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

Russian Academy of Sciences
Institute for Information Transmission Problems
(Kharkevich Institute)

Lexicographic Tools and Techniques

MONDILEX First Open Workshop
Moscow, Russia, 3—4 October, 2008

Proceedings

Moscow 2008

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

Russian Academy of Sciences
Institute for Information Transmission Problems
(Kharkevich Institute)

Lexicographic Tools and Techniques

MONDILEX First Open Workshop
Moscow, Russia, 3—4 October, 2008

Proceedings

Leonid Iomdin, Ludmila Dimitrova (Eds.)

The workshop is organized by the project
GA 211938 MONDILEX
Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources
supported by EU FP7 program
Capacities - Research Infrastructures
Design studies for research infrastructures in all S&T fields

Moscow 2008

Le63
УДК 80/81; 004.
ББК 81.1.

Lexicographic Tools and Techniques.
Moscow, IITP RAS, 2008. – 109 p.

The volume contains contributions presented at the First open workshop "Lexicographic tools and techniques", held in Moscow, Russia, on 3—4 October 2008. The workshop is organized by the international project GA 211938 MONDILEX *Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources,* Capacities – Research Infrastructures (Design studies for research infrastructures in all S&T fields) EU FP7 programme.

Editor of the volume**: Olga Shemanayeva,** Moscow, Russia

ISBN 978-5-9900813-6-9

Contents

FOREWORD

This volume contains contributions presented at the MONDILEX project First open workshop "Lexicographic tools and techniques", held in Moscow on 3—4 October 2008. The workshop is organized by the international project GA 211938 MONDILEX *Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources,* Capacities - Research Infrastructures (Design studies for research infrastructures in all S&T fields) Project developed under EU FP7 programme.

The purpose of this Workshop was to study, compare and generalise the partners' requirements for a common research infrastructure supporting research activities in digital lexicography, as well as offer recommendations for a common infrastructure supporting high-quality research in Slavic digital lexicography. The papers discuss current trends and achievements in the field of digital lexicography, especially for Slavic languages.

The first part of the volume, "Toward the Common Platform of Digital Slavic Lexicographic Resources", is dedicated to the creation of a multilingual multi-access resource that can stimulate various cross-linguistic activities. The paper by T. Erjavec and J. J. Javoršek discusses the requirements for grid computing to be applied to the digital lexicography and to enable the creation, annotation and querying of multilingual corpora. The paper by L. Dimitrova and R. Pavlov also deals with grid technologies and their contribution to natural language management, in particular lexicographic activities. Based on the authors' participation in the EC international project MULTEXT-East, some aspects of compatibility of language resources: unification and standardisation are presented. The paper by V. Shirokov considers integral multilingual lexicography in the context of the Mondilex project. It focuses on the relationship between the grammar and lexicographical type of the language system description.

The second part of the volume is "Maintenance and Optimisation of Multilingal Digital Environment". The paper by I. Boguslavsky and V. Dikonov is dedicated to the creation of universal dictionary of concepts that could serve as a semantic intermediary language for global information exchange. The paper by L. Iomdin and V. Sizov presents a sophisticated software tool which is used to expand and update bilingual and monolingual electronic dictionaries. The paper by R. Garabik describes the way of organizing morphology data into a form suitable to be kept as plain text files inside of a MoinMoin wiki engine. Certain practical results of managing Slovak morphology information are given. The paper by K. Simov and P. Osenova discusses the Bulgarian language resources infrastructure developed in several projects aimed at supporting information technology applications.

The four contributions of the third part of the volume are dedicated to the treatment of specific and universal linguistic phenomena in multilingual environment. The paper by I. Derzhanski and N. Kotsyba presents an overview of the category of non-verb predicatives, its definition and coverage in grammars, dictionaries and corpora of four Slavic languages (Russian, Ukrainian, Polish and Bulgarian). The paper by V. Koseska-Toszewa and R. Roszko presents various classifications of parts of speech for Polish, developed on the base of homogenous or mixed criteria: syntactic, semantic and grammatical (morphological) ones. The paper by L. Dimitrova and V. Koseska-Toszewa discusses some issues of entry classifiers in digital dictionaries, e.g., those emerging in the course of the development of a Bulgarian-Polish Digital Dictionary. Lexical specifications for Bulgarian in the EC international project MULTEXT-East, developed on the basis of a semantic and grammatical classification of Bulgarian word forms, are shown. The paper by A. Mazurkiewicz deals with Petri net formalism, showing how it can be used for defining temporal situations.

The preparation of these results has received funding from the EC's Seventh Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX.

We do hope this volume will be of interest to both lexicographers and computer scientists.

Ludmila Dimitrova, Leonid Iomdin

I. Toward the Common Platform of Digital Slavic Lexicographic Resources

Grid Infrastructure Requirements for Supporting Research Activities in Digital Lexicography[1]

Tomaž Erjavec, Jan Jona Javoršek
Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
tomaz.erjavec@ijs.si, jona.javorsek@ijs.si

Abstract
The paper discusses the requirements that need to be met in order for grid computing to be successfully applied to the field of digital lexicography, in particular to corpus processing. We explain the need for grid computing in this context, overview the current state of the grid, and discuss what special aspects are exhibited by grid-based corpus processing tools. We also provide a concrete proof-of-concept scenario, whereby a simple but still useful grid infrastructure would be implemented to enable the creation, annotation and querying of multilingual corpora.

Keywords: grid computing, digital lexicography, corpus processing, language resources, human language technologies

Introduction

Human Language Technologies (HLT), as well as related disciplines, such as digital lexicography, increasingly rely on large annotated corpora as the basic data source, serving such needs as datasets for training and testing language models or for lexical investigations based on naturally occurring data. While digital lexicography needs other tools, in particular those that deal directly with lexica or machine readable dictionaries, the focus of this paper is on the particular subfield of corpus investigations – be it from the viewpoint of easily producing specialised corpora, or from the viewpoint of exploitation of large, annotated, and sharable corpora.

Corpora for various languages can nowadays contain over a billion words, and annotations can increase their size by a factor of ten. Similarly, various automatic annotation tasks, such as word-alignment or semantic indexing can be computationally very expensive, both in the training and application phases; the sophisticated investigations of today's lexicographers are also computationally expensive, where complex searches or other computations need to be performed over such large and heavily annotated corpora. And while computational power of even personal computers has increased dramatically over the years, the sizes of corpora, and the computation resources needed to annotate, store, and investigate them have increased even more.

In addition to requiring large amounts of storage and computing power, lexicographers can also benefit from sharing resources, such as corpora. Of course, due to copyright and other factors, such sharing must be controlled via a system of access rights and permissions.

Grid computing is a form of distributed computing whereby a "virtual supercomputer" is composed of a cluster of networked, loosely-coupled computers, acting in concert to perform very large tasks. This technology has been applied to computationally-intensive scientific, mathematical, and academic problems, in areas such as physics, chemistry, biomedicine, pharmacology and meteorology. Grid computing thus offers the possibility to perform computationally intensive tasks, and offers also the possibility of storing and sharing large amounts of data. Additionally, various domains to which grid computing has already been applied, such as processing of data from medical records, demand a high level of data protection and controlled access: user authentication and digital rights management are a part of the grid infrastructure.

[1] The study and preparation of these results have received funding from the EC's Seventh Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX.

Given this overlap of requirements and their possible solution, it is natural that the paradigm of grid has started to be applied, albeit slowly, and with some time lag, to the area of Human Language Technologies, especially to the areas which deal with processing of large amounts of data, i.e. corpora.

In this paper we review the current state-of-the-art in grid based corpus processing, attempt to delineate what special requirements there are in applying the grid to this domain, and propose a specific, albeit simple scenario, which could be implemented in the scope of the MondiLex project as a test case. Hopefully, the actual implementation of even a proof-of-concept prototype will be already useful, as well as highlighting concrete barriers for the uptake of grid technology for digital lexicography and how to overcome them.

Enabling Grid Infrastructure for Digital Lexicography

Carroll et al. (2005) discuss some of the key challenges which need to be addressed in order to fully support a vision of 'Grid-enabled' NLP. They distinguish between issues relating to the underlying technological requirements of Grid NLP, focusing on data representation, system configuration and deployment on the Grid, and the delivery of this technology to different user groups: end users, service providers and NLP researchers. Recently, more specific initiatives are coming into being. As a part of the German Grid Initiative D-Grid (Neuroth et al., 2007) the TextGrid project aims to create a community grid for the collaborative editing, annotation, analysis and publication of specialist texts. It thus forms a cornerstone in the emerging e-Humanities, and has obvious applications for digital lexicography, although currently concrete results of this project are not yet known. A very interesting project, NLE-GRID, dealing with the grid infrastructure for Natural Language Engineering applications is currently underway at the L2F – Spoken Language Systems Laboratory, INESC ID Lisbon, and a number of concrete proposals for its architecture and other requirements have already been published (Martins de Matos et al., 2008a, 2008b, 2008c; Marujo et al., 2008, Luís, 2008). The main objective of the NLE-GRID project is to create a framework for high performance NLE computing on a computational grid by extending an already existing system of theirs, called Galinha, to include an interface to computational grid services so that the components and data can be geographically and organizationally distributed. The Galinha system consists of a web interface and a data repository. Applications were built through the web interface by creating service chains from a pool of re-usable components. Various tools were integrated in Galinha: these tools performed tasks such as morphological analysis, part-of-speech disambiguation, and syntactic analysis.

In this section we focus on the grid aspects of enabling a distributed infrastructure for corpus processing, including the establishment of the virtual organisation, rights and metadata management and corpus storage and processing.

Establishment of the Virtual Organization

Creation of Core Services. To support the Human Language Technologies Virtual Organization (HLT VO), we need to prepare a dedicated server that will run VOMS (Virtual Organization Membership Service). This is the central server for the VO user and server access control, including accreditation, authentication and authorization. To support a fully functioning VO, a number of additional services will be needed, such as site and job monitoring web services (developed as part of Slovenian National Grid Initiative) and a user front-end (both web and command-line) for Virtual Organization's users.

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. We are planning to support two different grid middleware solutions: NorduGrid and gLite. NorduGrid ARC is simpler and very efficient; it 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.

For NorduGrid ARC, we have established contacts with NorduNet to arrange for registration of HLT VO with relevant databases and services. Sites that already support ARC should be able to start supporting the new VO simply by editing the relevant setup files and installing the software base for the VO from its repository.

For gLite, the Jožef Stefan Institute (JSI) can, as a member of the project, register the new VO via the Central European Regional Operations Centre (CE ROC) and include it in central EGEE databases and service monitors.

Initial Web Page. A dedicated web site for information, documentation and user management of HTL VO will be set up at JSI as part of Slovenian National Grid Initiative effort. With the addition of basic job reporting, statistics and usage services, the central infrastructure will be sufficient for initial testing and evaluation for Human Language Technologies Grid.

In order to support full virtual organization usage, additional services will have to be then developed: web-based grid tasks with automatic job submission and control, data-set upload (including corpus upload, transformation, etc.) and data retrieval from finished jobs.

We are also planning to add some web-based interfaces to the resources incorporated in the grid. 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, we are planning to gridify the lemmatization and tagger tool totale. The tool has been extensively tested with Text Encoding Initiative Guidelines, TEI P5 (Sperberg-McQueen and Burnard, 2002) encoded corpora and MULTEXT-East tag sets. In addition, we are preparing a small test suite of generic *n*-gram processing tools 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.

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 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).

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 (Demchenko et al., 2006). 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 users 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.

We believe 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, we believe we can facilitate and simplify the process without jeopardizing the security of data and copyright protections in question.

Resource Catalogue Creation. The main resources for linguistic research and lexicography, electronic corpora, represent the basic resources that should be available in the HTL VO grid.

Compared to most other scientific disciplines, corpus metadata is rather complex. In the project, we plan to evaluate the level of detail needed to create a useful central metadata server with dataset selection, searching and extraction capabilities.

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 – we need to establish a system of data pools and replica servers 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.

Grid Enabled Corpus Processing for Digital Lexicography

This section details a particular scenario for enabling grid-based corpus processing, which could be, at least partially, implementable in the scope of the MondiLex project. The scenario concentrates on two main tasks: annotation of corpora, and querying of corpora. In this it is very similar to the functioning of the Sketch Engine (Kilgarriff et al., 2004), where such operation are also supported, albeit in a non-grid (and commercial) environment.

The scenario list below is composed as a sequence of tasks, where the preceding ones provide the environment for the latter ones. We first list the tasks and then explain them in more detail:

1. implement grid-totale
2. web interface for corpus processing
3. up-loaded corpus registry with metadata
4. grid concordancer
5. concordancer corpus extension with up-loaded files
6. text statistics over uploaded corpora: keyness, terms
7. access management

Corpus Annotation with Totale

We propose for the first prototype processing pipeline to implement in the scope of the VO the corpus annotation tool "totale" (Erjavec et al., 2005), which implements the following annotation steps, in a multilingual setting:

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

The program, written in Perl, implements a simple pipelined architecture, where plain Unicode (UTF-8) text is first tokenised, the word tokens (word-forms) are then tagged with their context-disambiguated part-of-speech, or, more accurately, morphosyntactic description (MSD), and the word-forms, given their MSD, are lemmatised to arrive at the canonical form of the word. The program can produce the output in several formats, in particular in tabular form or encoded in TEI-compliant XML.

The program is – once started – reasonably fast, i.e. it processes cca 100k words per minute. Starting time, however, is a problem. Partially this is to do with the system architecture of file-mediated sequential processing, and is partially due to the lemmatisation module for a language (with its possibly thousands of rules and exceptions) being loaded statically at the start of the program. In the rest of this section we explain the three annotation modules of totale.

The multilingual tokenisation module mlToken is written in Perl, and, in addition to splitting the text input string into tokens has also the following features:

- It assigns to each token its token type. The types distinguish not only between words and punctuation marks but also mark digits, abbreviations, left and right splits (i.e. clitics, e.g. *'s* , enumeration tokens (e.g. *a*) as well as URLs and email addresses
- It marks end of paragraphs, and end of sentence punctuation, where sentence internal periods are distinguished from sentence final ones.
- It preserves (subject to a flag) the inter-word spacing of the original document, so that the input can be reconstituted from the output – this consideration is important when several tokenisers are applied to a text, either for evaluation or production purposes.

mlToken stores the language dependent features in resource files, in particular a list of abbreviations and split/merge patterns.

In the absence of a certain language resource, the tokeniser uses default resource files – in order to achieve best results, however, resource files for a specific language have to be written – this task is helped by having pre-tokenised corpora for the language available.

For tagging words in the text with their context disambiguated morphosyntactic annotations we used a third-party tagger, namely TnT (Brants, 2000), a fast and robust tri-gram tagger. TnT is freely available (but distributed only in compiled code for Linux), has an unknown-word guessing module, and is able to accommodate the large morphosyntactic tagsets that we find in various EU languages.

The tagger uses two resources, namely a lexicon giving the weighed ambiguity class for each word and a table of trigrams of tags with weights assigned to the uni-, bi-, and tri-grams.

Both resources are acquired from a correctly annotated corpus, where the induced lexicon can of course also be further upgraded.

For our lemmatiser we used CLOG (Manandhar et al., 1998, Erjavec and Džeroski, 2004), which implements a machine learning approach to the automatic lemmatisation of (unknown) words. CLOG learns on the basis of input examples (pairs word-form/lemma, where each MDS is learnt separately) a first-order decision list, essentially a sequence of if-then-else clauses, where the defined operation is

string concatenation. The learnt structures are Prolog programs, but in order to minimise interface issues we made a converter from the Prolog program into one in Perl. In the final instance the usage for determining the lemma is simply the result of the function call $lemma = lemmatise($msd, $wordform); This function then calls the appropriate rule-set, which transforms the input wordform into its lemma.

The main feature of totale is the fact that it is multilingual and trainable for new languages, as the models for tagging and lemmatisation are induced from data. However, in order to make the tool useful, we first have to obtain such data, namely morphosyntactically annotated corpora and lexicons. It is an additional advantage if the multilingual training resources all follow the same guidelines for tagset and corpus annotation design.

The MULTEXT-East language resources, a multilingual dataset for language engineering research and development, first developed in the scope of the EU MULTEXT-East project, have now already reached the 3rd edition (Erjavec, 2004). MULTEXT-East is a freely available standardised (XML/TEI P4, (Sperberg-McQueen, and Burnard, 2002)) and linked set of resources, and covers a large number of mainly Central and Eastern European languages. It includes the EAGLES-based morphosyntactic specifications, defining the features that describe word-level syntactic annotations; medium scale morphosyntactic lexicons; and annotated parallel, comparable, and speech corpora. The most important component is the linguistically annotated corpus consisting of G. Orwell's novel "1984" in the English original and translations.

For training totale we used resources for the Czech, English, Estonian, Hungarian, Romanian, Serbian, and Slovene. The MULTEXT-East mtseg resource files were used as sources for the mlToken resource files; the annotated corpus for training the TnT tagger; and the lexicons to improve the performance of the tagger and for training the CLOG lemmatiser.

While training the tagger on this data is very fast, training the lemmatiser is much more process intensive, as each MSD is learned separately – so, for Slovene or Czech, this meant learning more than 1000 different classes for a language, and the training time is measured in days.

In the scope of this task, totale must be implemented in a grid environment. As the tool is able to operate on separate texts, there is no problem with parallelisation: the input corpus files can be split into chunks, and passed on to totale. The optimum split size, however, needs to be determined experimentally, trading start-up with running time.

Web Interface for Corpus Processing

In order to be able to process corpora with totale (or other programs) in a flexible manner, it will be necessary to develop a Web front-end to enable up-load of corpora, specifying the parameters of the annotation process (e.g. language, type of annotation), running the annotation process and downloading the results.

We have already implemented, albeit in a non-grid environment, and interactive Web based system for up-loading documents, processing them, and returning the results (Erjavec, 2007). The system supports uploading of compressed archives, as well as downloading of compressed files, a necessary requirement for transfers of large corpus files. The system also supports conversion to and from XML/TEI encoded files.

The system can also generate Excel (XML) documents which then serve as input to the editing process, and are uploaded to the server after they have been corrected. This option is interesting where the automatically generated annotations need to be manually corrected.

The implemented web service runs under Linux/Apache, using CGI/Perl. The Perl script:

1. takes the uploaded file, possibly compressed, with the archive containing multiple files;
2. calls various transformations with user-selected parameters;
3. returns the result, either directly via HTTP or as an archive file; and
4. logs each transaction, possibly archiving the input and output files.

The developed system needs to be upgraded to a grid-based environment, where the main task is the distribution of the input files to appropriate servers and then collection of the processed files. As an initial option, the uploaded archive could be decompressed, and the input files distributed among servers.

Of course, for testing purposes a window where text can be pasted as well as uploading a single text file should also be supported.

The download scenario is similar to the upload one – for small files a window where annotated text is displayed could be offered. For larger files and corpora, a compressed archive should be offered.

It is recommended that the default encoding of the annotated corpora is XML/TEI, however, a tabular file format is also an option. The tabular file contains in each line either a structural tag (such as or) or a corpus token, either a punctuation symbol or word. The word tokens are then, at least in totale, annotated with their context disambiguated MSD and lemma.

Corpus Registry

While in the initial stage it is sufficient to simply offer the processed corpora for download, in the longer term it would be interesting to explore the possibility of storing the up-loaded and annotated corpora in the grid environment, either permanently or with an expiration date. This would enable other users – of course, with appropriate authorisation – to also access interesting corpora. Additionally, the VO could also offer "pre-cooked" corpora, which are simply loaded into the corpus pool, to be used as required.

For all these tasks, a corpus registry needs to be established, which stores the name of the corpus, along with appropriate meta-data, e.g.:

- language of the corpus
- annotations of the corpus
- responsible person
- access rights to the corpus

For the basic meta-data storage it is suggested to use the TEI header, as it can store all the required information about the corpus. It is also simple to write a XSLT transform to convert the TEI header into HTML, which can be then displayed to the users, and to a database backend used to track access rights and permissions. For simple search and display tasks, it is also appropriate to enable a conversion from basic elements of the TEI header to a Dublin Core record, which should be sufficient for most requirements.

The most complicated aspect of the registry is its connection to other pieces of software, e.g. the concordancer, as the software must be aware of access rights of particular users and prevent access to restricted data for users that do not possess sufficient access rights. To manage this process, a central system for establishing the required agreements and keeping track of current status for all users will need to be created.

Grid Concordancer

If so far the discussion has concentrated on the process of corpus compilation, but the corpora should also be usable by the lexicographers directly in the grid environment. Concordancers are the basic tool for corpus exploration; in the context of the grid, we have to look at Linux based, Web enabled concordancing engines. Currently, there are two main candidates on offer. The first is CWB, the IMS Corpus Work Bench (Christ 1994) developed at Institut für Maschinelle Sprachverarbeitung at the University of Stuttgart, which is now also freely available as source. The CWB has a very rich query language, supports annotated corpora, and is fast over large corpora. The main problem with CWB is that it does not support Unicode, so the corpora have to be encoded in one of the 8-bit encodings. While this is acceptable for monolingual corpora of (contemporary) European languages, it presents problems with multilingual corpora and historical corpora. CWB is also only the back-end engine, which means that it is still necessary to write an appropriate Web interface to enable simple access to the corpus. Various front-ends exist, e.g. at JSI or the collective effort named CSAR, but none of them is very well documented and none of them can use the full potential of the functionality that CWB offers.

An alternative back-end is Manatee (Rychlý and Smrž, 2004), which was deliberately made CWB compliant, although it is a complete re-implementation. The advantage of Manatee is that it is Unicode aware; however, it does not support parallel corpora. There has been less work in developing front-ends for Manatee (although a client, called Bonito does exist), but one of them was developed by the Slovak partner of MondiLex.

In short, either CWB or Manatee would need to be implemented to work in a grid environment. The simplest and the most effective scenario would be to store particular corpora on a predefined machine in the grid, run the corpus query locally on the machine, and return the results to the front-end, where they are each displayed separately. Such an approach has been proposed by (Tamburini, 2004), and has the advantage of simplicity, as the front end needs only to know which corpora are to be queried, initiate the query processing over them, and then trivially assemble the results. More complex scenarios, whereby a single corpus would be split and distributed in different storage pools in advance, and the results seamlessly joined by the front-end are probably not feasible in the current time-frame. However, a system using complete replicas and metadata server lookup to locate them would not represent a significant complication while contributing to availability and responsiveness of the system.

