA: Russian Data Intensive Grid http://ca.grid.kiae.ru/RDIG/

Ukraine: National Academy of Sciences (Gennady Zinovjev) Ukrainian Grid CA https://ca.ugrid.org/

VO authorisation protocol: VOMS (Virtual Organization Membership Service) originally developed in the framework of EDG and DataTAG collaborations and maintained by the EGEE project, is the industry standard Virtual Organization management solution, shared by all current grid middleware implementations. VOMS identifies users using their personal grid certificates. The server classifies users that are part of a VO on the base of a set of attributes in its database and includes that information inside Globus-compatible proxy certificates generated from user certificates, which enable users and their jobs to fully identify the user and authorize their actions in the grid based on the attributes from VOMS. The system is fine-grained, reliable, scalable, highly secure, supported and widely used. VOMS attributes can be used to control user access both to VO-wide services and capabilities, such as software repositories and file pool servers, and to specific resources, such as executing grid jobs and files stored on the grid (where the ownership and permissions of the job and file are taken into account). Since stored files can be encrypted and all file transmissions are performed using secured encrypted links and PKI-based authentication, the system enables fine-grained and secure control over file storage, access and manipulation.

The system is sufficiently versatile and secure to enable us to share even those linguistic resources that require user agreements or contractual relationships to be used. Since the fine-grained controls will allow us to restrict all access to such resources (i.e. corpora) to only those users that have legal access rights, it should be possible to facilitate and simplify the process without jeopardizing the security of data and copyright protections in question.

Creation of the resource catalogue: the main resources for linguistic research and lexicography, electronic corpora, represent the basic resources that should be available in the HTL VO gird. Compared to most other scientific disciplines, corpus metadata is rather complex. So the level of detail needed to create a useful central metadata server with dataset selection, searching and extraction capabilities shall be evaluated.

Conclusion

The project MONDILEX provided a **venue for networking activities**, such as joint management and pooling of resources, implementation of standards for products of digital lexicography, and coordination with relevant international standards and practices. It demonstrated that unified strategies should contribute to reusability and interoperability of such resources so that researchers in the humanities and social sciences as well as business communities could have easy access to bilingual and multilingual dictionaries of Slavic languages.

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

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

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