Concordancer Corpus Extension with Up-loaded Files

The preceding section assumed that the corpus files that the concordancer is aware of are static, and have been added to the concordancer by the VO maintainer. In the longer perspective it is worth exploring the option whereby up-loaded and annotated corpora are also made automatically available for concordancing. This involves an extra processing step, whereby the new corpora are added to the concordancer registry and indexed, and expired corpora are removed from the concordancer store.

Text Statistics

While the functionality of concordancers is the basic one to offer to lexicographers, there are other processing tasks that are also very useful, and typically involve processing over complete corpora. As an initial offering, we plan to prepare a small test suite of generic *n*-gram processing tools for statistical analyses on available distributed data. Since *n*-gram processing of large datasets, especially for *n*-grams where $n > 3$, can be relatively resource-intensive, we are expecting a noticeable improvement, especially for tasks where development or refinement of algorithms requires several iterations of reprocessing of the same dataset to test the hypotheses and implementations on suitably large datasets.

Here we note two other tasks, keyness and term extraction. Keyness, implemented in e.g. the WordSmith tools (Scott, 2004), shows which words are particularly salient for a text, and presupposes a reference corpus, giving the frequency of general usage for particular words, and the target text, from which a lexicon is extracted, and compared to general usage. In this manner, the "keywords" of a particular text are identified, and can serve as a first approximation of interesting lexical entries, e.g. for a particular terminological domain. In a similar fashion, terms can be extracted from text, using either purely statistic methods or prepared templates of syntactic combinations, e.g. "Adjective + Noun". These methods, unlike concordances, concentrate on the lexis of particular texts or corpora, and identify interesting words or phrases for subsequent investigations. These kinds of tools are also a good candidate for gridification, as they typically need to process the entire corpus or corpora, and are thus computationally expensive.

Conclusion

The paper presented why digital lexicography could benefit from grid-enabled processing and sketched a number of requirements that need to be met in order for this vision to become a reality. In particular, we gave an overview of grid services that would be necessary for corpus processing to be implementable on the grid, and then sketched a number of services, mostly to do with corpus compilation and analysis, that could be implemented relatively easily, yet still be useful for practical work. In particular, we outlined the measures necessary for enabling the grid infrastructure for digital lexicography and then discussed a scenario that would enable corpus processing on the grid.

We are currently attempting to implement the vision sketched in this paper, and the reality of it will doubtless bring new insights into the usefulness of the grid for digital lexicography.

References

Brants, T. TnT-A Statistical Part-of-Speech Tagger. (2002). In Proceedings of the 6th Applied Natural Language Processing Conference ANLP-2000 (pp. 224-231). Seattle, WA.

Carroll, J., Evans, R., Klein, E. (2005). Supporting text mining for e-science: the challenges for grid-enabled natural language processing. In: Proceedings of the UK e-Science All Hands Meeting.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. COMPLEX'94, Budapest, 1994.

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

Ellert, M., et al., (2007). Advanced Resource Connector middleware for lightweight computational Grids. Future Generation Computer Systems 23, pp. 219-240.

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

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

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

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. Fourth International Conference on Language Resources and Evaluation, LREC'04. (pp. 1535-1538). ELRA, Paris.

Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). The Sketch Engine. Proceedings of EURALEX Lorient, France.

Luís, T., Martins de Matos, D., Paulo, S., Daniel Ribeiro, R., (2008). Natural Language Engineering on a Computational Grid (NLE-GRID) T5 - Performance Experiments, Tech. Rep. 35 / 2008 INESC-ID Lisboa, January 2008.

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) (pp. 135-144). Springer-Verlag, Berlin.

Martins de Matos, D., Tiago Luís, Daniel Ribeiro, R. (2008a). Natural Language Engineering on a Computational Grid (NLE-GRID) T1 - Architectural Model, Tech. Rep. 30 / 2008 INESC-ID Lisboa, January 2008.

Martins de Matos, D., Daniel Ribeiro, R., Paulo, S., Batista, F. Coheur, L., Paulo Pardal, J. (2008b). Natural Language Engineering on a Computational Grid (NLE-GRID) T2 - Encapsulation of Reusable Components, Tech. Rep. 31 / 2008 INESC-ID Lisboa, January 2008.

Martins de Matos, D., Daniel Ribeiro, R. (2008c). Natural Language Engineering on a Computational Grid (NLE-GRID) T2h - Encapsulation of Reusable Components: Lexicon Repository and Server, Tech. Rep. 32 / 2008 INESC-ID Lisboa, January 2008.

Marujo, L. Lin, W. Martins de Matos, D. (2008). Natural Language Engineering on a Computational Grid (NLE-GRID) T3 - Multi-Component Application Builder, Tech. Rep. 33 / 2008 INESC-ID Lisboa, January 2008.

Neuroth, H., Kerzel, M., Gentzsch, W. (eds.), (2007). German Grid Initiative D-Grid.

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

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

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

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

Web References

Language processing tools:

- Sketch Engine: http://www.sketchengine.co.uk/
- IMS Corpus Work Bench: http://cwb.sourceforge.net/
- CSAR interface to CWB: http://csar.sourceforge.net/
- JSI corpus front-end: http://nl2.ijs.si/
- Manatee Corpus Processor: http://www.textforge.cz/
- WordSmith Tools: http://www.lexically.net/wordsmith/
- TEI consortium: http://www.tei-c.org/

Grid Web references:

- Globus Toolkit: http://www.globus.org/toolkit/
- EGEE/LCG Technical Report: http://lcg.web.cern.ch/LCG/tdr/
- LCG Progress report: Q1, Q2 2005: http://lcg.web.cern.ch/LCG/PEB/Documens/LCG-ProgressReport-01Q05.pdf, http://lcg.web.cern.ch/LCG/PEB/Documens/LCG-ProgressReport1-02Q05_02aug05.pdf
- Central European Operations Centre EGEE (CE-ROC): http://grid.cyfronet.pl/egee/
- CE Federation EGEE: http://egee-intranet.web.cern.ch/egee-intranet/federations/central.html
- Achievements of projects collaborating with EGEE: http://egee-na2.web.cern.ch/egee-NA2/files/material/rpbooklet-final-for-print.pdf
- Trust on the Grid Goes Global (establishment of IGTF): http://www.eugridpma.org/igtf/igtf-newsrelease-20051005.pdf
- EUGridPMA: http://www.eugridpma.org/
- SiGNET CA web: http://signet-ca.ijs.si/
- EGEE in Slovenia: http://www-f9.ijs.si/
- European Grid Initiative Partners: http://web.eu-egi.eu/partners/ngi/

On Compatibility of Slavic Language Resources[1]

Ludmila Dimitrova, Radoslav Pavlov
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences, Sofia
ludmila@cc.bas.bg, radko@cc.bas.bg

Abstract
We describe in brief what grid technologies are and how they could contribute to the language technologies, in particular lexicographic activities. Based on our participation in the EC international project MULTEXT-East, we present some aspects of language resource compatibility: unification and standardisation. We underline the importance of the developed harmonised lexical (morphosyntactic) specifications and descriptions of language data in machine-readable form in a common standard encoding format – Corpus Encoding Standard format – for six Central and East European (CEE) languages, as well as the language-independence of the tools employed.

Keywords: language technologies, language resources, grid technologies, electronic corpora, lexicon and dictionaries

Introduction

Applications of language technologies (or natural language processing) have recently been extended in the areas of information research, machine translation, machine learning, speech technology, lexicography, terminological bank servicing, etc.

In a situation of extended applications, language technologies are provoked by new technological decisions (tools) that the information technologies offer recently. A grid, or more precisely, a knowledge grid is such a decision.

What are grid technologies and how could they contribute to the language technologies in particular lexicographic activities?

A grid is a network or collection of distributed computer resources, which are accessible through local or global networks and are presented to the users via an enormous virtual computer system. In a nutshell, a grid is a virtual, dynamically changing organization of structured resources, which are shared among individuals, institutions or systems. Some of the main advantages of the grid technology are: virtual organisation of digital resources; optimized access and enhanced management of these resources; ability to be used worldwide, etc. Knowledge grids offer high-level approaches, techniques and tools for distributed mining and extraction of knowledge from data, processing and accessing of data from the repositories available on the grid, leveraging semantic descriptions of components and data. These functions allow scientists and professionals to compose workflows that integrate data sets and store them, and to create and manage complex knowledge applications. A knowledge grid uses knowledge-based methodologies and technologies to answer much harder questions and to find the appropriate answers in the required form. It joins technologies for data mining, ontologies, intelligent portals, workflow reasoning, etc., for supporting the way knowledge is acquired, used, retrieved, published and maintained.

The relationships between the described features of knowledge grids and lexicographic activities can be briefly formulated as follows:

- Typical knowledge grid objects and language technologies objects (for example, electronic dictionaries and corpora) share some specifications, like: the structural complexity of mono-, bi- and multilingual dictionaries, the great volume of the dictionaries, the internal structure of the dictionaries as a sequence of well-defined tagged-tree lexical entries, etc.
- The knowledge 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 (based on open

[1] The study and preparation of these results have received funding from the EC's Seventh Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX.

standards and software tools) and openness.

- The knowledge grid allows 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 knowledge grid technology could provide for the creation of a general lexical database with a rich system of links between forms and meanings of the words; the users could search in any language that already has a digital dictionary.
- The knowledge grid provides infrastructure for the creation and support of a network of high-quality multi-language resources. Many digital lexicographic resources, developed by different research groups or scientists, could be active on the same shared knowledge grid resources at the same time. The research groups could create in collaboration, regardless of distance and time, new digital lexicographic resources that could meet the requirements of the current information space.
- The lexicographic resources (file archives or databases) can be of very different nature, but they must have a standard description, be presented in a standard form in order to be used by standard software tools. This means that the knowledge grid-based infrastructure will support and manage a network of shared resources (e.g., archives, repositories, database, and software tools).
- The high power and secure services of knowledge grid will provide computational techniques for solving some digital lexicographic problems, such as interoperability, ontology integration, content-based automatic selection, automatic content source description, resources preservation, etc.

Annotated electronic Bulgarian language resources

The first annotated electronic resources for Bulgarian language were developed during the EC project MULTEXT-East *Multilingual Text Tools and Corpora for Central and Eastern European Languages* (Dimitrova et al. 1998). Here we give a brief account of our work, as participants in this EC project. We believe that the programming tools used in the MULTEXT-East project (MTE for short) and multi-language resources developed represent a good sample of a **research infrastructure**.

The MULTEXT-East is a continuation of MULTEXT project under the INCO-Copernicus programme. Project MULTEXT produced the language resources and a freely available set of tools that is extensible, coherent, and language independent, for six western European languages (English, French, Dutch, Italian, German, and Spanish) (Ide, Veronis 1994).

MTE project developed significant language resources for six Central and Eastern European (CEE) languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene, and for English. Three of these languages belong to the Slavic language group: Bulgarian, Czech, and Slovene. The MTE electronic lexical resources include **multilingual MTE corpus**, produced as a well-structured and lemmatized *CES-corpus:* in Corpus Encoding Standard (CES) (Ide 1998, http://www.cs.vassar.edu/CES/), and a dataset of **language specific resources** (for Bulgarian see Dimitrova 1998, Dimitrova et al. 2005).

The results of the two multi-language projects MULTEXT and MULTEXT-East – as resources and experience of using the same program tools – show how important the development of **common harmonised lexical specifications** in *CES-format* for different European languages and the **language-independence of the tools** employed are.

MTE multilingual corpus comprises three corpora: *parallel corpus*, based on George Orwell's novel "1984", *comparable corpus* – newspaper excerpts and texts from CEE literatures, and a small *speech corpus*. The texts of the parallel corpus have been produced as well-structured, lemmatized documents in CES-format.

The language specific resources that MTE project developed are:

- Lexical (morphosyntactic) specifications;
- Language-specific data;
- Lexicons.

The texts and the lexicons produced serve as input data for experiments with the tools created for processing Western-European languages in MULTEXT, but also serve as resources for building

lexical databases for the six CEE languages (EC project CONCEDE). The MULTEXT tools were implemented under UNIX. They could be distributed in two main types: corpus annotation tools and corpus exploitation tools – *segmenter, morphological analyser, part-of-speech disambiguator, aligner, etc*. All tools are integrated via a common user interface into a general-purpose manipulation system suitable for natural language processing research. The MULTEXT tools were designed with an engine-based approach where all language-dependent materials are provided as data (in a form of the tables or rules).

MTE language specific resources

In 1995, the Text Encoding Initiative (TEI), an international project, aimed to produce a guide for preparation and exchange of electronic texts for scientific purposes, using the standards for text representation (http://www.tei-c.org/Guidelines/P5/get_p5.xml).

The TEI-group chose SGML, a metalanguage defined in 1986 with an international standard, ISO 8879, because of its important application to language engineering. SGML makes a text available to many different types of processing or using, because it defines the document's contents entirely and independently of the language. Such text serves as a reusable resource for the purposes of many multilingual systems. The TEI-conformant mark-up techniques (SGML), ensures the efficiency of the electronic exchange of information, large corpora, and lingualware between the scientists of linguistic research. The SGML technique was applied to create language specific resources for the 6th CEE language of the project. The MULTEXT methodology (harmonization of the resources and usage of common tools), used in MTE, has provided producing portable uniform SGML multilingual resources.

The resources for texts processing developed in MTE project are language-specific data. These resources are files required by the MULTEXT project tools for segmentation, tagging, and disambiguation. In this way the language independence of the tools was provided. Each partner has developed a set of resource files for their language and a lexicon according to the common specifications in the MULTEXT format.

MTE lexical specifications

The MTE languages use different character sets and the originals of texts contain symbols not present in ASCII. All MTE electronic texts use 8-bit encoding defined in one of the ISO 8859 standards: Bulgarian uses ISO 8859-5 (Cyrillic), Czech, Hungarian, and Romanian, for example – ISO 8859-2 (Latin 2). The free word order and rich inflection of CEE languages (especially three Slavic: Bulgarian, Czech, and Slovene) presented significantly different linguistic problems than do those of Western Europe. For a description of specific languages phenomena, MULTEXT's specifications were enlarged by an addition of new attributes and values for each Central and Eastern European (CEE) language. So at its first phase the MTE project has developed **harmonised** lexical specifications for six CEE languages and for English (Ide, Veronis, Erjavec (Eds.) 1997).

The specifications are presented as sets of attribute-values, with their corresponding codes used to mark them in the lexicons. The features that are shared by all MTE languages (so-called **core features**) were determined. In such manner the **comparability** of the information encoded in the lexicons across the MTE languages was provided. Except these "general properties" the so-called language-specific features were defined, which describe language-specific morphosyntactic phenomena.

Language-specific data

Sets of segmentation and morphological rules and data for use with the various annotation tools were developed. Segmentation rules describe the form of sentence boundaries, quotations, numbers, punctuation, capitalization, etc. Morphological rules, needed by the morphological tools, provide exhaustive treatment of inflection and minimal derivation. Other language-specific data, the so-called special tokens, required by the segmenter, includes lists of special tokens (frequent abbreviations and names, titles, patterns for proper names, etc.) with their types. For maximum flexibility and to retain **language independence**, all such information is provided directly to the subtools via **external resource files,** for example, Bulgarian external files are: tbl.punct.Bg, tbl.abbrev.Bg, tbl.compound.Bg, etc.

MTE lexicons

The MTE lexicons have the **standard** form of the MULTEXT lexica.

Each lexicon entry includes the following information: inflected form; lemma; morphological information for this inflected form encoded in its morphosyntactic description:

wordform <TAB> lemma <TAB> MSD

Examples of Bulgarian lexicon entry:

май	**=**	**Qgs**	**(Particle, general, simple)**
май	**мая**	**Vmm-2s**	**(Verb, main, imperative, 2nd person, singular)**

In fact, Bulgarian MTE lexicons are three and mostly cover the available texts (Dimitrova et al. 2005):

1. Bulgarian translation of G. Orwell's "1984";
2. Bulgarian corpora
   2.1. *Fiction* (two novels)
   2.2. *Newspaper* (excerpts).

The lexicon of Bulgarian translation of G. Orwell's "1984" contains 17567 lemmas and 295431 word-forms for these lemmas.

The table below shows a number of lemmas and word forms in the Bulgarian lexicon:

Part of Speech	Lemmas	Entries
Nouns	9891	47969
(masculine	4180	25100)
(feminine	4120	16493)
(neuter	1591	6376)
Verbs	4140	226666
Adjectives	2155	19397
Pronouns	92	110
Adverbs	790	790
Adpositions	98	98
Conjunctions	76	76
Numerals	67	67
Interjections	172	172
Particles	86	86
Total	17567	295431

The MTE results

The MULTEXT-East project developed three multilingual corpora:

(1) Parallel Corpus,
(2) Comparable Corpus,
(3) Speech Corpus.

There are **four versions** of **MTE parallel corpus**, corresponding to four different levels of annotation.

For Bulgarian these versions (differently encoded documents) are:

- **Original text –** Bulgarian translation of G. Orwell's novel "1984", includes 86020 words (lexical items, excluding punctuation), 101173 tokens (words and punctuations);
- **CesDOC-encoding** of the Bulgarian text of the novel (SGML mark-up of the text up to the sentence-level), includes 1322 paragraphs, 6682 sentences;
- **CesANA-encoding**, containing word-level morpho-syntactic mark-up (undisambiguated lexical information for 156002 words, 156002 occurrences of MSD, and disambiguated lexical information for the 86020 words of the novel);
- **CesAlign-encoding:** Bulgarian-English aligned texts, containing links to the aligned sentences.

The software tools, with which the below-mentioned encoded documents were carried out, were developed within the MULTEXT project, but the data input came from MTE language-specific resources.

To arrive at the tokenised and tagged document (for example, G. Orwell's "1984" in Bulgarian) the following steps have been performed:

1. cesDoc version has been simplified and converted to cesAna encoding;
2. the text (the result of step 1) was tokenized;
3. the tokens (the result of step 2) were annotated with lexical (ambiguous MSDs) lemmas and tags;
4. lexical information was disambiguated.

At first, the Bulgarian translation of G. Orwell's "1984" was segmented by means of the segmenter MTSeg – a tokenizer. The segmenter MTSeg is a language-independent and configurable processor used to tokenize input text, given in one of the three possible formats: plain text, a normalized SGML form (nSGML) as output by another MULTEXT tool (MTSgmlQl), or a tabular format (also specific to MULTEXT processing chain). The output of the segmenter is a tokenized form of the input text, with paragraph and sentence boundaries marked-up. Punctuation, lexical items, numbers and several alphanumeric sequences (such as dates and hours) are annotated with various tags out of a hierarchy class structured tag set. The language specific behavior of the segmenter is driven by several language resources (abbreviations, compounds, split words, etc.), incl. segmentation rules and special tokens.

To explain the structure of the final documents, first consider a fragment of the **Bulgarian cesDoc Orwell**:

```
<p id="Obg.1.1.2">
  ...
   <s id="Obg.1.1.2.10">Портретът бе нарисуван така, че очите да те следват, накъдето и да се обърнеш. </s>
  ...
 </p>
```

At the S (Sentence) level the documents have been tokenised according the lexical resources of the language and are encoded as TOKen elements. Tokens are either "normal" words, compounds, separable parts of words ("clitics"), or punctuation marks. They are distinguished by the value of the token's TYPE attribute. WORD is the values used for words, and PUNCT for punctuation marks. The word or punctuation mark is contained in the ORTH element. The punctuation tokens are annotated with (unambiguous) corpus tags, which identical across the languages of MULTEXT-East.
The following example illustrates this markup:

```
<par from="Obg.1.1.1">
   <s from="Obg.1.1.2.10">
    <tok type=WORD><orth>Портретът</orth></tok>
    <tok type=WORD><orth>бе</orth></tok>
    <tok type=WORD><orth>нарисуван</orth></tok>
    <tok type=WORD><orth> така </orth></tok>
    <tok type=PUNCT><orth>,</orth><ctag>COMMA</ctag></tok>
    <tok type=WORD><orth>че</orth></tok>
    <tok type=WORD><orth>очите</orth></tok>
    <tok type=WORD><orth>да</orth></tok>
    <tok type=WORD><orth>те</orth></tok>
    <tok type=WORD><orth>следват</orth></tok>
    <tok type=PUNCT><orth>,</orth><ctag>COMMA</ctag></tok>
    <tok type=WORD><orth>накъдето</orth></tok>
    <tok type=WORD><orth>и</orth></tok>
    <tok type=WORD><orth>да</orth></tok>
    <tok type=WORD><orth>се</orth></tok>
    <tok type=WORD><orth>обърнеш</orth></tok>
    <tok type=PUNCT><orth>.</orth><ctag>PERIOD</ctag></ctag>
    </tok>
   </s>
```

When the **input text was segmented**, the next tool – **MTLex** (from MULTEXT tools) – was used: a dictionary look-up procedure assigns to each lexical token all its possible morpho-syntactic descriptors (MSDs). Corresponding lines for morphosyntactic annotation of the Bulgarian phrase "портретът бе" in output of MTLex (in English "picture was" – from the tenth sentence of the "1984": "It was one of those pictures which are so contrived that the eyes follow you about when you move.") are:

1.1.2.10\1 TOK Портретът портрет\Ncmsf\NCMS-F
1.1.2.10\11' TOK бе бе\Qgs\QGS|съм\Vaia2s\VAIA2S|съм\Vaia3s\VAIA3S

At the next step **the text was tokenized**. The **word tokens are annotated both with ambiguous lexical information** (in the <lex> elements of the token), and with disambiguated, context-dependent, information (in the <disamb> element(s)). Both elements contain the <base> (lemma) of the token, its morphosyntactic description <msd>, and its language depended corpus tag – <ctag> – as illustrated in the following example, the tenth sentence of the Bulgarian translation of "1984" – *Портретът бе нарисуван така, че очите да те следват, накъдето и да се обърнеш.* (In English: *It was one of those pictures which are so contrived that the eyes follow you about when you move.*).

```
<par from='Obg.1.1.1'>
……..
<s from='Obg.1.1.2.10'>
   <tok type=WORD from='Obg.1.1.2.10\1'>
    <orth>Портретът</orth>
    <disamb><base>портрет</base><ctag>NCMS-F</ctag></disamb>
    <lex><base>портрет</base><msd>Ncms-f</msd><ctag>NCMS-F</ctag></lex>
   </tok>
   <tok type=WORD from='Obg.1.1.2.10\11'>
    <orth>бе</orth>
    <disamb><base>съм</base><ctag>VAIA3S</ctag></disamb>
    <lex><base>бе</base><msd>Qgs</msd><ctag>QG</ctag></lex>
    <lex><base>съм</base><msd>Vaia2s</msd><ctag>VAIA2S</ctag></lex>
    <lex><base>съм</base><msd>Vaia3s</msd><ctag>VAIA3S</ctag></lex>
   </tok>
   <tok type=WORD from='Obg.1.1.2.10\14'>
    <orth>нарисуван</orth>
    <disamb><base>нарисувам</base><ctag>VMPS-SM</ctag></disamb>
    <lex><base>нарисувам</base><msd>Vmps-smp-n</msd><ctag>VMPS-SM</ctag></lex>
   </tok>
   <tok type=WORD from='Obg.1.1.2.10\24'>
    <orth>така</orth>
    <disamb><base>така</base><ctag>QG</ctag></disamb>
    <lex><base>така</base><msd>Qgs</msd><ctag>QG</ctag></lex>
    <lex><base>така</base><msd>Rg</msd><ctag>RG</ctag></lex>
   </tok>
   <tok type=PUNCT from='Obg.1.1.2.10\28'>
    <orth>,</orth>
    <ctag>COMMA</ctag>
   </tok>
   <tok type=WORD from='Obg.1.1.2.10\30'>
    <orth>&chcy;&iecy;</orth>
    <disamb><base>че</base><ctag>QG</ctag></disamb>
    <lex><base>че</base><msd>Ccs</msd><ctag>CC</ctag></lex>
    <lex><base>че</base><msd>Css</msd><ctag>CS</ctag></lex>
    <lex><base>че</base><msd>Qgs</msd><ctag>QG</ctag></lex>
   </tok>
```

```
<tok type=WORD from='Obg.1.1.2.10\33'>
  <orth>очите</orth>
  <disamb><base>око</base><ctag>NCNP-Y</ctag></disamb>
  <lex><base>око</base><msd>Ncnp-y</msd><ctag>NCNP-Y</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.2.10\39'>
  <orth>да</orth>
  <disamb><base>да</base><ctag>QV</ctag></disamb>
  <lex><base>да</base><msd>Ccs</msd><ctag>CC</ctag></lex>
  <lex><base>да</base><msd>Qgs</msd><ctag>QG</ctag></lex>
  <lex><base>да</base><msd>Qvs</msd><ctag>QV</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.2.10\42'>
  <orth>те</orth>
  <disamb><base>ти</base><ctag>PP2</ctag></disamb>
  <lex><base>ти</base><msd>Pp2-sa--y</msd><ctag>PP2</ctag></lex>
  <lex><base>те</base><msd>Pp3-pn</msd><ctag>PP3</ctag></lex>
  <lex><base>те</base><msd>Qgs</msd><ctag>QG</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.2.10\45'>
  <orth>следват</orth>
  <disamb><base>следвам</base><ctag>VMIP3P</ctag></disamb>
  <lex><base>следвам</base><msd>Vmip3p</msd><ctag>VMIP3P</ctag></lex>
</tok>
<tok type=PUNCT from='Obg.1.1.2.10\52'>
 <orth>,</orth>
 <ctag>COMMA</ctag>
</tok>
<tok type=WORD class=COMP from='Obg.1.1.2.10\55'>
  <orth>накъдето_и_да</orth>
  <disamb><base>накъдето_и_да</base><ctag>RG</ctag></disamb>
  <lex><base> накъдето_и_да </base><msd>Rg</msd><ctag>RG</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.2.10\69'>
  <orth>се</orth>
  <disamb><base>се</base><ctag>QV</ctag></disamb>
  <lex><base>се</base><msd>Px---a--yp</msd><ctag>PX</ctag></lex>
  <lex><base>се</base><msd>Qvs</msd><ctag>QV</ctag></lex>
</tok>
<tok type=WORD from='Obg.1.1.2.10\72'>
  <orth>обърнеш</orth>
  <disamb><base>обърна</base><ctag>VMIP2S</ctag></disamb>
  <lex><base>обърна</base><msd>Vmip2s</msd><ctag>VMIP2S</ctag></lex>
</tok>
<tok type=PUNCT from='Obg.1.1.2.10\79'>
 <orth>.</orth>
 <ctag>PERIOD</ctag>
</tok>
</s>
```

Conclusion

As the above examples show, the compatibility of digital resources in Slavic languages (corpora, lexicons, mono-, bi- and multilingual dictionaries) can be achieved through carrying out two major tasks:

- development of standardised and unified lexical descriptions for Slavic languages to annotate texts and word-forms in corpora; lexicon lines; dictionary entries, headwords, etc.,
- use of language-independent programming tools for processing of annotated in such manner language resources.

The grid technologies give us possibilities to:

- transfer and exchange tools and high-volume data (such as digital corpora and dictionaries)
- process in parallel unified data in different Slavic languages by same tools.

The usage of Slavic language resources annotated with standardised and unified lexical descriptions, and the possibilities offered by grid technologies will help linguists in their work to produce new bi- and multilingual Slavic lexical resources and to offer them to the research, education, business communities and to the wide public.

References

Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.J., Petkevic, V., Tufis, D. (1998). Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *COLING-ACL'98*, pages 315-319, Montréal, Québec, Canada.

Dimitrova L. (1998). Lexical Resource Standards and Bulgarian Language. *In International Journal Information Theories & Applications,* Vol. 5, No. 1. pp. 27-34.

Dimitrova, L, Pavlov, R, Simov, K, Sinapova L. (2005). Bulgarian MULTEXT-East Corpus – Structure and Content. In *Cybernetics and Information Technologies*. Volume 5. Number 1. pp. 67-73.

Ide, N. and Véronis, J. (1994). Multext (multilingual tools and corpora). In *COLING'94*, Kyoto, Japan, pp. 90-96.

Ide, N., Veronis, J., Erjavec, T. (Eds.) (1997). Specifications and Notation for Lexicon Encoding. MULTEXT-East Deliverable D1.1F, Institute Jozef Stefan, Ljubljana, Slovenia. http://nl.ijs.si/ME/CD/docs/mte-d11f/

Ide, N. (1998) Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In *First International Conference on Language Resources and Evaluation, LREC'98*, Granada, Spain, pp. 463-470.

http://www.tei-c.org/Guidelines/P5/get_p5.xml

http://www.cs.vassar.edu/CES/

Integral Slavic Lexicography in the Linguotechnological Context[1]

Volodymir Shyrokov
Ukrainian Lingvo-information Fund
National Academy of Science of Ukraine
vshirokov48@mail.ru

Abstract
Integral multilingual lexicography is considered in the context of the International FP7 project, Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources. The relationship between the grammar and lexicographical type of the language system description is analyzed. A tool is proposed to construct a virtual lexicographical laboratory aimed at implementing the project.
***Keywords*:** integrity, multilingual lexicography, linguistic picture of the world, language system

Preface

At the present time we witness two scientific and technical revolutions simultaneously: a digital one and the one related to communication. In a foreseable future we can expect the world of communications and digital technologies to become a mirror of the evolutional paradigm of the modern civilization – such a prospect looks more and more actual because there is a mechanism of information-energy transformations[2] making the natural-science foundation of the information society and the knowledge society. It is possible to make a certain futurological forecast: if some methods to connect the information-technological evolution of human society and the biological evolution of matter are invented, i.e. if these two lines of evolution meet at some stage (note that the contemporary development of genetic engineering, microelectronics, nano-technologies, neurophysiology and cognitive science makes this scenario even more substantiated) – then the appearance of a new form of reasonable life to integrate biological and technotronic substance in a united cognitive organism can be considered quite probable.

However, such a futurology can be upset by a specific "limitation of the integrity" of the world information system connected with the human multilinguism. Indeed the segmentation of the information space into separate language segments (and therefore, into separate pictures of the universe) is an established fact, while the contemporary technologies of interlingual communication are still very far from being perfect and are incapable to overcome the "Babylonian syndrome" caused with the simultaneous active operation of tens and hundreds of human languages that strongly differ from each other in both their constructions and information statuses. Thus, a question of interlingual adaptation rather unexpectedly takes a different turn and changes its perception in the context of the civilizational shifts of nowadays, and as a result, linguistic problems, or to be more precise, questions of the development of linguistic technology are moving from the periphery to the focal point of scientific and technical development.

From the above we can conclude the necessity of purposeful fundamental studies of the language system in order to obtain results ready for implementation into highly effective intellectual language technologies. As the applied technological aspects are the top priority for the modern linguistic studies, a lexicographical description of the language system becomes especially important. Indeed, the effectiveness of linguistic technologies depends finally on the quantitative and qualitative parameters of the lexicographic description of units, relations and levels of language those technologies are based on. And if we take into account the importance of finding a solution to the problem of multilinguism in the global information medium then we quite sensibly come up to the task of integration of the lexicographical descriptions for all the languages, i.e., the compiling of a dictionary for the entire mankind, a unique Summa lexicographiae.

[1] The study and preparation of these results have received funding from the EC's Seventh Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX.
[2] This mechanism is described in (Широков 1996). It is applied to substantiate the principles of functioning of the information society in (Широков 1996a, 1998, 2004, 2004a).

It is clear that at the moment this task looks somewhat fantastic since any complete lexicographical description does not exist even for one separate language now. However, this does not mean that such tasks – at least theoretically – should not be set at all. We are convinced that general statements of problems stimulate elaboration of general theoretical approaches, methods and concepts.

On a practical plane the given problem has arisen for us in connection with the commencement in 2008 of an international linguistic project in the seventh frame program of scientific studies by the European Union FP73, with the purpose to create a united lexicographical system for a number of Slavic languages (Bulgarian, Polish, Russian, Slovak, Slovenian and Ukrainian). We think that the following stage of the project should embrace all the Slavic languages. Thus a basic system for the All-Slavic linguistic dialogue would be founded.

1. What approaches to the solution of this problem do we consider to be realistic?

To answer this question let us turn to the experience of the Ukrainian Lingual- Information Fund, National Academy of the Sciences (NAS) of Ukraine regarding the creation of the National Lexical Base of Ukraine. In our institution during a relatively short period of time (approximately 15 years) a theory of the computer lexicographical systems (L-systems) has been developed and a series of such systems fundamental for the Ukrainian lexicography concerning their status, functionalities and exhaustiveness of the language material involved have been created. These L- systems, according to the nature of systems engineering, are instrumental, i.e. they are oriented to the support of the compiling of new lexicographical works. Among these L-systems we can mention the grammatical system of Ukrainian and Russian languages, the system for the composition of intelligent, etymological, synonymous, phraseological dictionaries and some others. The L-systems mentioned above were created by investigating the structures of the fundamental traditional dictionaries, first of all, Ukrainian4 ones, on the basis of the information theory of lexicographical systems we had developed (Широков 1996а, 1998, 2004). The application of the given theory to the study of the structures of traditional lexicographical objects enabled us not only to find new types of regularities in the Ukrainian language system5, but also to create an effective computer technology aimed at the development of large lexicographical projects (Широков 1996а, 1998, 2004, 2004а; Русанівський, Широков 2002). This application made it possible, in particular, to compile a 20-volume explanatory dictionary of the Ukrainian language actually within five years. In this case the most effective modes of use of this technology were experimentally determined. Particularly, the lexicographers work immediately with the computer system to use all the advantages of the new information-lexicographical resources (among them the own resources of the Ukrainian lingual-information fund – a corpus which amounts to more than 58 million word usages (Широков и др. 2002а) and numerous computer dictionaries) as well as the lingual resources of the Internet. The characteristic feature of systems engineering in this technology is its orientation to the functioning in the network mode which in principle permits for the dictionary compilers – linguists from different institutions or even countries – to work simultaneously at the development of common lexicographical projects. This technology was named the "Virtual Lexicographical Laboratory". At present, besides the already mentioned L-system of the Dictionary of Ukrainian Language, we should mention the following L-systems ready to operate in the mode of the Virtual Lexicographical Laboratory ([VLL]): grammatical dictionaries for a number of languages (Ukrainian, Russian, English, German, French, Spanish, Turkish and, partially, Polish), explanatory dictionaries (Russian and Turkish languages), dictionaries of synonyms (Ukrainian and Russian languages), the etymological dictionary of Ukrainian language. Furthermore, a technology of the computer-aided transfer of hard-copy lexicographical works into computer L-systems with their subsequent integration into the VLL-structures is being worked out. One should note, however, that the practical application of the VLL-ideology to implementation of

[3] Project acronym: MONDILEX. Project full title: Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources. Grant agreement no.: 211938.

[4] Dictionaries: Explanatory Dictionary of Ukrainian language in 11 volumes, Etymological Dictionary of Ukrainian language in 7 volumes (5 volumes are already printed, vol. 6 – in print, vol. 7 in edition), Dictionary of Synonyms of Ukrainian language, Phraseological Dictionary of Ukrainian language.

[5] Shirokov V.A. New classification of Ukrainian verbs // Collected papers Actual problems of Ukrainian linguistics: theory and practice. 2003. – issue VIII. – P.113-125. Latent symmetry of lexicographical system of Dictionary of Ukrainian Language (in print).

international lexicographical projects is still impossible without finding solutions to a number of problems – both organizational and engineering as well as pure linguistic ones.

2. To illustrate the latter point let us dwell on the problem of the integral lexicographical description of the language system. Theoretically it is, in our opinion, most deeply developed in the works of Yu. D. Apresyan and other scholars of the Moscow semantic school (Апресян 1995, Апресян и др. 2006). Its basic features are as follows: orientation to the reconstruction of the language picture of the world, the principle of integrated lexicographic representation[6] (we shall call it the principle of linguistic integrity (LI)), the concept of lexeme and its integral lexicographical idea, the concept of lexicographical type.

The definition of the integral lexicographical description by Yu.D.Apresyan implies the requirement for lexicographical representation of different ethnically specific language pictures of the world as well as different realizations of the LI principle, specific for each language. The picture of the integral lexicographical description of the language system is, in fact, still more complicated what follows from the analysis of interaction of the grammatical and lexicographical description in the language system. Linguistics has long ago affirmed the opinion that the basic description of any language composes of a vocabulary and a grammar, although they function differently and deal with different objects: in the vocabulary the units of language are described traditionally while in the grammar rules for their variation and combination are formulated to make morphology and syntax as commonly understood. Vocabulary and grammar in the above sense present an opposition and simultaneously supplement each other to compose a complete pattern of the lexicographical description. Its effectiveness depends on the interconsistency between lexicographical and grammatical components. It means that, on the one hand, charasteristics referred to in the rules of the grammar should be explicitly ascribed to units of lexicographical description in the dictionary, and on the other hand, the grammar must describe (as far as possible) all types of language units' peculiarities that the dictionary does not take into account. Otherwise, the dictionary and the grammar cannot effectively interact and offer a correct and inherently coordinated description of the language system.

Thus, an ideal linguistic description conforming to the above conditions should be systemically integrated to unite both methods of description of the language system – lexicographical and grammatical ones – although the provision of integrity, i.e. a comprehensive, multi-level and multi-aspect description of the language, quite natural as it is, turns to be theoretical and does not have a simple realization.

At the same time it seems this provision is indeed a consequence of some more general and fundamental principle or it does have a deeper phenomenological nature. This is confirmed with the fact of two vocabularies, non-terminal and terminal ones, present in the definition both of the concept and the construction of formal grammars. Consequently, a formal grammar must contain at least two lexicographic objects and cannot function without them at all. On the other hand, the theory of lexicographic systems used by us as a formal basis of the lexicographical description of language and presenting a far advanced generalization of the concept of vocabulary, has in its structure some compulsory elements and meanings naturally interpreted as grammatical ones. Therefore, here we can say about grammar as a natural and compulsory structural element of the lexicographical system. We can affirm that both types of description of the language system – grammatical and lexicographical ones – prove to be complementary according to N. Bohr, and therefore a theoretical scheme should exist to join both types mentioned in a united conceptual object. We consider that the theory of semantic states whose initial principles are developed in the works (Широков 1996a, 2004, 2005, 2005a, Русанівський, Широков 2002) could be regarded as such a conceptual diagram.

At the same time the grammatical and lexicographical description of language – each of them – is integrated by its nature and must present the appropriate, specific integrated medium in which different levels, units, relations of language etc. are represented. An integral vocabulary and an integral grammar in theory must give an interpretation not only of the separate, specially chosen and

[6] According to the principle of integrity by Yu.D.Apresyan, when formulating a certain rule of the language the grammatist should deal with the whole set of the language lexemes and take into account all those lexemes submitted to the rule if their special feature is not fixed explicitly in the dictionary entry. Sometimes it makes it necessary to include the information about some distinct lexemes immediately into the rules. However, when describing just another lexeme the lexicographer should deal with the whole set of the language rules and ascribe to the lexeme all the characteristics mentioned in the rules; sometimes it necessitates the inclusion of the information about the rules into the dictionary entries.

prepared elements of the language system, but also contain some adequate tools to process integral objects of language (texts). In our opinion, a universal conceptual means for mapping of this situation is the lexicographical effect in the information systems, the concept we have formulated in a number of works ((Широков 1996, 1996a, 1998)). It explains the mechanism to generate a system of discrete units for each relatively stable, systemically defined level of the perception of the language substance. The patterns of the integration of different grammatical and lexicographical systems can strongly differ. Let us give one of them, maybe the simplest. Let us designate by the symbol

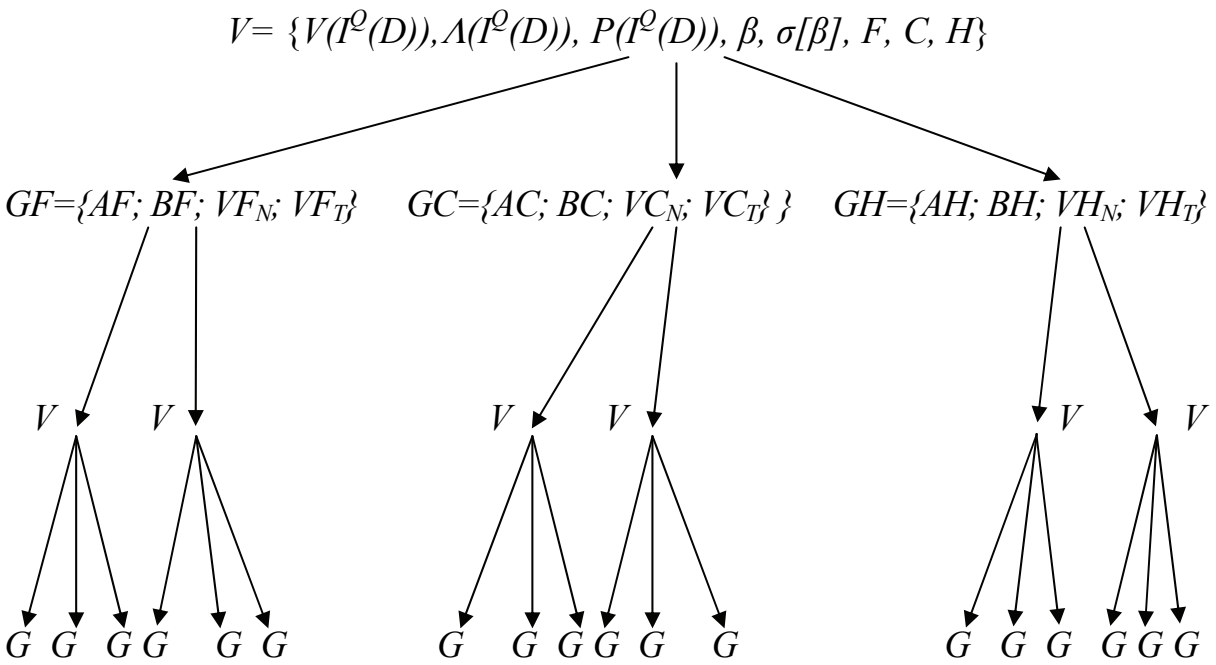
$$G = \{A; B; VN; VT\} \quad (1)$$

a certain formal grammar; in this formula, as always, the symbol A designates the class of the initial elements (axioms) of grammar, B – the set of the basic rules, and V_N and V_T – non-terminal and terminal vocabularies, respectively. By the symbol

$$V = \{V(I^Q(D)), \Lambda(I^Q(D)), P(I^Q(D)), \beta, \sigma[\beta], F, C, H\} \quad (2)$$

we designate in following (Широков 2004) a certain elementary L-system, where $I^Q(D)$ – a class of elementary information units in respect of the lexicographical effect Q over D; $\Lambda(I^Q(D) = FV(I^Q(D))$ and $P(I^Q(D)) = CV(I^Q(D))$ – elements of formal and substantial parts of the lexicographical description $V(I^Q(D))$, respectively; β and $\sigma[\beta]$ represent the microstructure of the L-system; operator H carries out the connection between $\Lambda(I^Q(D)$ and $P(I^Q(D)$.

The scheme of the integrated grammatico-lexicographical description consists in the consequent interpretation of the operators F, C, H as certain grammars (G-systems), and lexicographical elements V_N and V_T – as certain L-systems:



The Ukrainian Lingua-Information Fund has acquired specific, technologically verified experience in developing integrated lexicographical systems, in particular the integrated lexicographical system "Dictionaries of Ukraine" (Palagin et al. 2000) that unites the relations of word declination, transcription, phraseology, synonymy and antonymy for the Ukrainian language. It has been developed and will be released as an industrial product within several years. This experience enables us to make a carefully optimistic conclusion as to the theory of lexicographical systems and the theory of semantic states having a sufficient potential to construct integrated grammar-lexicographical objects (we will henceforth call them GL-systems) actually unlimited in volume and complexity.

3. Let us dwell at some general methodological observations concerning the principles of the simulation of the language system. The first question to appear when formulating these principles, is a question about the objects of simulation, namely: what are the objects of language and what, strictly, we intend to simulate. As a starting point we take the assertion regarding the fact that the language's own objects are some specific psychophysical states of the human thought-speech apparatus and

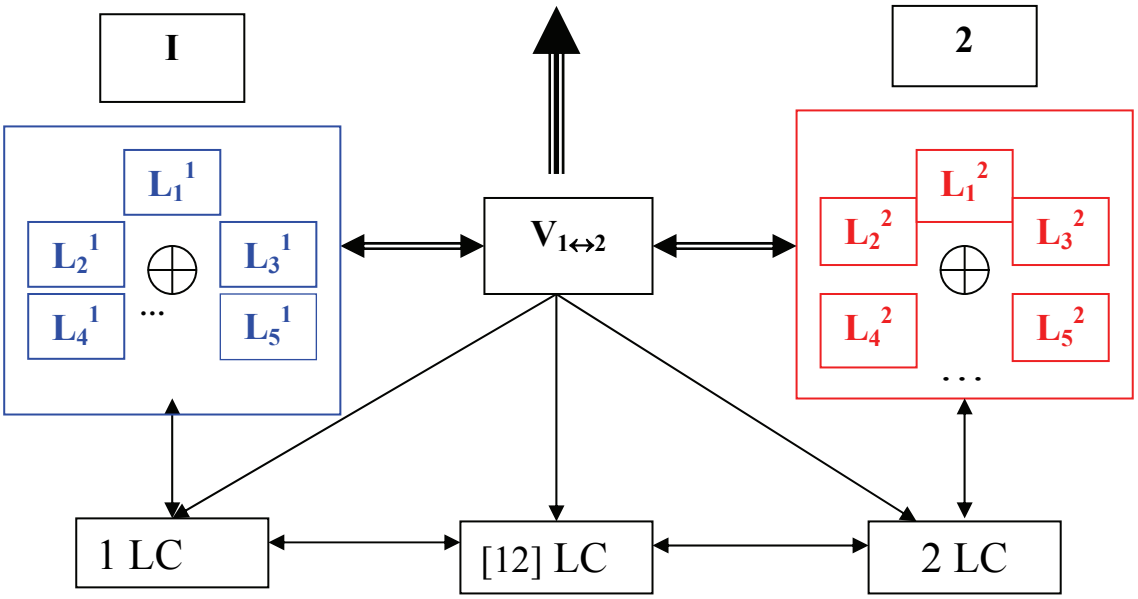
processes going on in it, while the oral and writing forms of the language serve as elements of the infrastructure of the language process. Both the abovementioned psychophysical states and processes and their infrastructure are to be described and simulated. Proceeding from this principle we will try to explain their role in the processes of the simulation of language.

It is obvious that the thought-speech process is integrated in itself since it contains both the language and the mental components. It is realized in the thought-speech apparatus in the form of a dynamic system of the interdependent reflexes, content and nature of which are investigated, for example, in the book by V.M. Bekhterev (Бехтерев 1991) that has not lost its great value until now. According to V.M. Bekhterev's theory one of the so-called combined reflexes proceeding in human brain is the natural language. In this connection, we consider the separation of language processes from mental ones accepted by many linguists as well as attempts to study language "in itself" to be an unjustified and methodologically incorrect simplification. The language system should be regarded as an open one that permits a considerable expansion of both its phenomenological basis and the corresponding conceptual apparatus.

The oral and written forms of language in this sense play the part of models of thought-speech processes and at the same time of a communicative medium for them. Accepting such a factorization it is possible to assert that they present the language periphery. We must warn, however, against a possible underestimation of infrastructural components of language which can arise because we consider psychophysical states and processes of the thought-speech apparatus to be "basic", primary language objects. The point is that contemporary data attest to the fact that language is not an innate property of human being. The capability for language alone is innate. But the process of the "installation" of language in human inevitably requires the availability of such infrastructural elements as the so-called "external" language and the "egocentric" language that function even at the early stages of the phylogenetic development of the language system in child and come to an end with the formation of the "internal" language in him/her to crown the process of creating a full-blooded language apparatus (Широков, Широков 1996). Thus, the language periphery is an inherent element of the language system. Furthermore, it ensures the openness of the language system.

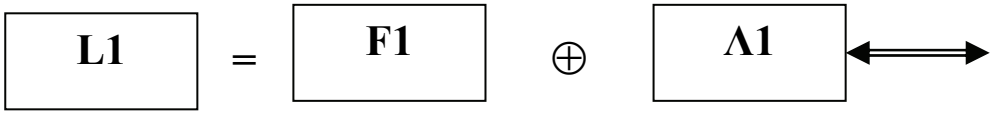
4. It should, however, be noted that the approach described in **2** is realized (and not more than in a few examples) within the limits of one language only. If we speak about the necessity of creating an integrated GL-system to describe several languages, then a set of additional problems emerges. Let us give the overall diagram of the construction of the bilingual GL-system we shall use to try to model the problems of the integral linguistic description of multilingualism.

A General Structural Diagram of the Integral Bilingual GL-System



In this diagram we have used the following designations: L_1^1, L_2^1 , … – GL-systems in the language 1; L_1^2, L_2^2 , … – GL-systems in the language 2; ⊕ – operation of the integration of the GL-

systems; $\mathbf{V_{1\leftrightarrow 2}}$ – GL-system – interface between 1 and 2; 1 LC, 2 LC, [12] LC – linguistic corpora in the languages 1,2 and the parallel corpus (1,2). The construction of the integrated GL-system in every language is standard and if having restructured the integrated GL-systems showed on the diagram in the squares 1 and 2 has the following structure:



Here the symbol **Λ** designates the lexical subsystem of the respective language (for instance, the language 1) chosen as an integrating one for the functional subsystem **F** that in its turn unites the functions of phonetical, morphonological, semantic, phraseological, etymological and other kinds of description for each language. The double arrow shows the connection with the interface subsystem $\mathbf{V_{1\leftrightarrow 2}}$ that joins together the integrated GL-systems of the languages 1 and 2 organized in the same way. The respective linguistic corpora (1 LC, 2 LC, [12] LC) serve as a source to compile elements of the integrated system with the help of the experimental language material.

The construction of the interface subsystem $\mathbf{V_{1\leftrightarrow 2}}$ is especially important for the functioning of the bilingual GL-systems. Its principal feature is a symmetric property, i.e. a symmetry between the entry and the exit of the $\mathbf{V_{1\leftrightarrow 2}}$. It means that if the entry element *x* in the language 1 returns the element *y* in the language 2 we without fail get the element *x* when choosing the element *y* as an entry. The symmetry of the interface can be ensured in using the operator of symmetrization, for instance, in such a way:

$$V_{1\leftrightarrow 2} \equiv S\,[V_{12}] = V_{12} \oplus V_{21} = V_1 \rightarrow V_2 \oplus V_2 \rightarrow V_1$$

where the symbol V_{12} ($V_1 \rightarrow V_2$) designates the bilingual dictionary directed from the language 1 to the language 2, V_{21} ($V_2 \rightarrow V_1$) – the dictionary directed from the language 2 to the language 1; $\oplus$ – operation of integration, S $[V_{12}]$ symmertical L-system between the languages 1 and 2. Similarly, the symmetrization of the GL-systems is carried out. Noteworthy, a generalization of this scheme over three and more languages is by no means a trivial operation. For example, in the case of three languages at least two schemes of symmetrization, each of them not at all simple in realization, can be proposed:

$$1.\quad S[V_{123}] = S[V_{12}] \oplus S[V_{23}] \oplus S[V_{31}] =$$

$$= V_{12} \oplus V_{21} \oplus V_{23} \oplus V_{32} \oplus V_{13} \oplus V_{31} =$$

$$= V_1 \rightarrow V_2 \oplus V_2 \rightarrow V_1 \oplus V_2 \rightarrow V_3 \oplus V_3 \rightarrow V_2 \oplus \oplus V_3 \rightarrow V_1 \oplus V_1 \rightarrow V_3$$

$$S[V_{123}] = S[V_{p(123)}]$$

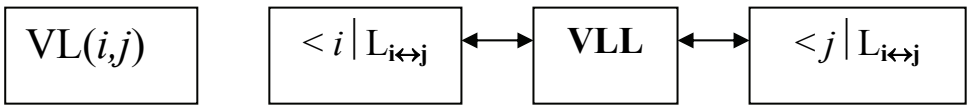
p(123) any rearrangement of symbols 1, 2, 3.

and

$$2.\ S''[V_{123}] = \begin{array}{c} V_1 \rightarrow V_2 \oplus V_2 \rightarrow V_1 \\ \downarrow \quad \downarrow \quad \uparrow \quad \uparrow \\ V_3 \rightarrow V_2 \oplus V_2 \rightarrow V_3 \end{array}$$

For a greater number of languages the situation is getting even more complicated. Therefore, in our case when 6 languages are the object of operation from the very beginning a scheme of the interaction of two languages is the only possible. As we have 15 different pairs, for the project to carry out it is necessary to create 15 interface systems S $[V_{ij}]$, $1 \leq i < j \leq 6$. Simultaneously, computer instruments for creating the integrated GL-systems in every language involved in the project are to be formed and schemes and methods of their network interaction worked out. In such a way 15 virtual

lexicographical laboratories are to be made, each of them embracing a certain pair of languages. As a result, each of the participant institutions of the project would maintain the functioning of five VLLs (obviously, in cooperation with each of its VLL-partners). In a diagram form, the mechanism can be presented as follows:



where the symbol VL(i,j) designates the VLL to unite the languages *i* и *j* , and respectively institutions *i* и *j* connected with the VLL engineering. Thus, a virtual linguistic medium will be created. In it every participant will be connected to the remaining five partners through the VLL instruments for carrying out its respective task. Besides, every participant of the project will get the information access to each of those VLLs any other pair of participants is taking part in as well as will be able to estimate at any time the state of the project and the progress in its realization thanks to the common information-reference system.

Acknowledgments
I would like to thank Dr. Igor Shevchenko for valuable discussions and for his help during the work.

References

Palagin, A.V., Shirokov, V.A. (2000). Principles of cognitive lexicography // International journal «Informational theories & application». Vol. 9, № 2, pp. 43–51.

Апресян Ю. Д. (1995). Интегральное описание языка и системная лексикография. Избранные труды. Т. 2. М.

Русская языковая картина мира и системная лексикография (2006). В. Ю. Апресян, Ю. Д. Апресян, Е. Э. Бабаева, О. Ю. Богуславская, Б. Л. Иомдин, Т. В. Крылова, И. Б. Левонтина, А. В. Санников, Е. В. Урысон; Отв. ред Ю. Д. Апресян. — М.: Языки славянских культур.— 912 с.

Бехтерев В.М. (1991). Объективная психология. – М.: Наука. – 480 с.

Выготский Л.С. (1996). Мышление и речь. – М. – 416 с.

Гейзенберг В. (1989). Физика и философия. Часть и целое. Пер. с нем. –М.: Наука. С.191–196.

Колмогоров А.М. (1987). Три подхода к определению понятия количества информации. //Теория информации и теория алгоритмов /Колмогоров А.Н. –М., Наука – 304 с.

Русанівський В.М., Широков В.А. (2002). Інформаційно-лінгвістичні основи сучасної тлумачної лексикографії//Мовознавство, № 6.сс.7-48.

Успенский В.А. (1957). К определению падежа по А.Н.Колмогорову. Бюллетень Объединения по проблемам машинного перевода. — № 5. — М.: [I МГПИИЯ] — сс. 11 – 18.

Широков В.А. (1996). Информационно-энергетические переходы: механизм действия информационных ресурсов в производственных системах. //III Международная научно-практическая конференция "Информационные ресурсы: создание, интеграция и использование" – Ивано-Франковск.

Широков В.А. (1996а) Інформаційно-енергетичні трансформації та інформаційне суспільство. «Наука. Інновація. Інформація.», Українсько-польський науково-практичний журнал. №1, сс. 48-66.

Широков В.А. (1998). Інформаційна теорія лексикографічних систем. –К.: Довіра. 331 с.

Широков В.А., О.Г.Рабулець, І.В.Шевченко, О.М.Костишин, К.М.Якименко. Інтегрована лексикографічна система “Словники України”, версія 3.0, ISBN 966-507-201-3, 2001–2006.

Широков В.А., Рабулець О.Г., Костишин О.М., Шевченко І.В., Якименко К.М.. (2002) Технологічні основи сучасної тлумачної лексикографії// Мовознавство, № 6. сс. 49-86.

Широков В.А.. (2003). Всеукраїнський лінгвістичний діалог у контексті теорії лексикографічних систем. // Мовознавство, № 6.сс.3–7.

Широков В.А. (2004) Феноменологія лексикографічних систем. – К.: Наукова думка. - 331 с.

Широков В.А. (2003a). О проекте создания украинско-македонского словаря на базе виртуальной лексикографической лаборатории.//Зборник од научната конференција одржана во Охрид на 21-23 октомври 2003 година. Скопје. 2004. сс.187-208.

Широков В.А. (2005). Семантические состояния языковых единиц. Труды международной конференции "MegaLing'2005. Прикладная лингвистика в поиске новых путей"/ Отв. ред. В.П. Захаров, С.С.Дикарева. - Спб.: Издательство "Осипов", сс. 147 – 162.

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

Широков В.А. та ін. (2005b). Корпусна лінгвістика. – К.: Довіра. 471 с.

Широков В.А. (2005c). Елементи лексикографії. –К.: Довіра. 304 с.

Широков В.А. Очерк основных принципов квантовой лингвистики. Бионика интеллекта.

Широков К.В., Широков В.А. (2005). Застосування формалізму нечітких множин для визначення граматичних станів турецьких слів// Мовознавство, № 5, сс.51-56.

II. Maintenance and Optimisation of Multilingal Digital Environment

Universal Dictionary of Concepts[1]

Igor Boguslavsky, Vyacheslav Dikonov
UPM/IITP RAS, IITP RAS
Moscow-Madrid
bogus@iitp.ru, dikonov@iitp.ru

Abstract
A universal dictionary of concepts, developed as a part of the ongoing effort to create a semantic intermediary language for global information exchange, is presented. The article describes basic principles and contents of the dictionary and outlines the current state of the project. The dictionary can evolve into an open and freely available language-independent resource with many potential applications. For example, the extensible dictionary of concepts can serve as a pivot to uniformly record and link meanings of words of different languages and facilitate creation of bi- and multilingual dictionaries. Another possible use is word sense markup of corpora. It could bring rich extra benefits due to the fact that the same set of concepts is going to be linked with major world languages including Russian, English, Spanish etc. and supported by multiple text analysis tools. There is a possibility of cooperation and exchange between this dictionary project and other projects, which could enhance the output and eventually spare a lot of parallel effort.

Keywords*:* Universal Networking Language, universal dictionary of concepts, universal word, ontological structure, argument structure, semantic web, conceptual network

1. Introduction

This article is dedicated to the creation of a new linguistic resource – the Universal Dictionary of Concepts (UDC), also known as the UNL Dictionary. It is a part of a broader international effort to develop a semantic intermediary language named the Universal Networking Language (UNL) (Boguslavsky et al., 2005; Iraola, 2003). Although the dictionary is closely associated with the UNL language, it has considerable value of its own and can be used as a standalone resource for different scientific and practical tasks not related with UNL.

1.1. What is UNL?

UNL is an artificial language for global information exchange in computer networks (http://www.undl.org). Unlike Esperanto, it is not a language for direct oral communication, but a formal way to record the meaning of a natural language text. The goal of the UNL project is to produce a worldwide standard for language-neutral storage and exchange of textual information in multilingual environment. A document written in UNL can be automatically deconverted into a text in any language. Traditional automatic translation systems often fail to produce correct translation because of inherent ambiguity of the source natural language. UNL offers a possibility to edit the intermediate representation of text and/or interactively guide an enconversion system to achieve practically unambiguous representation of the source text. When used as a pivot, it ensures that the meaning of the document is always adequately expressed. UNL is a powerful tool to capture the meaning of a text and preserve it through translation and linguistic processing. It is also well suited for precise search, knowledge extraction, and AI applications.

[1] The study and preparation of these results have received funding from the EC's Seventh Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX. The authors are also grateful to the Russian Foundation of Basic Research for partially supporting this research (grant No. 08-06-00367).

The UNL project offers much more than the dictionary. Other linguistic resources include specifications of the language and multiple software tools, which provide translation to UNL (conversion) and from UNL (deconversion) into different languages of the world. There are several groups of linguists and computer scientists participating in the UNL project and supporting different natural languages. Such groups work in Russia (English, Russian), Spain (Spanish), France (French), Egypt (Arabic), India (Hindi, Marathi, Urdu), Brazil (Portuguese) and several other countries.

1.2. UNL Representation of Text

The UNL representation of a text is a semantic hypergraph. It consists of nodes linked with semantic role relations and embellished with attributes, which convey various grammatical meanings and attitudes of the author. A node can contain either a single lexical unit of UNL or another graph, as shown in Figure 1. The latter type is known as hypernodes.

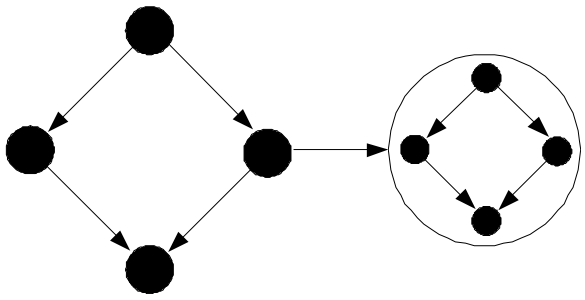


Fig.1. A possible structure of a UNL graph

The basic lexical units of UNL are called universal words (UW). Each UW stands for one single concept.

Although, the principal elements of UNL graphs (UWs, relations and attributes) are technically different in form and function, all of them are just different ways to represent semantic concepts. In some cases it is even possible to choose between using an UW or an attribute, e.g. to express a modal meaning, or prepositional UW and a relation, e.g. for space and time circumstantials. For example, the UW *to(icl>how,plt<uw,obj>thing)* is equivalent to the relation *plt* (target place) and *allow(icl>do,equ>permit,agt>volitional_thing,obj>uw,ben>volitional_thing)* can be an equivalent of the modal attribute of permission. Thus, a UNL graph can be viewed as a pure set of interconnected concepts.

2. Concepts

The concepts of UNL represented in the Universal Dictionary of Concepts are equivalent to the word senses commonly distinguished by explanatory dictionaries. For example, according to the Merriam-Webster, Collins Cobuild, Oxford and other dictionaries of the English language the word *baby* can be used to express the following five concepts:

a human child,
a cub of a mammal animal,
an attractive girl,
a childish person,
a favorite thing, idea or project.

Each of them is a separate lexical unit in UNL and has a unique identifier (UW). This may seem simple enough, but in fact it is not.

If we take several explanatory dictionaries of the same language, it becomes obvious that there is no unity between the authors in how many senses each word really has and how to define them. As of today, there is simply no exact scientific method to draw borders between different concepts pertaining to the same word of a natural language. The only guide here is common lexicographic practice and practical need to distinguish between different ideas, objects and phenomena of the real world. Therefore, a concept is a word sense ascribed to a natural language word in a set of typical contexts.

It is possible to argue that the concepts from the example above are not elementary and should be viewed as compositional constructs containing simpler elements, e.g. *"baby of a human", "baby of an animal", "woman whom I treat as gently as a baby"*, etc. UNL does not follow this approach and refrains from any attempts to decompose the word senses into smaller semantic units. There are both practical and theoretical reasons for this decision. An essential goal of UNL is to provide a simple and easy to understand and edit representation of the text meaning. Disassembling of every word into a plethora of primitives does not help to achieve it. From a theoretical point of view UNL is a shallow semantic language, which presupposes the possibility of deeper (more detailed) semantic analysis in accordance with the principles of stratification and compositionality. The notion of concepts adopted by UNL and UDC fits well with the lexicographic tradition and facilitates the reuse of data already collected in explanatory dictionaries, thesauri and wordnets.

3. Universal Dictionary of Concepts

UDC describes the inventory of concepts used by UNL and serves as the authoritative and exhaustive lexicon of that language. A UW which is not present in UDC should not be used. Any new UW must be submitted to the dictionary. This is an important point for maintaining the lexical compatibility of UNL documents and software tools for automatic translation into natural languages.

3.1. Highlighted Features

- The Universal Dictionary of concepts strives to include and integrate conceptual lexicons of all natural languages.
- The dictionary is characterized by total absence of polysemy.
- Each concept is represented by a universal word (UW). Normally, there should be only one UW per concept.
- The dictionary does not tolerate homonymy, i.e. when one UW is used to express several different concepts.
- The dictionary does not provide any kind of grammatical or morphological information for the simple reason that there is no use for it in UNL.
- All concepts are derived from natural languages. None of them may be invented artificially and the existence of each concept must be justified by some practical need or supported by lexicographic evidence in some natural language. A small number of special abstract concepts, such as *uw*, *thing(icl>uw)*, *abstract_thing(icl>thing)*, etc. have to be privileged because of internal needs.
- If the dictionary lacks a concept, a new UW can be created on demand.
- The dictionary is more than a simple list. It organizes the concepts into a complex semantic network. The structure of this network is outlined in section 5.2.

3.2. Bringing All Tongues Together

It is a common linguistic fact that each natural language has its own unique set of concepts and there are concepts which are specific to certain languages. In fact, we should not expect that concepts which are truly identical for several languages will constitute the majority. Even very common facts and notions can be treated differently by other languages. For example, the English general concept of *"grandmother"* (*the mother of one of the parents*) does not exist in Swedish. Instead, two different words and concepts are used: *"mormor"* (*the mother's mother*) and *"farmor"* (*the father's mother*). UDC will include all three concepts and many more.

In order to be able to record all natural languages accurately the Universal Dictionary of Concepts should grow into the “Summa Lexicographica” of the human kind. This is an immense challenge, which no single group of linguists can meet. The Universal Dictionary can never be considered complete and can grow forever, because the scientific and cultural progress always adds new concepts. However, a dictionary does not have to be complete in order to be usable. There is a practical threshold where the number of registered concepts becomes sufficient for adequate recording of most texts.

4. Universal Words

This section provides only a brief overview of the UW format. More information and rules for UW construction can be found in (Boguslavsky, manuscript).

Universal Words (UW) are used in the dictionary in order to represent the concepts unambiguously. The inventor of the UW format H. Uchida made a lot of effort to achieve intuitive understanding of the concepts on the basis of the UWs alone, without any additional explanation. Nevertheless, most UWs are supplied with a short definition and an example (currently only in English).

A UW consists of a headword and a list of constraints used to differentiate between different concepts associated with the headword and provide additional information. A constraint consists of a UNL relation and another UW, usually reduced to its headword. The general UW format is:

headword(relation>uw>uw,relation>uw,...)

The headword is usually an English word.

cut*(icl>wound>thing)*

If the new concept is expressed by a phrase, the phrase becomes the headword. Spaces are replaced with underscores.

morse_code*(icl>code>thing,equ>morse)*

If there is no corresponding word in English and the concept is a hyponym of some already existing one, we should only change or add constraints. The first of the following three UWs stands for a general concept of entering into a marriage. The other two are its hyponyms describing two aspects of the action differentiated by some languages.

marry(icl>do,agt>person,obj>person)
*marry(icl>do,****agt>man,obj>woman****)* *marry(icl>do,****agt>woman,obj>man****)*

If the new concept is culture-specific and has no hypernym in English, we can use the native word transliterated into Latin and supplement it with constraints that would link it with the nearest commonly known class of objects.

tarator*(icl>****soup(icl>food)****>matter)*
lapot*(icl>****footwear****>..,equ>bast_sandal,com>russian_peasantry)*

UW constraints convey only a minimal amount of information required for identification of concepts. There are three types of constraints: ontological, semantic and argument.

Ontological constraints reflect the most important links between concepts: hypernymy (icl), meronymy (pof), instantiation (iof).

*tongue(****icl****>concrete_thing,****pof****>body)*
*madrid(****iof****>city)*

Semantic constraints are used to show the difference between several concepts associated with one headword: synonymy (equ), antonymy (ant), association (com).

*ably(icl>how,****equ****>competently,****ant****>incompetently,****com****>able)*

Argument constraints reflect the semantic frame of the concept: agent (agt), object (obj), second object (cob), source (src).

*buy(icl>get>do,****agt****>person,****obj****>thing,****cob****>thing,****src****>thing)*

More detailed information about the relations between UWs is going to be stored in the semantic network of the Universal Dictionary of Concepts.

5. Structure of the Dictionary

The Universal Dictionary of Concepts must include three principal components:

1. the repository of concepts, commonly referred to as the dictionary of UNL;
2. the network of relations between concepts, which is known as the UNL Knowledge Base (UNLKB)[2];
3. the local dictionaries, which link concepts with words of various natural languages.

5.1. Inventory of Concepts

The inventory of concepts is a collection of all concepts available in the dictionary and the UNL language in the form of a flat list of UWs. There is no distinction between UWs for concepts coming from different languages. **All concepts are equal as separate lexical units** of UNL and listed together.

In principle one concept should be represented by only one UW. However, it is hardly possible to avoid a situation when several different UWs for the same concept appear. It may happen due to technical and organizational reasons in a decentralized community and the dictionary must provide adequate means to handle this situation.

The first and easiest case is when an already existing UW is modified in order to correct an error, achieve better disambiguation or supply missing information. The old version of the UW cannot be deleted immediately, because it can be used by existing UNL documents (or linked to by other resources). Simple deletion would render such documents incompatible with the dictionary. Although all UNL-related software tools must be able to process documents with unknown UWs, the percentage of such UWs should not exceed the level when it starts to affect the quality of translation. The dictionary has to support per-UW history of changes, allowing to trace any registered version of the UW and prevent reintroduction of deprecated UWs in the same version of the dictionary.

The second source of different UWs for the same concept is the very nature of human language and categorization processes. Each natural language contains a certain amount of exact synonyms which may or may not drift apart with time, e.g. *everyone* and *everybody* in English. It is extremely difficult to build a definitive list of them. Therefore, people will keep adding multiple UWs based on such words even if the corresponding concept already has an UW.

Both processes effectively create groups of UWs resembling synsets used by the Wordnet family of dictionaries. Such groups could be distinguished among all synonyms, viewed as close yet different concepts.

5.2. Network of Concepts

The concepts create a semantic network linked by the relations of hypernymy, meronymy, instantiation, synonymy, antonymy, association and various other relations describing argument frames. The goal of the semantic network is to provide description of the links between concepts, that exist in the human languages and minds, and make it as objective as possible.

The network of concepts consists of three separate structures formed by a) the ontological relations, which link the concepts with different semantic classes, b) semantic relations, which reflect similarity or contrast between concepts, and c) argument relations, which specify what classes of concepts can fill argument slots of each concept.

5.2.1. Ontological Structure

The ontological structure consists of the **icl** (hypernymy), **pof** (meronymy) and **iof** (instantiation) relations. They can be supplemented with some other types of relations, such as **val** (value of) and **scn** (domain of).

---

[2] In older UNL publications UNLKB can be referred to as the Master Entries dictionary. This name is related with the idea of Master Definitions of UWs – an extended form of UWs, which contains full set of relations with any other concepts. Currently the master definitions are not used, but they can easily be derived from UNLKB.

The *icl* and *iof* relations have a privileged status because it is obligatory for every UW to specify at least one more general ontological class through these relations. A concept should be linked to all classes, an immediate member of which the concept is. The result is a hierarchy of ontological relations embedded into a network of other relations. Hypernymic classes are hierarchical by nature and with certain approximation can be arranged in the form of a tree, although the real relations between them can be more complex (see Figure 6). UDC offers a more robust and realistic way to represent the relations between classes of concepts than a regular tree. The resulting base structure is a hybrid one. It combines features of a tree and a network. The branches may split and later join, as shown in Figure 2, yet there is a common root.

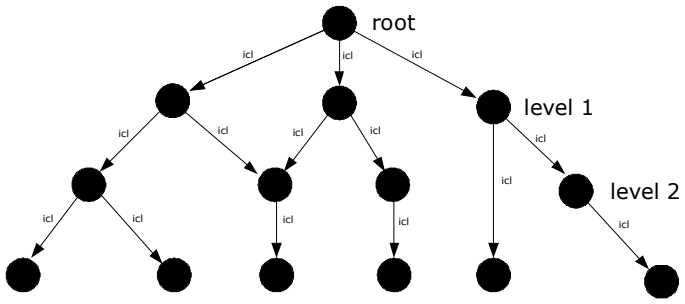


Fig.2. Ontological structure

The abstract root class is named "uw" (any universal word) and divided into further abstract classes of objects, attributes, actions, states, etc. It is possible to talk about different levels of the ontological structure, but a concept in UDC may belong to more than one level or branch.

Ontological relations make it possible to trace the relative semantic volumes of concepts and find more general terms if no direct translation is possible into the target language. For example: while translating the Russian word *жениться*, which means literally *"to acquire a wife"* and has no exact equivalent in English, we should replace it with the more general concept *"to become married"*, which has a straightforward translation.

5.2.2. Semantic Structure

The semantic structure has a different layout. It consists of the semantic relations **equ** (synonymy), **ant** (antonymy) and **com** (association). The equ relation does not distinguish between real and quasi-synonyms and can be supplemented with other technical means to mark sets of UWs denoting the same concept. The semantic relations unite groups of concepts and do not form any hierarchy. Therefore, the resulting structure is a pure decentralized network, as shown in Figure 3.

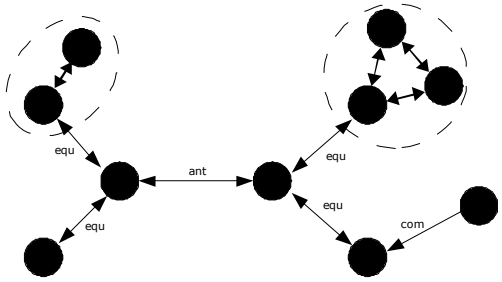


Fig.3. A fragment of semantic structure

There is no requirement for the semantic structure to be connected, unlike the ontological one. It may consist of multiple isolated fragments.

5.2.3. Argument Structure

The argument structure is a collection of argument relations, e.g. **agt** (agent), **obj** (object), **ptn** (partner), **ben** (beneficiary), **plt** (target place), **src** (source), **gol** (resulting state), etc., connecting each concept with an argument frame and general class concepts, which unite all specific concepts that normally fill respective argument slots. In most cases the argument relations point to concepts which belong to a relatively compact group of the most general ontological classes, which occupy the topmost levels of the ontological structure (Figure 4).

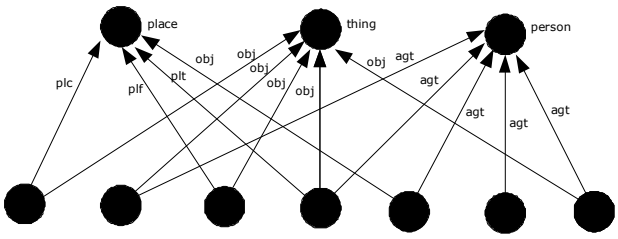


Fig.4. Argument structure

All three structures link the same concepts and are superimposed on each other, forming the network of concepts of UDC.

5.3. Local Dictionaries

Local dictionaries are optional parts of the Universal Dictionary. They are used to connect concepts with the vocabularies of different natural languages. Each language should have a local dictionary in order to be supported. The local dictionaries can be just flat lists enumerating pairs of concepts and their translations into the target language. The natural language words may be supplied with grammatic information.

A translation does not have to be one word. Some concepts represented by a single word in one language may be translated into another by multiword phrases and abbreviations, e.g. *senior pupil* or *VIP*.

However, not all concepts can be translated into all languages even descriptively. If there is a need to translate such a concept, a nearest general term or a more specific one can be found via the network of concepts. Figure 5 provides an example. It outlines relations between Russian (left) and Bulgarian (right) words for *pen*, *handle*, *knob*, *stem* and *tiller* with UWs as a pivot. There is no direct equivalent in Russian for the Bulgarian word *дръжка* in the sense of *stem of a plant*. The translation must be chosen by tracing the ontological (icl) links between *stem of a fruit* and *stem of a flower*. Additionally, there are two alternative Bulgarian translations for the concept *pen.*

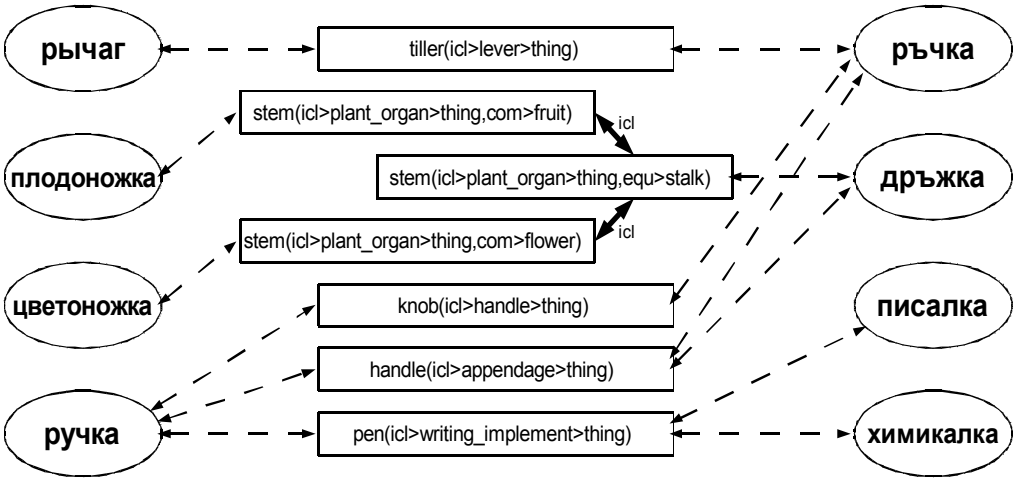


Fig.5. Concepts and possible links between some Russian and Bulgarian words

6. Universal Dictionary of Concepts and Wordnet

The Universal Dictionary of Concepts is quite similar to the well-known Wordnet family of dictionaries in many important aspects. Both have concepts as their basic units and define similar relations between them. A lot of data have been imported from Princeton Wordnet (Bekios et al., 2007). Even more information, including concepts and relations (Iraola 2003), can be imported from different existing Wordnets into the Universal Dictionary of Concepts. However, there are some important differences between UDC and Wordnets.

6.1. Relation to Natural Languages

Each Wordnet describes the lexical system of a particular language and each language is maintained separately. Wordnets may be interconnected by means of the Inter-Language-Indexes (ILI), which describe the relations between the concepts of certain versions of the original Princeton Wordnet (typically 1.5 or 1.6) and concepts of other national Wordnets. However ILIs play a subsidiary role. Only some non-English Wordnets are linked to the original Princeton Wordnet and such links get outdated as soon as a new version of it is released.

The Universal Dictionary of Concepts can be compared to several Wordnets linked through ILI, but it has no bias towards any particular language. The emphasis is given to the unified inventory of concepts and their relations. Links to vocabularies of natural languages are provided through optional local dictionaries and do not have to be discarded when changes are made in the repository of concepts and the semantic network.

The fact that most of the UW headwords come from English and the constraints in so many UWs are motivated by the need to describe the polysemy of English words, might suggest that the dictionary uses English as a pivot or "gold standard" to describe other languages. However it is not quite true. English headwords and constraints were chosen for mere practicality, because most linguists understand this language and it uses the most common and well supported A-Z script in the world. It is also a fact that not all UW headwords are English.

Concepts coming from any language receive identical status. Concepts originating from different languages can have direct links between each other. Non-English concepts may also be used as a base for modification and as constraints to describe other concepts. For example:

samovar*(icl>boiler>concrete_thin,com>tea)*
*tula_samovar(icl>****samovar****>concrete_thing,com>tula(iof>city))*

sauna*(icl>sweating_room>place,com>finnish,com>dry)*
parilka*(icl>sweating_room>place,com>russian,com>steam)*
*venik(icl>massage_tool>...com>****parilka(icl>sweating_room)****)*

If the number of concepts unique to other languages increases, the statement about the special role of English in UDC will lose ground.

6.2. Hierarchical Structures

Wordnets organize the noun and verbal concepts into hypero-hyponymic hierarchies represented as trees. Such structures are easy to search and analyze, but pure tree classification does not support partially intersecting classes and works well only for the top classes of ontology. For example, Princeton Wordnet has concepts of (*tennis) racket*, and *(hockey) puck* as well as a class for "*sports implements*". However, *racket* is a member of the class of sports implements and *puck* is not. Instead it is a member of the class of "*disk objects*". Moving *puck* to the "*sports implements*" class in a pure tree structure would cause losing information that it is a disk.

UDC is able and strives to accommodate a different less formally hierarchical approach. The basic ontological structure is a network graph which has only some features of a tree. It is normal to have multiple parents to the same daughter node, which allows for more complex relations and more fine-grained classification. Every concept should be linked to all possible immediate hypernyms. For example, the word *sushi* in Wordnet is a direct daughter of the concept *dish* (food). Suppose that we want to introduce further ontological divisions by nationality (sushi is a Japanese dish) and primary

ingredient (sushi is made of fish). It is not possible to decide which of the two classes has to be placed higher in the hierarchy, because these classes specify intersecting sets of concepts (Figure 6)[3].

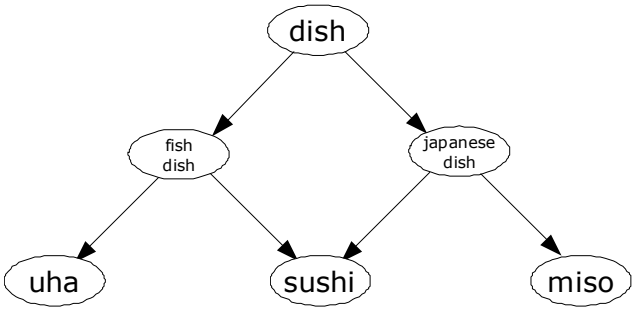


Fig.6. Multiple parent classes

Using a network instead of a tree has some implications. A tree structure allows to trace every concept to its deepest root classes with full confidence, whereas the hybrid network structure permits multiple paths, leading to different high-level classes for the same concept, even when it creates confusion. For example, the class "*functional thing*", which includes the concept of *hammer*, is a daughter of both "*abstract thing*" and "*concrete thing*", thus making *hammer* a possibly non physical object! This problem can be remedied in UWs by providing a secondary direct link to the relevant top class.

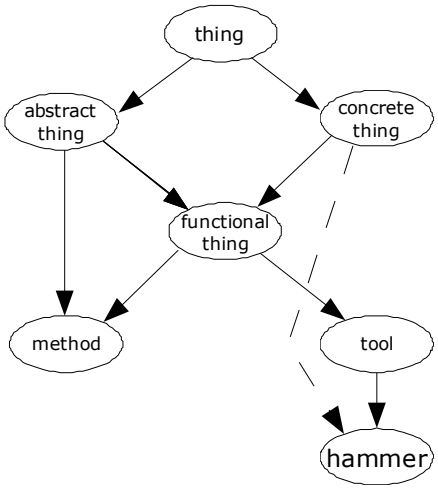


Fig.7. Additional link to the relevant top class

According to Figure 7, the UW for the concept *hammer* should be *hammer(icl>tool>**concrete_thing**)*. Knowing two ends allows to trace the ontological relations between any concept and the relevant top class and produce full hierarchy.

6.3. Other Features

Wordnet does not make the difference between hypernymy as a relation between classes (e.g. the class of "*living things*" includes the class of "*plants*") and instantiation as the relation between an individual and a class to which it belongs, e.g. *Deli* is a member of the class "*cities*". In UDC two different relations are used for such cases: **icl** for hypernymy in *plant(icl>living_thing)* and **iof** for instantiation in *Deli(iof>city)*.

[3] Princeton Wordnet provides a way to include a synset into several classes at the same level of its hierarchy too, but this is not common. For example, *key* in the sense of "*a kilogram of a narcotic drug*" is described as both "*a mass unit*" and "*a metric unit*" at the same level and this split is immediately joined at the next level under the "*units of measurement*" class.

UDC does not limit itself by certain parts of speech like Princeton Wordnet and provides full set of concepts for prepositions, conjunctions and some words with special grammatical functions, e.g. modal verbs.

UDC provides more detailed semantic frame information, not limited to the verbal concepts. The roles are annotated with UNL relations and prototype semantic classes of the arguments are given where Princeton Wordnet offers only "somebody" and "something".

Some wordnets preserve syntactic information about the words, such as part of speech, gender, animacy, etc. (Сухоногов, Яблонский, 2004), while other are coupled with morphology engines. This is not the case in the Universal Dictionary because such information is unneeded in the UNL language. Its proper place is in the local dictionaries.

7. Development of the Dictionary

The development process should follow the essential principles of **division of labor, gradual development, reuse of existing data and decentralization**. A community model, where everyone checks everyone and all significant disputes are resolved by experts, is the best option, because no single authority can have enough resources and expertize to verify everything.

Every time when a significant amount of changes is done and no formal objections received, a snapshot of the dictionary should be taken and released as a new version. From that moment all participating parties must update their tools to use the new dictionary. An automated system to propagate UW changes to local copies utilized by linguistic processors supporting UNL is required to ensure smooth transition to any new versions of the dictionary.

7.1. Current Status

At the moment of writing the Universal Dictionary is under active development. It has already passed a number of important milestones including: adoption of the common UW guidelines (Boguslavsky, manuscript) and creation of the initial set of UWs completely covering the general vocabulary of English. The current version of the dictionary includes about 200 000 UWs generated on the basis of the Princeton Wordnet (Bekios et al., 2007) and about 9 000 UWs (Диконов, 2008) created manually to fill in the gaps found in Wordnet. The manually written UWs cover English prepositions, conjunctions, and certain other words left out of Wordnet. A significant portion of them replaces the automatically generated UWs for the most frequent English verbs and nouns in order to improve the quality of the UWs.

The existing inventory of UWs was merged with the dictionaries of the linguistic processor ETAP (Диконов, 2008), developed by the members of the Russian group, and is used for text conversion from and deconversion to English and Russian. The automatically generated UWs are available online at http://www.unl.fi.upm.es/unlweb.

The French group develops an infrastructure for the central data repository and exchange of data between different groups. Considerable effort is made by different participants towards massive revision and correction of the generated UWs.

The next step can be enriching the semantic network beyond the links already available in the form of UW constraints.

7.2. Availability

The Universal Dictionary is going to be released to the public under a free license as soon as the first version will be ready, which presupposes merging in more UWs from other UNL groups and putting in operation the infrastructure for automated data exchange.

The essential principles to be maintained are:

- The Universal Dictionary of Concepts will be available to the public free of charge.
- The data may be used freely for any purpose, though commercial use may be a subject to special conditions.
- Everyone will be given the right to expand the resource and fix errors, provided that all modifications will be returned to the community of dictionary users and editors.

8. Possible Uses and Related Projects

The dictionary of concepts can be used as a standalone resource to match words of different languages for automatic generation of multilingual dictionaries, provided that all such languages have local dictionaries.

Universidad Politécnica de Madrid (UPM) runs a project named Patrilex (Boguslavsky et al., in print) which is aimed at experimental verification of this approach. The practical goal is to produce a multilingual dictionary of terminology in the domain of culture and national heritage for the Spanish Ministry of culture. A special custom set of UWs for the relevant terms is being built and independently translated into English, Spanish, Russian and Arabic. The translators receive flat lists of the UWs without any additional information and independently write local dictionaries for their languages. The resulting multilingual dictionary will be assembled automatically and verified to detect any problems.

Another possible use is to annotate lexical meanings after word sense disambiguation, e.g. for semantic annotation of corpora. There is a need for a reference corpus of UNL, but it is not yet created. The most relevant effort in this field is the project to translate the Encyclopedia Of Life Support Systems (EOLSS) into several languages via UNL.

The overall progress of the UNL project may seem slow, but current projects show that it is real. A quantum leap is expected as soon as the first public version of the Universal Dictionary is released and the tools for automatic conversion of text into UNL documents reach industrial quality. Every new related project and contribution make this perspective closer.

References

Bekios, J., Boguslavsky, I., Cardeñosa, J., Gallardo, C. (2007). "An Efficient Method for Building Multilingual Lexical Resources" In: Proceedings of the *Fifth International Conference Information Research and Applications* i.TECH 2007, Varna, Bulgaria, v.1. Sofia, Ithea, pp. 39-45.

Boguslavsky, I., Cardeñosa, J., Gallardo, C. (in print). "A Novel Approach to Creating Disambiguated Multilingual Dictionaries ".

Boguslavsky, I., Cardeñosa, J., Gallardo, C., and Iraola, L. (2005). "The UNL Initiative: An Overview" Computational Linguistics and Intelligent Text Processing.

Boguslavsky, I.M. (*manuscript*). "Guidelines for UW construction".

Диконов, В. (2008). Развитие системы построения семантического представления текста с использованием языка-посредника UNL на базе лингвистического процессора ЭТАП-3. *Информационные технологии и системы*, ITiS'08, Геленджик.

Iraola, L. (2003). "Using WordNet for linking UWs to the UNL UW" *International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies*, December 2 - 6, 2003, Alexandria, EGYPT.

Сухоногов, А.М., Яблонский, С.А. (2004). Разработка русского WordNet. RCDL2004, Пущино, Россия.

Web site of the UNL project, http://www.undl.org

Lexicographer's Companion: a User-Friendly Software System for Enlarging and Updating High-Profile Computerized Bilingual Dictionaries[1]

Leonid Iomdin, Victor Sizov
IITP RAS
iomdin@iitp.ru, sizov@iitp.ru

Abstract
A sophisticated software tool is presented which is used to expand and update electronic dictionaries.The tool operates interactively and enables the introduction of new entries into bilingual and monolingual dictionaries as well as collocations and their translation equivalents. At the moment, the Lexicographer's Companion is tuned to the ETAP-3 linguistic processor and works with Russian and English but it can be easily modified to meet the needs of any multilingual linguistic dictionary resources.

Keywords: lexicographic resources, lexicographic tools, digital bilingual and multilingual dictionaries, update and maintenance of digital dictionaries

1. Introductory Remarks

Digital bilingual dictionaries, both human-oriented and application-oriented ones, require convenient tools that should ensure their smooth, fast and error-free update and expansion. Such tools are especially important if these dictionaries are managed by end users rather than by their authors (which may be the case if an application is maintained and further developed outside of the development team) but can also be of great help to the original developers.

In what follows, a software tool designed for these purposes in the Laboratory of Computational Linguistics, Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, by the developers of a multipurpose linguistic processor, ETAP-3, is described in detail.

Digitalized dictionaries are a vital part of the ETAP-3 processor. In one of the main options of ETAP-3, the machine translation system, bilingual dictionaries contain versatile data on lexical units of the languages concerned: these data must be perfectly matched with the data presented in the other linguistic component underlying the ETAP-3 system – the grammar.

The creation, update and expansion of bilingual dictionaries require enormous effort on the part of the system developers and/or end users. The tool, Lexicographer's Companion, offers the developers a considerable level of automation in dictionary management.

The authors firmly believe that the Lexicographer's Companion can be extremely useful in a multilingual environment of lexicographic resources to be developed in the MONDILEX project: even though the Companion is currently oriented to the Russian/English language pair, its adaptation to other language pairs, or groups, can be performed rather easily.

2. The Structure of ETAP-3 Dictionaries

The ETAP-3 linguistic processor uses two kinds of dictionaries – morphological and combinatorial ones.

2.1. Morphological Dictionaries

Morphological dictionaries are used for morphological analysis and generation. Each entry of the morphological dictionary specifies a word's paradigm: the complete list of word forms matched with the word's name (the lexeme) supplied with a string of morphological features. To give an example, if the entries of this dictionary were written in a straightforward manner,

[1] The study and preparation of these results have received funding from the EC's Seventh Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX. This study has also received partial funding from the Russian Foundation of Basic Research (grant No. 08-06-00373), which is gratefully acknowledged.

then the entry of the English morphological dictionary for GIRL and WOMAN would look, respectively, as sets of strings

girl	GIRL, sg	*woman*	WOMAN, sg
girls	GIRL, pl	*women*	WOMAN, pl
girl's	GIRL, sg, poss	*woman's*	WOMAN, sg, poss
girls'	GIRL, pl, poss	*women's*	WOMAN, pl, poss

where lowercase elements in italic represent word forms, uppercase elements represent the lexemes' names, and sg, pl and poss are morphological features that stand, respectively, for singular, plural and the possessive case.

In morphological analysis, such entries are used to match the word forms of the text processed with the lexeme's names supplemented with feature strings; in morphological generation, or synthesis, the reverse operation is performed: for the lexeme accompanied by the feature string, a word form is produced.

It goes without saying that the creation and management of morphological dictionaries in which entries are straightforward lists of paradigms would be extremely cumbersome and virtually impossible. It might be tolerable for languages with little inflexion such as English where most paradigms consist of no more than 5 or 6 strings, but totally unacceptable for Slavic languages with their rich morphology: suffice it to say that the paradigm of a regular transitive two-aspectual Russian verb consists of 225 elements.

Accordingly, real morphological dictionaries resort to specially designed standard objects representing types of paradigms or parts thereof. In ETAP-3 dictionaries, such objects include lists of endings for different paradigm types, lists of alternations, formats, templates, and masks[2]. These objects enable the creation of compact representations of entries.

The English dictionary entry for GIRL, for example, will look like

```
girl
bas:= f:10
```

(format No. 10 is reserved for regular English nouns that have regular forms of plural ending with –s, and the expression "**bas:=**" means that the stem of all word forms coincides with the lexeme's name).

The Russian dictionary entry for ДЕВОЧКА 'girl' will look like

```
девочка
осн:девоч(е)к чер:4 т:104
```

(here, the stem of the word can be either *девочк* or *девочек*, the template No 104 lists endings for the first declination animate feminine nouns whose stems terminate with letters к, г, х and the alternation type No. 4 specifies the elements of the paradigm where the stem is presented in its longer form *девочек* – this is genitive plural and accusative plural).

The Russian dictionary entry for ПУТЬ 'way', which is an irregularly formed noun, will look like

```
пут|ь
осн:= ф:1,сок:5/2,ок:'ем'твор,ед
```

In this entry, the stem of the word is *нут* (which is shown by the vertical bar in the lexeme's name denoting that the stem is on its left). Format No. 1 indicates that the word is a masculine inanimate noun. The list of endings No. 5 is normally reserved for standard Russian feminine inanimate nouns ending with *–ь* but can be used for *путь* with the exception of one word form – instrumental singular, which is formed differently. An adequate morphological description is achieved through using mask No.2 (written following a slash after the list of endings), which disables the respective word form, and straightforwardly adding the missing form by specifying the ending 'ем' and features "твор,ед" (instrumental singular).

[2] Masks are conditions imposed on standard paradigms stating that certain word forms are to be eliminated from these paradigms – they may be formed differently or be absent altogether.

For a number of technical reasons, entries of ETAP-3 morphological dictionaries each include a line of a simple translation equivalent into the opposite language within the Russian/English pair.

Despite the fact that the methods and ways of representing morphological information in the morphological dictionaries are rather standard and compact, it is not at all easy for a developer (let alone an end user) to manage these dictionaries, especially in cases of complicated types of paradigms like irregular Slavic verbs.

2.2. Combinatorial Dictionaries

Combinatorial dictionaries of the ETAP-3 system are high-level lexicographic resources that convey a variety of data types for each word: sophisticated syntactic features, semantic classes, subcategorization frames (government patterns), lexical functions, certain word combinations, different types of linguistic rules or references to these rules, etc. They are linked to morphological dictionaries through the lexeme name: in the majority of cases, each entry of the combinatorial dictionary has its morphological counterpart. As a matter of fact, the combinatorial dictionary is a somewhat simplified version of the explanatory combinatorial dictionary (ECD) of the Meaning ⇔ Text theory (see e.g. Мельчук 1974, Мельчук-Жолковский 1984) on which the ETAP-3 system is largely based. The main difference between the ECD and the combinatorial dictionary is the absence of regular lexicographic definitions in the latter.

A combinatorial dictionary entry has a rather complex hierarchical structure. It is composed of a general zone representing the source language without regard to other languages involved in translation options, and a number of target language zones offering translations of the headword and word combinations involving this word into the particular target language.

The general zone and target language zones may include subareas, called subject domains, which handle behavior peculiarities of the headword lexeme in texts belonging to particular subject domains, like science, sports, TV, or even more specific domains.

Zones of the dictionary, in particular the general zone, are structured into fields, such as syntactic feature field, semantic field, subcategorization frame field, etc. In most cases, linguistic data recorded in the fields are presented declaratively. However, a considerable part of the data may only be conveyed algorithmically, in the form of linguistic rules composed of two parts: an inventory of conditions to be checked and a list of actions to be performed if the conditions of the rule are satisfied. In the case if identical rules should be introduced in many entries, such a rule is stored in a separate file and the entry is supplied with a reference to this rule, with individual parameters specified for each entry if need be.

Schematically, a combinatorial dictionary entry looks as follows:

- General zone
  Headword field specifying the lexeme name and the time of the last update (mandatory)
  Subject domain for the lexeme (optional)
  Field of discriminating comments enabling the discrimination between word senses in case of lexical ambiguity (optional)
  Part of speech (mandatory)
  Syntactic features field (optional)
  Semantic descriptors field (optional)
  Subcategorization frame field (optional)
  Lexical functions field (optional)
  Operational fields containing rules and references to template rules (optional)
    - Specialized subject domains (optional)

  Operational fields containing rules and references to template rules relevant for the subject domain (optional)
- Target language zone
    - Specialized subject domains (optional)

  Default translation field (optional) containing one or more single word translational equivalents in the target language (optional)
  Field of discriminating comments enabling the discrimination between translational equivalents (in the case of multiple translational equivalents in the previous field, optional)

Operational fields containing rules and references to template rules relevant for the subject domain (optional).

A fragment of a simple English combinatorial dictionary entry for the noun INFLUENCE is given below for illustration.

```
1   03744 21:37:39 15-05-2008          INFLUENCE1
2   COMMENT:"NOUN"
3   EXAMPLE:"DARWIN'S INFLUENCE ON MODERN THOUGHT"
4   POR:S
5    SYNT:COUNT,VOC
6   DES:'ДЕЙСТВИЕ','ОТНОШЕНИЕ','ФАКТ','ПРОЦЕСС','АБСТРАКТ'
7   D1.1:OF
8   D2.1:ON1
9   D2.2:OVER1
10  _SYN1:IMPACT1
11  _V0:INFLUENCE2
12  _MAGN:POWERFUL/PROFOUND/STRONG/FAR-REACHING
13  _BON:GOOD1/POSITIVE1/SALUTARY/WHOLESOME
14  _ANTIBON:BAD/NEGATIVE1/PERNICIOUS/UNWHOLESOME
15  _OPER1:HAVE/EXERT
16  _OPER2:BE<UNDER1>
17  **************************
18  ZONE:RU
19  TRANS:ВЛИЯНИЕ
20  TRAF:TRADUCT1.25
21  LA:OUTSIDE2,LR:ВНЕШНИЙ
22  LA:BACKSTAIRS1,LR:ЗАКУЛИСНЫЙ
23  LA:BAD,LR:ДУРНОЙ
24  LA:BANEFUL,LR:ПАГУБНЫЙ
25  LA:CIRCUMAMBIENT,LR:ВСЕСТОРОННИЙ
26  LA:COMMAND2,LR:РЕШАЮЩИЙ
27  LA:DEEP1,LR:СИЛЬНЫЙ1
28  LA:DELETERIOUS,LR:ПАГУБНЫЙ
29  LA:DIRECT2,LR:НЕПОСРЕДСТВЕННЫЙ
30  LA:DISRUPTIVE,LR:ДУРНОЙ
31  LA:EVIL1,LR:ДУРНОЙ
32  LA:EXTRANEOUS,LR:ПОСТОРОННИЙ1
33  LA:GENIAL,LR:БЛАГОТВОРНЫЙ
34  LA:GOOD1,LR:БЛАГОТВОРНЫЙ
35  LA:HEALTHY,LR:БЛАГОТВОРНЫЙ
36  LA:MALIGN1,LR:ДУРНОЙ
37  LA:NEGATIVE1,LR:ДУРНОЙ
38  LA:OUTWARD1,LR:ВНЕШНИЙ
39  LA:PERNICIOUS,LR:ДУРНОЙ
40  LA:POISONOUS,LR:ГУБИТЕЛЬНЫЙ
41  LA:POSITIVE2,LR:ПОЛОЖИТЕЛЬНЫЙ
42  LA:SALUTARY,LR:БЛАГОТВОРНЫЙ
43  LA:SECRET2,LR:СКРЫТЫЙ
44  LA:SINISTER,LR:ДУРНОЙ
45  LA:UNWHOLESOME,LR:НЕЗДОРОВЫЙ
46  LA:WHOLESOME,LR:БЛАГОТВОРНЫЙ
TRAF:TRADUCT1.28
47  LA:BEAR1,LR:ОКАЗЫВАТЬ
48  LA:BRING,LR:ОКАЗЫВАТЬ
49  LA:CONSOLIDATE,LR:УКРЕПЛЯТЬ
```

```
50  LA:EXERCISE2,LR:ОКАЗЫВАТЬ
51  LA:EXERT,LR:ОКАЗЫВАТЬ
52  LA:EXTEND,LR:РАСПРОСТРАНЯТЬ
53  LA:GAIN2,LR:ПРИОБРЕТАТЬ
54  LA:WIELD,LR:ПОЛЬЗОВАТЬСЯ
55  TRAF:TRADUCT1.2A
56  LA:DECREASE2,LR:УМЕНЬШАТЬСЯ
57  LA:GROW,LR:РАСТИ
58  TRAF:TRADUCT1.29
59  LA:ABJECTIVE,LR:ДЕМОРАЛИЗОВАТЬ,T1:НЕСОВ,T2:НЕПРОШ,T3:*
60  LA:CORROSIVE1,LR:РАЗЛАГАТЬ2,T1:НЕСОВ,T2:НЕПРОШ,T3:*
61  LA:FORMATIVE,LR:ФОРМИРОВАТЬ,T1:НЕСОВ,T2:НЕПРОШ,T3:*
62  LA:PREPONDERANT,LR:ГОСПОДСТВОВАТЬ,T1:НЕСОВ,T2:НЕПРОШ,T3:*
63  REG:TRADUCT2.D0
64  LOC:R
65  R:QUASIAGENT/1-COMPL
66  N:01
67  CHECK
68  1.1 =(X,PL)&DEP-EQU(X,*,R,'СОВОКУПНОСТЬ')
69  DO
70  1 ZAMRUZ:X(ФАКТОР)
71  TRAF:RA-EXPANS.12
72  LA:ON1
73  TRAF:RA-EXPANS.50
74  LA:INFLUENCE2
```

In this entry, line 1 represents the headword field; lines 2-3 – the discriminating comment; line 4 – the part of speech field; line 5 – the syntactic features field; line 6 – the semantic features field (the descriptors are coded with Russian words denoting ‘action’, ‘relation’, ‘fact’, process’, and ‘abstract thing’); lines 7-9 – the subcategorization frame, lines 10-16 – the lexical finctions field. In the Russian zone starting from line 18, line 19 is the default translation field providing the Russian translational equivalent of the headword; lines 20 to 62 list template translation rules responsible for rendering multiword collocations, lines 63-70 accommodate a dictionary translation rule ensuring that in certain contexts the headword is translated into Russian as *факторы* rather than *влияние*; lines 71-74 are used to list two template rules of Russian-English translation (the first one introduces the strongly governed preposition *on* and the second ensures the substitution of the noun *influence* with the verb *influence* in certain conditions).

3. Typical Tasks of Enhancing Dictionary Coverage of the System

3.1. Introduction of New Dictionary Entries

The following phases in dictionary expansion within the ETAP-3 system could be specified.

- Determine the list of new lexemes to be entered.
- For each such lexeme, determine the concrete dictionaries where it should be entered, and the subject domain of the lexeme (if needed).
- Enter the lexeme into the morphological dictionary.
- Enter the lexeme into the combinatorial dictionary.
- Enter the lexeme(s) of the translational equivalents of the given lexeme in the target language morphological and combinatorial dictionaries (if needed).

By way of example, we will describe a typical set of actions to be performed by an ETAP developer trying to enhance the dictionaries when working on a specific text.

Imagine that a large natural text has been submitted for processing to the ETAP-3 machine translation system. The analyzer may show that some of the words could not be treated properly by the dictionary. In particular, it may happen that

- a word form could not be recognized by the dictionary at all;

- a word form receives an erroneous parse because the corresponding lexeme is missing from the dictionary but the word form is homonymous with some other word form belonging to the paradigm of another lexeme (e.g. the Russian word form *Красноярском* which appears in the text as the prepositional case of the adjective *красноярский* 'pertaining to Krasnoyarsk' is analyzed as the instrumental case of the noun *Красноярск* 'Krasnoyarsk'.

In the first case, the dictionary gap can be detected automatically. The software environment of ETAP-3 has a special program called Textan that lists all words of a text to be processed which are totally absent from the dictionary[3].

In the second case, deep manual analysis of parsing results is necessary.

When entering a word into the morphological dictionary, the operator must determine what standard objects correspond to this word. For richly inflected languages, this task is far from trivial even for an experienced worker. To facilitate the work, ETAP-3 morphological dictionary compiler offers a number of useful features, including the option of paradigm browsing which allows the operator to see the results of his work immediately and correct the mistake if it is made. Another useful feature is the possibility to visualize the decomposition of complex standard objects (e.g. templates or formats) into elementary ones (like lists of endings) so that the author could detect a probable error more easily.

Further, the introduction of a new entry is greatly facilitated if the author is able to select an existing word with the inflectional properties identical to those of the word to be entered. In this case, the author will only need to supply the headword and stems, and provide the default translation into the target language.

The introduction of a new word into the combinatorial dictionary is a task by far more complicated and important. The quality of an entry in this dictionary often affects the successful processing of the whole sentence containing the headword of this entry. The author's work is however complicated by the scarcity of formal criteria prompting the creation of an entry, even though some formal prompts do exist. A good way to proceed is to find an existing entry with properties close to the word to be added – but this is not an easy task. Some entries may differ from even the seemingly close existing ones in the list of syntactic features; others have peculiar values of lexical functions, still others may require a lot of nontrivial translational equivalents or a plethora of idiomatic word combinations, etc.

Besides, there is no debugging method that could help the author of a newly introduced entry to reveal all or even most of the errors that could have been made.

3.2. Representation of Multiword Terms and Phraseological Units in the Combinatorial Dictionary

Representation of such units may be viewed as part of work devoted to the creation of new entries or an independent task. ETAP-3 does not have a separate dictionary of idioms; instead such a unit is represented in the entry of one of the words comprising the unit (in most cases, the syntactic head of the unit).

The data on multiword terms and phraseological units (henceforth, MPU for short) is important as it helps to optimize text processing at least at two different stages. First, this information is taken into account during text analysis when ambiguous lexemes are assigned priorities: it is naturally assumed that the co-occurrence probability of lexically bound words is higher than that of unbound ones. Second, it ensures an adequate translation of an MPU into the target language. ETAP-3 rules responsible for MPU handling can be summarized as follows.

- The CHECK zone of such a rule verifies whether the sentence being processed has the respective MPU. The conditions are considered to be holding if the algorithm has been able to identify a group of lexemes having the respective names which are linked by the specified syntactic relations.
- The actions prescribed in the DO zone of such a rule depend on the specific task solved by the rule. If a rule is applied at the parsing phase, it increases the priorities assigned to syntactic

[3] An additional peculiarity of the ETAP-3 system is that its Russian morphological dictionary is noticeably larger than the combinatorial dictionary (for a number of technical reasons, the main one being that the dictionary incorporates the whole Grammatical Dictionary of the Russian Language by A.A. Zaliznyak with all its obsolete words, see Зализняк 2003). If some word form is represented in the morphological dictionary but absent from the combinatorial dictionary, this program will reveal this fact.

links that connect MPU elements. If it is applied at the transfer phase, the actions include substitutions of the source language lexemes by the target language ones as well as connecting the target language lexemes with the specified syntactic links.

Normally, a MPU can be included into a lexical entry either by adding a reference to a template rule or by adding a unique dictionary rule.

The former method is used if a MPU has a syntactically typical and simple pattern (e.g. it consists of a noun modified by an adjective (*Советский Союз, Soviet Union*), a noun modified by two adjectives (*большой адронный коллайдер*) or a compound modified by an adjective (*large hadron collider*), etc. – and if its translation equivalent in the target language also has a syntactically typical and simple pattern (non necessarily the same one). This method has an obvious advantage of description simplicity and compactness. Due to this fact it proved possible to automatically generate and activate a pre-syntactic rule of priority assignment to MPU elements based on the CHECK zone of a template transfer rule for a given MPU. However, a serious disadvantage of the method is that in order to effectively use it the lexicographer has to remember the names of template rules, at least of the most popular ones (there are almost 300 MPU rules in the ETAP-3 processor today).

The latter method is used if a MPU and/or its translation equivalent has a complex structure and cannot (or should not) be listed among typical patterns, An obvious drawback of the method is that writing unique rules may be cumbersome and labor-consuming.

It follows from the above that much of dictionary work can hardly be automated and requires manual work of skilled and experienced lexicographers.

On the other hand, the maximally possible automation of this work is a very important task. In many cases, this can be achieved using the Lexicographer's Companion.

4. Lexicographer's Companion as a Means of Automating Dictionary Management

In what follows, we will describe how the Lexicographer's Companion is used for some of the dictionary management tasks.

4.1. Initial Data for the Lexicographer's Companion

The initial data unit for the Lexicographer's Companion (henceforth, LexiComp for short) is a pair of translation equivalents "source lexeme/MPU – target lexeme/MPU". At the moment, LexiComp supports two directions of translation: English-to-Russian and Russian-to-English. The user may submit both elements of the pair to LexiComp in any order, as the system is able to define the language automatically.

Pairs of translational equivalents may be submitted to LexiComp one by one or as a list (such lists may for example be produced by the Textan program, see Section 3.1 above, or constructed manually). Lists are tables (in text format) having four columns: 1) source lexeme/MPU; 2) target lexeme/MPU; 3) information on whether the translation of source language unit into target language unit is available; 4) information on whether the reverse translation is available. In every row, only the first column is mandatory; others are optional. A sample list is given below:

Source Lexeme/MPU	Target Lexeme/ MPU	Direct Translation Available?	Reverse Translation Available?
McCain	Маккейн	No	No
Obama	Обама	No	No
Балабанов	Balabanov	No	No
Дулево	Dulevo	No	No
crab catching	лов краба	No	Yes
deathly hallows	роковые мощи	No	No
hysteresis loop	петля гистерезиса	Yes	No

4.2. Processing of the Lexeme/MPU Pair in the LexiComp

LexiComp starts this processing by checking whether any of the elements of the pair needs entering into the dictionaries of ETAP-3 at all. This is done by running the ETAP-3 MT system on the source and/or the target lexeme/MPU in the background. The result is compared with the counterpart element of the pair and if they coincide the LexiComp informs the user that the unit is already covered by the dictionary. If a difference has been revealed, the LexiComp starts the procedure of entering the pair into the dictionary.

Each of the two elements of the pair is consecutively processed by the morphological analyzer and the syntactic parser of the respective language. The first phase is to determine whether the word form or word forms are present in the morphological dictionary and, if the answer is yes, whether they are covered by the combinatorial dictionary.

After the dictionary status of the pair elements to be entered is established, the LexiComp offers an interactive dialogue to the user, who is instructed to choose the correct option of the parse for each word form (Fig. 1).

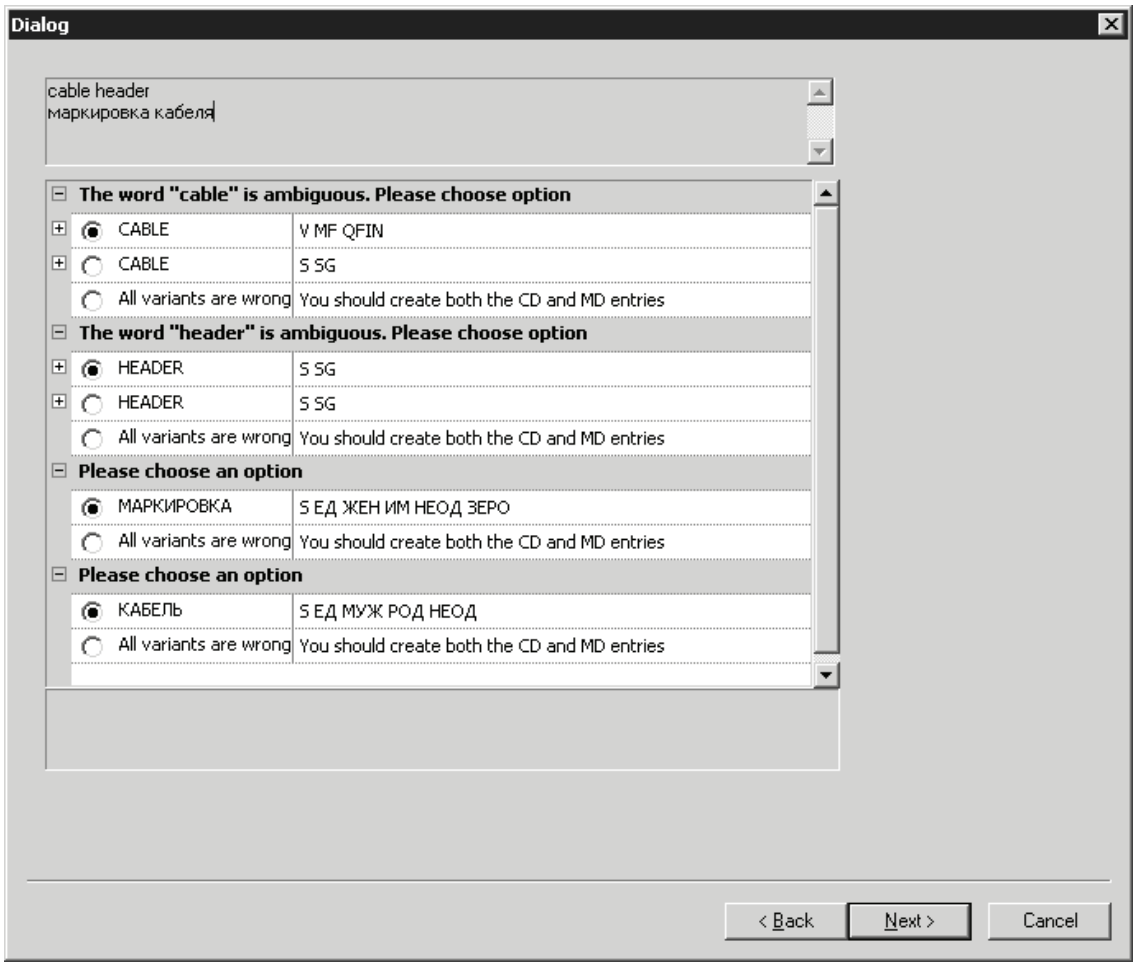


Fig.1. LexiComp interactive window for the user entering the pair
cable header ⇔ маркировка кабеля

The following options are possible for word forms submitted to LexiComp:

- No lexeme has been found that could correspond to the word form. The only possible course of action is to enter a new lexeme into the morphological and the combinatorial dictionary.
- In the morphological and/or combinatorial dictionaries one or more matches for the word form have been found. In this case, the LexiComp user may choose one of the matches or decide that none of the listed matches are appropriate for the given word form.

Further LexiComp actions depend on whether it has been decided that one or more words have to be entered into the dictionaries. In this case, the LexiComp will start entering the words (see Section

4.3 below), whereupon it will proceed with handling the translational equivalents (Section 4.4). In the opposite case, the LexiComp will start handling these equivalents at once.

4.3. Entering New Words

In most cases, the lacking lexeme has to be entered both into the morphological and the combinatorial dictionaries. We will therefore go on to describe the most typical course of action where the word is first entered into the morphological and then to the combinatorial dictionary.

Automation of morphological dictionary expansion is based on the empirical fact that, for the great majority of cases, it is sufficient to specify only a few word forms (normally, four or less) and/or morphological properties (like animacy or absence of certain ending types) of a Russian lexeme in order to determine its full paradigm. The same goes for other Slavic languages, whilst the situation with English is even simpler. So, instead of straightforward generation of the paradigm, which is extremely time-consuming and subject to error, or application of standard morphological objects, which requires special knowledge and skills, LexiComp asks the user to answer several questions easily understandable by any native speaker of the respective language.

By way of illustration, we will demonstrate LexiComp operation for the creation of a Russian morphological entry for the lexeme *пятикурсница* 'fifth-year female student' and an English morphological entry *polltracker*.

In the first case, LexiComp will start by finding out that the lexeme *пятикурсница* is not present in the Russian dictionary (and that words constituting the LFU offered are present in the English dictionary).

Then, consecutively, LexiComp will ask the user to answer the following questions: (1) what part of speech does the word to be entered belong to (Fig. 2); (2) what particular word form does Nominative Singular have (Fig. 3); (3) does the word inflect for case (Fig. 4); (4) what particular form does Accusative Plural have (Fig.4); importantly, the LexiComp automatically fills in the required slot with the probable stem of the word which in many cases, including the present one, coincides with Accusative Plural. After this, the new dictionary entry is ready and waiting to be compiled:

```
ENTRY:RU_M:пятикурсниц|а
      осн:= т:114
      trs:fifth-year_female_student
```

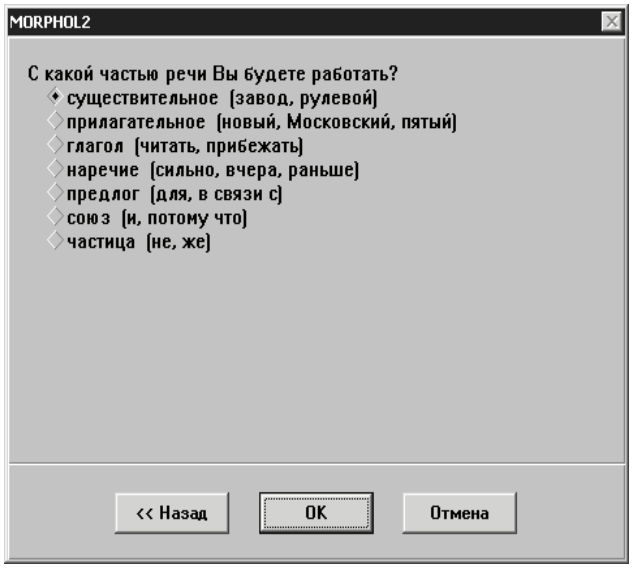


Fig. 2. LexiComp's Multiple Choice Window for Part-of-Speech Assignment

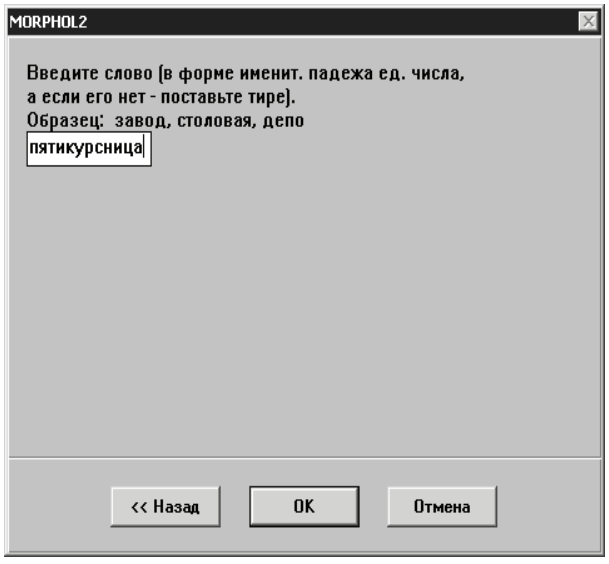


Fig. 3. LexiComp's Window Asking to Give Nominative Singular if it Exists

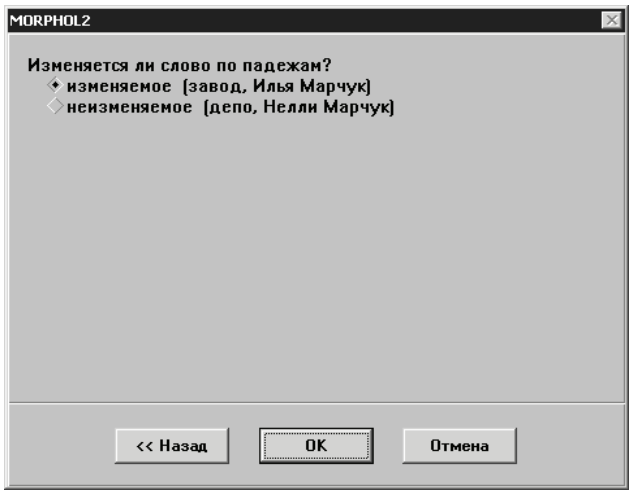


Fig. 4. LexiComp's Window Asking whether the Word Inflects for Case

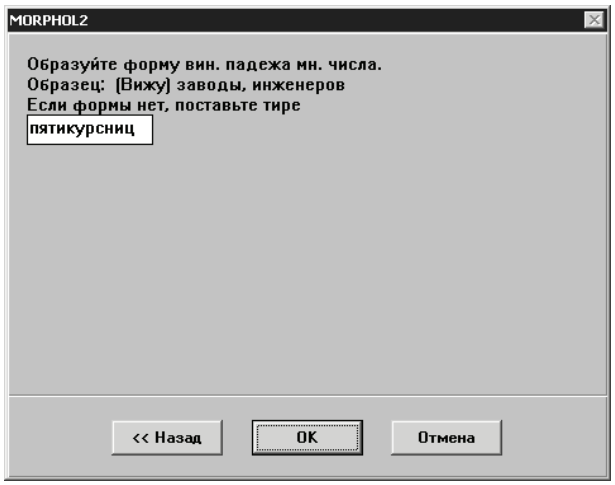


Fig. 5. LexiComp's Window Asking to Give Accusative Plural if it Exists

In the second case, LexiComp will state, likewise, that the word *polltracker* is absent from the English morphological dictionary. The user will be asked to select the part of speech (Fig. 6) and fill in additional slots asking about whether the noun lacks plural or singular (Fig. 7). This information is sufficient to create the morphological dictionary entry

ENTRY:EN_M@polltracker
bas:= f:10
trs:индикатор_выборов

Fig. 6. LexiComp's Multiple Choice Window for English Part-of-Speech Assignment

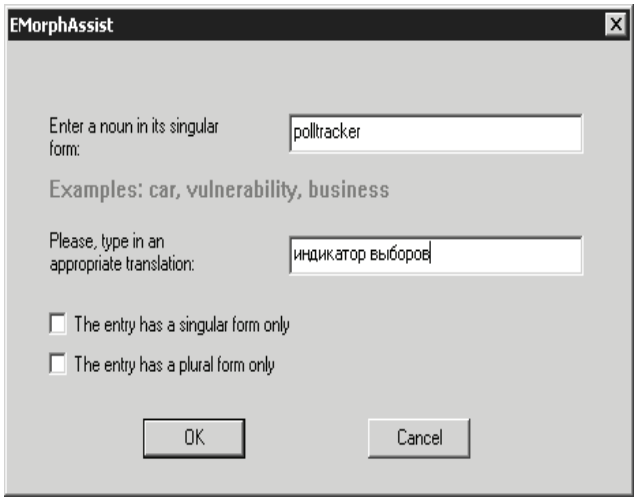


Fig. 7. LexiComp's Window Asking to Give Nominative Singular and Determine Whether the Word is a *Singularia* or *Pluralia Tantum*

In many cases, the morphological dictionary entry is used to create a partial draft entry of the combinatorial dictionary: morphological information serves as basis for predicting some of the syntactic or semantic properties of the word. To give a simple example, English nouns ending with *–zation* are likely to denote an action, have the agentive and objective valences; and be associated with transitive verbs ending with *–ize*; Russian animate nouns ending with *–ер*, *–ор, –чик, –щик* tend to denote an active agent, and the great majority of Russian verbs whose paradigms contain passive forms are transitive and may have a direct complement in the accusative case. In our example, the draft entry of the combinatorial dictionary for the word *пятикурсница* will look as follows (only the source language zones are given):

```
00000                    ПЯТИКУРСНИЦА
    POR:S
    SYNT:ОДУШ,ЖЕНСК,ИСЧИСЛ,АГЕНС
    DES:'ЛИЦО'
```

The entry contains a list of syntactic features (animateness, feminine gender, countability, and agentivity) and one semantic feature (‘person’). Of course, further editing of this word must be done manually. As a result, the entry will acquire additional features (a syntactic feature describing the word’s ability to act as an appellative, semantic features ‘woman’, ‘position’ and ‘human’, a subcategorization frame, and a list of references to three template syntactic rules). The resulting entry will look like

```
00000                    ПЯТИКУРСНИЦА
    POR:S
    SYNT:ОДУШ,ЖЕНСК,ИСЧИСЛ,АГЕНС,ОБРАЩ
    DES:'ЖЕНЩИНА','ДОЛЖНОСТЬ','ЛИЦО','ПРЕДМЕТ','ЧЕЛОВЕК'
    D2.1:РОД
TRAF:1-КОМПЛ.20
TRAF:ВВОДН.20
TRAF:АППОЗ.10
```

In many cases the words to be entered belong to particular morphological and/or lexical types (e.g. names of cities, human given names, surnames, etc.). If such a group is specified by the user, LexiComp will use simplified procedures to create new entries, resorting to fewer and simpler questions asked of the user. A typical example is processing of regular Russian surnames, for which LexiComp creates 8 entries (for the male and female variants in two languages both in the morphological and the combinatorial dictionary) just asking two simple questions. The following example lists all entries triggered by a Russian surname *Барабанщиков*: every entry of the morphological dictionary is followed by the respective combinatorial dictionary entry.

```
ENTRY:RU_M@Барабанщиков
    осн:= т:145
    trs:Barabanshchikov

ENTRY:RU@БАРАБАНЩИКОВ
    COMMENT:"ФАМИЛИЯ (МУЖСКАЯ)"
    POR:S
    SYNT:МУЖСК,АГЕНС,ОДУШ,ЗАГЛАВН,СОБСТ,ИСЧИСЛ,АДЪЕКТПЛЮС
    DES:'ПРЕДМЕТ','ЛИЦО','ФАМИЛИЯ','ЧЕЛОВЕК'
TRAF:ВВОДН.20
    **************************
ZONE:EN
    TRANS:BARABANSHCHIKOV

ENTRY:EN_M@Barabanshchikov
    bas:=  f:10
    trs:Барабанщиков

ENTRY:EN@BARABANSHCHIKOV
    COMMENT:"SURNAME (MALE)"
    POR:S
    SYNT:COUNT,AGENS,PROP
    DES:'ПРЕДМЕТ','ЛИЦО','ФАМИЛИЯ','ЧЕЛОВЕК'
TRAF:APPOS.10
TRAF:PARENTH.20
    **************************
ZONE:RU
    TRANS:БАРАБАНЩИКОВ
```

```
ENTRY:RU_M@Барабанщиков|а
    осн:= т:144
    trs:Barabanshchikova

ENTRY:RU@БАРАБАНЩИКОВА 13:44:18 02-11-2008 (00000)
    COMMENT:"ФАМИЛИЯ (ЖЕНСКАЯ)"
    POR:S
    SYNT:ЖЕНСК,АГЕНС,ОДУШ,ЗАГЛАВН,СОБСТ,ИСЧИСЛ,АДЪЕКТПЛЮС
    DES:'ПРЕДМЕТ','ЛИЦО','ФАМИЛИЯ','ЧЕЛОВЕК','ЖЕНЩИНА'
TRAF:ВВОДН.20
    **************************

ZONE:EN
    TRANS:BARABANSHCHIKOVA

ENTRY:EN_M@Barabanshchikova
    bas:= f:10
    trs:Барабанщикова

ENTRY:EN@BARABANSHCHIKOVA 13:44:18 02-11-2008 (00000)
    COMMENT:"SURNAME (FEMALE)"
    POR:S
    SYNT:COUNT,AGENS,PROP
    DES:'ПРЕДМЕТ','ЛИЦО','ФАМИЛИЯ','ЧЕЛОВЕК','ЖЕНЩИНА'
TRAF:APPOS.10
TRAF:PARENTH.20
    **************************

ZONE:RU
    TRANS:БАРАБАНЩИКОВА
```

4.4. Entering Translational Equivalents

Once all entries that constitute the pair of translation equivalents are in the morphological and the combinatorial dictionaries (irrespective of whether they were present before the LexiComp session or added during this session), LexiComp starts the concluding phase of the the work, i.e. introduction of translation equivalents themselves.

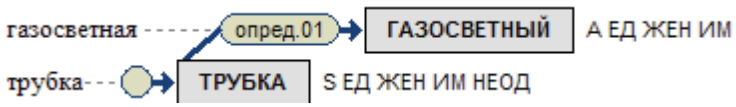
In the simplest case, i.e. when the pair consists of one source language lexeme and one target language lexeme and they both belong to the same part of speech, the only thing that remains to be done is adding the respective translations in the default translation fields of the target language zone of the dictionaries.

In more complex cases, LexiComp starts by trying to find an appropriate template translation rule (for each direction of translation) and, if successful, inserts a reference to this translation rule into one of the entries of the words constituting the element of the pair. Usually this is the syntactic head of an MPU[4]. In the great majority of cases, such translation rules require parameters (names of lexemes constituting the MPU, morphological features, etc.) which have to be identified and registered in the reference to this rule. If LexiComp cannot find a template rule, it generates a unique dictionary translation rule and adds it to the entry.

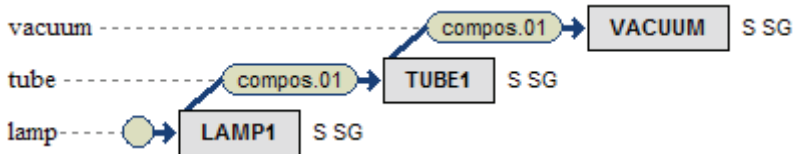
As mentioned in Section 3.2, template translation rules may be accompanied by pre-syntactic rules of priority assignment. These are also template rules determined automatically with the help of a special table of correspondences. For a unique dictionary rule, LexiComp may also generate a unique pre-syntactic priority assignment rule and add it to the entry.

We will briefly illustrate the LexiComp mechanism of template rule selection by describing the processing of translation pair *газосветная трубка – vacuum tube lamp.* The software starts by parsing both MPUs, obtaining the Russian syntactic structure

[4] A notable exception is the case of a prepositional group, whose head is the preposition: translations like *в понедельник ↔ on Monday, на Пасху ↔ at Easter, на Кубе ↔ in Cuba,* etc. are more appropriate in the entries of the nouns rather than in those of the prepositions.



and the English syntactic structure



Then the LexiComp is trying to find a template Russian-English translation rule that accepts a configuration

$$\text{Adjective} \Rightarrow \text{Noun} \quad (1)$$

and translates it into a configuration

$$\text{Noun} \Leftarrow \text{Noun} \Leftarrow \text{Noun} \quad (2)$$

This is achievable by a specific template translation rule oriented to these types of configuration, which should be referred to in the entry of the word *трубка*:

00000 ТРУБКА
……..
TRAF:RA-TRADUCT2.81
LR:ГАЗОСВЕТНЫЙ,LA1:LAMP1,LA2:TUBE1,LA3:VACUUM

It can be seen that the reference mentions four lexical parameters: one Russian word and three English words.

The reverse translation is achievable by a template English-Russian translation rule that accepts configuration (2) and translates it into a configuration (1). The respective rule is entered into the entry of the word *lamp:*

00000 LAMP1
……..
TRAF:TRADUCT2.08
LA1:TUBE1,LA2:VACUUM,LR1:ТРУБКА,LR2:ГАЗОСВЕТНЫЙ

This reference also mentions four lexical parameters: two English words and two Russian words.

Technically, this is achieved due to the fact that every template translation rule is supplied with a special recognition zone (RECOG), which describes additional requirements to the source word combination to which the template rule has to be applied, and requirements to the target word combination (a fragment of syntactic structure obtained after this rule has been applied).

5. Future Prospects

The experience that the ETAP-3 developers have gained from using the Lexicographer's Companion shows that the system increases the lexicographer's output and precision. It is especially important when specialized entries are produced on a mass scale. We are planning therefore to extend the set of specialized lexicographic types in the nearest future. We are also planning to improve the choice of correct parses of translational equivalent pairs taking full account of those cases where the lexemes as elements of those pairs stand in the basic form. We hope that this software system may be of much help if applied to a multilingual lexicographic resource.

6. References

Зализняк А.А. (2003). Грамматический словарь русского языка. М., «Русские словари».
Мельчук И.А. (1974). Опыт теории лингвистических моделей "Смысл <=> Текст". М.: Наука.
Мельчук И.А., Жолковский А.К. (1984). Толково-комбинаторный словарь современного русского языка. Wiener Slawistischer Almanach, Sonderband 14, 992 p.

Storing Morphology Information in a Wiki[1]

Radovan Garabík
Ľ. Štúr Institute of Linguistics
Slovak Academy of Sciences
813 64 Bratislava, Slovakia
korpus@korpus.juls.savba.sk, http://korpus.juls.savba.sk

Abstract
There are different ways of storing information about morphology. We describe the way of organizing morphology data in a form suitable to be kept as plain text files inside of a MoinMoin wiki engine and the practical results of keeping information about Slovak morphology.

***Keywords*:** morphology analyser, wiki, MoinMoin, tagset, part of speech, scalability

Introduction

We successfully developed the Slovak language morphology analyser based on Levenshtein edit operations (Garabík, 2006; Левенштейн, 1965). The original aim was to cover all the words present in the Short Dictionary of the Slovak Language (KSSJ) (2003) and some additional frequent words. The Levenshtein edit operation based paradigm classes proved very useful for quick semi-automatized construction of all the word forms derived from a given lemma. However, as the number of words included reached the originally projected goal, the project entered a maintenance phase with new words being added only sporadically with focus towards long term storage and reutilization of the data, and consequently a new approach appeared to be desirable.

Wiki Software

We settled on the MoinMoin (http://moinmo.in) wiki engine (presently we are using version 1.6.3). MoinMoin is a wiki written completely in the Python programming language (http://www.python.org) using plain text files as a storage backend rather than a database. This makes it particularly attractive for our needs 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 and I/O are 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, we extended the default parser to display the morphology tags in better human-readable forms (with explanation of different grammar categories), while keeping the original data intact and in a terse form suitable for computerized parsing.

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.

Our MoinMoin server runs on a modest Intel Pentium 4 server with 2 GB of RAM, two IDE disks assembled into a RAID 1. The operating system is GNU/Linux (Ubuntu Hoary) and the wiki data is kept on a ReiserFS filesystem.

Data Structure

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

The basic unit of the wiki data is called a page (using MoinMoin terminology). Each page contains data pertaining to one lexeme, i.e. a lemma with full paradigm and morphology annotation. Each page

[1] The study and preparation of these results have received funding from the EC's Seventh Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX

name is equal to the lemma taking into account common capitalization of words in Slovak (proper nouns)[2]. In the case of lexical homonymy, pages are named by the lemmas with the part of speech tag attached in parentheses[3].

We strived to keep the page structure to be both human-readable and human-editable as well as being easy to parse automatically.

The page body has a form like this:

```
== Lema ==
ucho

== Paradigma ==
SSns1: ucho
SSns2: ucha
SSns3: uchu
SSns4: ucho
SSns5: ucho
SSns6: uchu
SSns7: uchom
SSnp1: uši, uchá
SSnp2: úch, ušú, uší
SSnp3: ušiam, uchám
SSnp4: uši, uchá
SSnp5: uši, uchá
SSnp6: uchách, ušiach
SSnp7: ušami, uchami
----
[[Kategória:Substantíva]]
```

Text 1: Example of a wiki page (lemma *ucho*)

The page body contains several sections. The first one is the *Lema*, which contains just one word, the lemma. Then the *Paradigma* section follows, containing the inflectional paradigm spelt out in full. For each grammar category there is one corresponding line, with the 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

We are talking here about the basic homonymy only, where lemmas for two different words (two different parts of speech) are identical. 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, we create a special disambiguation page linking to all the possible lemmas.

[2] An important point, because by design the final morphology analyser disregards the capital letters and gives all the lemmas in lowercase.

[3] E.g. *mať_(V)* for a verb, *mať_(S)* for a noun.

```
== Lema ==
mať

== Pozri ==
[[mať_(S)]] [[mať_(V)]]
----
[[Kategória:Dezambiguácia]]
```

Text 2: Example of a disambiguation wiki page (lemma *mať*)

Reflexive Verbs

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[4] (Dvonč et al., 1966). As there exists a reflexive/non-reflexive dichotomy (i.e. reflexive verbs almost always have their non reflexive counterpart), we decided to keep 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 – we could denote them equally good as particles, indeed this is how they are sometimes classified in the traditional Czech grammars. We simplified our task by assigning the *sa* and *si* a special morphology tag **R**, regardless of their semantic use.

Part of Speech Distribution

Currently, the wiki contains 77567 entries. Categorised by the POS type, we have the following distributions:

[4] Unlike other languages, e.g. in Russian the reflexive pronoun/particle takes a form of a clitic inseparably bound to the verb.

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[5]
26	part of multiword expression[6]
2	*sa/si*
1	*by*[7]
716	disambiguation pages

Table 1: Distribution of parts of speech

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 underlaying file system to be able to cope with many thousands of entries per directory. All the major modern Linux file systems have no problems with this usage pattern (Piszcz, 2006). Probably the best file system for these purposes at the moment is ReiserFS, which also has 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. The total size of our data is 1.2 GB of disk storage.

Basic usage works well, and 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 searches for an inflected word form typically takes many long 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 the 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).

[5] "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 our wiki, only a few such words are included.

[6] Used to mark standalone morphemes that are a part of multiword expressions – these are in fact just a remnant of our tokenization.

[7] Special conditional morpheme, traditionally classified as a particle.

Usage

The wiki can be used directly as a reference dictionary of inflectional data. However, we are using it mostly as a source of data for a morphology analyser, transforming the data from the wiki into constant database tables for quick retrieval (http://cr.yp.to/cdb.html), further independent of the wiki software.

We also convert the data into a nicer looking format for the DICT server (RFC 2229) for a quick web-based search, integrated with several other Slovak language dictionaries (Faith, Martin, 1997).

Conclusion

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 number of pages such as Category pages – in our wiki containing just a static description of each part of the speech category, not the list of all pages belonging to a given category. The wiki is used as a source of data for various morphology related automatized tasks, as well as a source for a human-readable dictionary of Slovak morphology.

References

Dvonč, L. et al. (1966). Morfológia slovenského jazyka. Vydavateľstvo Slovenskej akadémie vied, Bratislava. pp. 377–388.

Faith, R., Martin, B. (1997). "A Dictionary Server Protocol", Request for Comments 2229, Network Working Group, ftp://ftp.isi.edu/in-notes/rfc2229.txt

Garabík, R. (2006). Slovak morphology analyzer based on Levenshtein edit operations. In: Proceedings of the WIKT'06 conference, pp. 2—5. Bratislava, Slovakia.

Krátky slovník slovenského jazyka. Ed. M. Považaj. Veda, Bratislava, 2003.

Piszcz, J. (2006). Benchmarking Filesystems Part II. Linux Gazette, 122.

Левенштейн, В. И. (1965). Двоичные коды с исправлением выпадений, вставок и замещений символов, Докл. АН СССР, 163, 4, pp. 845—848.

Web References

1. http://moinmo.in
2. http://www.python.org
3. http://www.xapian.org
4. http://cr.yp.to/cdb.html

Bulgarian Language Resources for Information Technology[1]

Kiril Simov, Petya Osenova
BulTreeBank Project
http://www.BulTreeBank.org
Linguistic Modelling Laboratory, Bulgarian Academy of Sciences
Acad. G. Bonchev St. 25A, 1113 Sofia, Bulgaria
kivs@bultreebank.org, petya@bultreebank.org

Abstract

In the paper we will discuss the Bulgarian language resources infrastructure developed in several projects to support information technology applications in contrast to supporting Bulgarian language studies. In many aspects the two usages overlap, but our aim is to cover the necessary resources for real applications. In our work we follow two main principles: (1) minimization of construction effort – use and reuse of language tools and resources for creation of other resources; and (2) taking pragmatic decisions where linguistics can not provide solutions. We report on resources that we have already implemented and resources which are under development. These resources are considered as elements of Basic Language Resources Kit (BLARK) for Bulgarian.

Keywords: Basic Language Resource Kit, treebank, Bulgarian, language resources, NLP applications.

1. Introduction

One of the central questions within Human Language Technologies discusses "what is minimally required to guarantee an adequate digital language infrastructure for a language"? (Bimnnenpoorte et al. 2002). Thus the notion of Basic Language Resources Kit (BLARK) was introduced and discussed within NLP community. Its definition and scope have been considered in several European initiatives; see ENABLER Network, Dutch LT Platform, and CLARIN among others. BLARK is defined as a set of three distinct groups: *applications*, *processing modules* and *language data*. (Strik et al. 2002).

This paper aims at localizing the BulTreeBank (a project devoted to the creation of an HPSG-based treebank of Bulgarian) language resources for Bulgarian within the notion of BLARK. Several problematic issues are addressed:

- How close are the language resources (LRs) to the BLARK requirements?
- How can a more advanced resource like a treebank give rise to a number of basic language resources, which lack in this language?
- How can the existent LRs be turned into a solid basis for the development of other LRs?
- Is it always the case that basic language resources are produced first, and the more advanced ones afterwards?
- How to test the resources in real application?

We consider the creation of such a complex language resource an application which tests the availability of other resources and processing modules. During the project we have discovered the *white spaces* in the resources for Bulgarian. We have been trying to fill these gaps developing LRs with respect to the actual work on the treebank. However, in this creation we have not restricted ourselves to the needs of the treebank development only, but we also have envisaged wider range of NLP tasks, such as information extraction, grammar checker and parsing.

2. Treebanking as Basic Language Resources Compiler

The creation of a treebank for a "less-spoken" language like Bulgarian imposes certain challenges due to the limited scientific, technological and financial resources. As a central task we considered the organization of the work with respect to the minimization of human intervention and the achievement

[1] The study and preparation of these results have received funding from the EC's Seventh Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX.

of the project goal. The greatest problem appeared to be the lack of already available set of language resources, which to serve as a base for the treebank compilation. Thus, on the one hand, we were aware before the start of the project that most of the required resources had to be produced by us. On the other hand, we have used the situation as a possibility to construct a variety of resources to support the creation of the treebank and to be extensively tested within the project. As a result, we have produced a basic set of language resources for Bulgarian, which are easily adaptable for different mono- and multilingual NLP tasks.

Generally, two strategies have been applied:

1. *Before starting the treebank creation*, we have implemented basic processing modules: tokenization, morphological analyzer, disambiguator, named entities recognition modules, partial grammars, the text corpus.
2. *Parallel to the treebank creation*, we have compiled resources, which need more elaborate and high quality information: specific lexicons of verb frames, lists of fixed phrases, specific introductory patterns for newspaper texts, parenthetical expressions.

Note that the time distinction (before and parallel to the treebank creation) is a relative one, because all the primary resources have been further developed and tested.

The creation of the resources is governed by two principles:

Bootstrapping principle: Its aim is to obtain as much information as possible at the very basic processing levels. For instance, in the tokenization case, according to "a general token classification" (Osenova and Simov 2002), the tokens are divided into the potential classes of common words, abbreviations, names etc.

Corpus-driven principle: Several results are simultaneously obtained by using extraction and observation procedures: the gazetteers are compiled, the dictionaries are improved and the tools are tested against unrestricted data.

The creation of our resources has always been in close connection to the overall annotation process of data. It comprises the following steps:

Sentence/Text Extraction from the Corpus

The source of the sentence extraction is the BulTreeBank text corpus (100 mln. running words at the moment). We aimed at sentences with different lengths and from different genres. Sentence extraction was combined with text extraction, which means that whole newspaper articles or book chapters have been selected for annotation. As supporting modules, the CLaRK concordancer and grammar engine have been relied upon.

Automatic Pre-processing

Each sentence needs first to be pre-processed at all the levels, that precede deeper syntactic annotation. These include: (1) Morphosyntactic tagging; (2) Named entity recognition; (3) Morphosyntactic disambiguation; (4) Partial parsing (chunking). We aim at a result of a 100 % accurate partial parse of a sentence. The accuracy is checked and validated by a human annotator with the assistance of the CLaRK System (Simov et al. 2001).

HPSG Step

The result from the previous step is encoded into an HPSG compatible representation. Then HPSG parsing takes place. The output is encoded as a parse forest.

Resolution Step

The parse selection is performed by supplying partial information and navigation in the parse forest. However, relevant for this paper is the first step, because the pre-processing module comprises most of the basic LRs. The result is manually checked, the lexicons are extended to cover the tokens within the corpus, the phenomena with bigger impact over the corpus are considered.

Applying the above principles and the steps of annotation, we have created a set of basic resources and tools for Bulgarian. Later we have used them in other projects for implementation of systems that

need language technologies. Such systems are: BulQA – system for Bulgarian Question Answering; LT4eL system (within LT4eL project[2]) to support semantic annotation for eLearning tasks; AsIsKnown system (within AsIsKnown project[3]) to support semantic annotation for domain text mining and within LTfLL for semantic annotation[4]. In fact, using of the resources created and tested during compilation of the Bulgarian treebank was a further step of creation of additional basic resources and tools as it will be described below.

3. The BulTreeBank LRs in the Context of BLARK

Language technology is supposed to include the following modules: tokenization and named entities recognition, morphological analyzer and disambiguator, syntactic and semantic analyzer. The data is supposed to consist of: a mono-lingual lexicon, annotated corpus of texts (a treebank with syntactic, morphological and semantic structures) and benchmarks for evaluation. Within this BLARK notion a priority list can be proposed depending on what exists and what needs development in a certain language. In our case the priority was to create all the complementary resources which would support the treebank creation.

3.1. BulTreeBank Language Technology

Tokenization

There is a hierarchy of tokenizers within the CLaRK system, which tokenize the texts in an appropriate way. Additionally, the user can decide what the category of the token is and to assign it.

Morphosyntactic analyzer

It assigns all possible analyses to the word tokens. The lexicon is too large to be loaded as one grammar in CLaRK and this is why we have divided it into several grammars which are applied simultaneously. The separation of the lexicon is on the basis of the frequencies of the word forms within the corpus. In this way the application has been sped up. As it was mentioned above, together with the morphosyntactic analyzer we use the gazetteers. They are also implemented within the CLaRK system. In the places where competing analyses arise between a common word and a name or an abbreviation, we try to use the token classification strategy and the prompts of the context. If there is no clear preference, we leave the decision to the human annotator.

MorphoSyntactic Disambiguation

We have already implemented a rule-based morphosyntactic disambiguator, encoded as a set of constraints within the CLaRK system. This rule-based disambiguator exploits context information like *agreement between an adjective and a noun in a noun phrase*, specific positions like *a noun after a preposition*, but it also deals with some fixed phrases. The disambiguator does not try to solve unsure cases, but leaves them for further processing. Its coverage is about 80 %. For next step of disambiguation we have developed a neural-network-based disambiguator (Simov and Osenova 2001). It achieves accuracy of 95.25 % for part-of-speech and 93.17 % for complete morphosyntactic disambiguation.

Partial Grammars

We have constructed such grammars for:

- *Sentence splitting.* At the moment it is fully automated and reliable only for the basic and clear cases. For solving complex and ambiguous cases this grammar is combined with supporting modules for abbreviation detection.
- *Named-entity recognition*. Identifying numerical expressions, names, abbreviations, special symbols (Ivanova and Doikoff 2002) and (Osenova and Kolkovska 2002). They are designed to work in cooperation with the morphosyntactic analyzer. If necessary, the grammars can overwrite the analysis of the morphosyntactic analyzer.

[2] http://www.lt4el.eu/
[3] http://www.asisknown.org/
[4] http://partners.ltfll-project.org/

- *Chunking*. Two basic modules have been developed: an NP chunker (Osenova 2002) and (Osenova and Kolkovska 2002). Generally speaking, the chunking process conforms to the following requirements: it deals with non-recursive constituents; relies on a clear-indicator strategy; delays the attachment decisions; ignores the semantic information; aims at accuracy, not coverage. Additionally, there are chunk grammars for APs, AdvPs, PPs and some non-problematic clauses.

3.2. BulTreeBank Language Data

Text archive

It is intended to yield the size of a national corpus, that is, 100 million words. In order to compile a representative and balanced corpus of Bulgarian texts, we tried to gather a variety of different genres: 15 % fiction, 78 % newspapers and 7 % legal texts, government bulletins and others.

Morphologically annotated corpus

Over 1 000 000 running words are morphosyntactically manually disambiguated. This part of the text archive is used in two ways within the project: (1) as a source of sentences and articles which to be annotated syntactically and included in the treebank, and (2) as training and testing data for POS disambiguation of Bulgarian texts.

Treebank

Currently the BulTreeBank (Simov et al. 2005; Simov and Osenova 2003) comprises 215 000 tokens, a little more than 15,000 sentences. Each token is annotated with elaborate morphosyntactic information. The BulTreeBank is based on HPSG. Syntactic structure is encoded using a set of constituents with head-dependent markings. The phrasal constituents contain two types of information: the domain of the constituent (*NP*, *VP*, etc.) and the type of the phrase: head-complement (*NPC*, *VPC*, etc.), head-subject (*VPS*), head-adjunct (*NPA*, *VPA*, etc.). In every constituent the head daughter could be determined unambiguously. Coordinations are considered to be non-headed phrases, where the grammatical function overrides the syntactic labels.

Morphological Dictionary

The dictionary is an electronic version of a paper dictionary (Popov, Simov and Vidinska 1998) extended with new words from the corpus. It covers the grammatical information of about 100 000 lexemes (1 600 000 word forms) and serves as a basis for the morphological analyzer.

Gazetteers

Three basic lists with items, missing in the morphological dictionary, have been compiled with respect to their frequency:

Gazetteers of names. These consist of 25 000 items and include Bulgarian as well as foreign person names, international and national locations, organizations. The most frequent names are additionally classified according to three criteria: (1) grammatical (gender and number); (2) semantic - with respect to ontology (names for different types of locations, organizations, artifacts, persons' social roles etc.) and (3) ontological – some person names were connected with specific individuals in the world and thus some encyclopedic information was provided in addition to the semantic classification.

Gazetteers of the most frequent abbreviations. They consist of 1500 acronyms and graphical abbreviations. The acronyms' extensions were mapped against the names (mostly organizations) and therefore, assigned the same semantic and grammatical label. In cases of idiosyncratic grammatical behaviour, the relevant patterns have been added as well.

Gazetteers of the most frequent introductory expressions and parentheticals. This is considered to be a step towards a basic list of collocations. They were classified according to their morphological type or behavior: verbal, adverbial, linking (for conjunctions), nominal (vocatives), idiomatic etc. We use them as an extended supplementary lexicon during the phase of the syntactic annotation.

Valence Dictionary

It consists of 1000 most frequent verbs and their valence frames and it is based on a paper dictionary. Each frame defines the number and the kind of the arguments and imposes morphosyntactic and semantic restrictions over them. The semantic restrictions over the arguments are extracted and matched against ontology.

Semantic Dictionary

Semantic information plays a crucial role in the process of parse discrimination on which the construction of our treebank depends. Thus, in order to support the selectional restrictions imposed by the valence dictionary and to facilitate its usage, we decided to compile a semantic dictionary related to ontology.

The main strategy we have adhered to in our work is the preparation of a minimum set of resources with substantial impact over the text archive.

4. Usage of BulTreeBank HLT in Other Projects

In this section we present the application of the above tools and their extensions in two tasks necessary for construction of question answering system for Bulgarian and for semantic annotation of Bulgarian text.

4.1. Bulgarian Question Answering System

Since 2004 we are responsible for preparation of Bulgarian data for CLEF initiative. These data include a newspaper corpus consisting of 18000 articles, question-answer pairs (500) and topic descriptions (200). The questions and topics ensure compatibility with corpora in other languages supported by CLEF, Thus the data could support monolingual Bulgarian-Bulgarian question answering and information retrieval, but also cross-lingual task such as Bulgarian-English question answering. In this section we describe the language analysis of question in the question answering system BulQA in which Bulgarian is the source language (the language of the questions) and the target language (the language of the answers) could be Bulgarian or English. The described processing is to demonstrate the facility of the resources we described above in such a system. BulQA and the preparation of the necessary data are described in (Osenova et al. 2005) and (Simov and Osenova 2006).

Although the above listed language processing tools were extensively tested during the compilation of our treebank, they needed some additional tuning to the task of question analysis. The main difference is that most of them were implemented in such a way that in unsure cases the ambiguity remained unresolved or the analysis was not produced. This tools' application was required when an annotator had to inspect the result of the processing.

With respect to the Question Answering task some ambiguities were resolved in the following way: (1) in ambiguities between 2nd and 3rd person or 1st and 3rd person, always the 3rd person was selected; (2) in ambiguities between present and past verb tense, the past tense was selected, etc. The first ambiguity was resolved because the questions given in CLEF are never in 1st or 2nd person. Resolving between the different tenses in the question with respect to validation of the found answers is not currently supported by the Answer extraction module. Some other ambiguities were resolved on a frequency basis only – for each ambiguity class the most frequent option was selected.

The major addition with respect to the available tools was the construction of a lemmatizer for Bulgarian. We defined the lemma to be functionally determined by the wordform and its morphosyntactic characteristics. The cases of ambiguous lemmas were not resolved and all possible lemmas were assigned to the corresponding wordform. Lemmas are also used later to access the semantic information from the semantic dictionary and the English equivalents in the Bulgarian-English dictionary.

Here is an example of the analysis of the question “През коя година Томас Ман получи Нобелова награда?” “*Prez koya godina Tomas Man poluchi Nobelova nagrada*?” (in English: *Which year did Thomas Mann receive the Nobel Prize*?):

```
<analysis group="BTB">
  <PP>
    <Prep><w ana="R" bf="prez">Prez</w></Prep>
    <NPA>
      <Pron><w ana="Pie-os-f" bf="koya">koya</w></Pron>
      <N><w ana="Ncfsi" bf="godina">godina</w></N>
    </NPA>
  </PP>
  <NPA sort="NE-Pers">
    <N><name ana="Npmsi" sort="PersNE">Tomas</name></N>
    <H><name ana="Hmsi" sort="PersNE">Man</name></H>
  </NPA>
  <V><w ana="Vpptf-o3s" bf="polucha">poluchi</w></V>
  <NPA>
    <A><w ana="Hfsi" bf="nobelov">Nobelova</w></A>
    <N><w ana="Ncfsi" bf="nagrada">nagrada</w></N>
  </NPA>
  <pt>?</pt>
</analysis>
```

Here each common word is annotated within the following XML element *<w ana="MSD" bf="LemmaList">wordform</w>*, where the value of attribute *ana* is the correct morphosyntactic tag for the wordform in the given context. The value of the attribute *bf* is a list of the lemmas assigned to the wordform. Names are annotated within the following XML element *<name ana="MSD" sort="Sort">Name</name>*, where the value of the attribute *ana* is the same as above. The value of the attribute *sort* determines whether this is a name of a person, a location, an organization or some other entity.

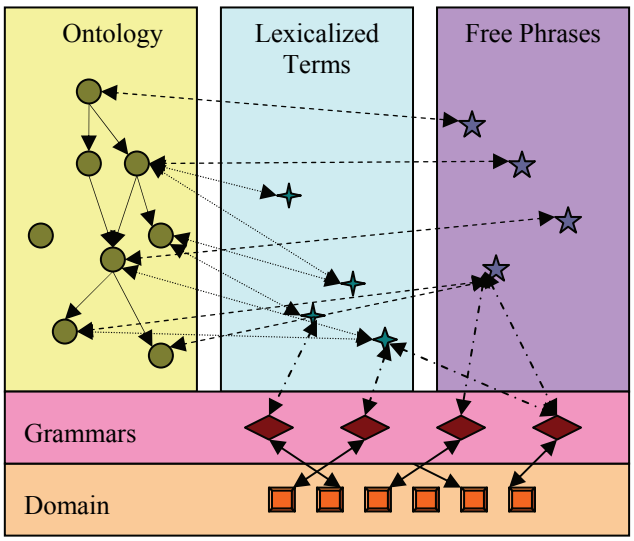
The next level of analysis is the result of the chunk grammars. In the example there are three *NPA* elements (NPA stands for a noun phrase of head-adjunct type) and one *PP* element. Also, one of the noun phrases is annotated as a name with a sort attribute with value: *NE-Pers*.

The result of this analysis had to be translated into the format which the answer extraction module uses as an input.

The corpus from which the answers are extracted is processed in similar way. Thus, the real work is in connection with the answer extraction procedure and answer validation.

4.2. Semantic Annotation of Bulgarian Text

Within three European projects (LT4eL, AsIsKnown, LTfLL) we have developed an Ontology-to-Text relation which supports the semantic annotation of domain texts and semantic search. In this section we present briefly the linguistic model of the Ontology-to-Text relation. We assume that the ontology is the repository of the lexical meaning of the language. Thus, we have started with a concept in the ontology and we searched for lexical items and non-lexical phrases that convey the content of the concept. There are two possible problems here: (1) there is no lexical item for some of the concepts in the ontology, and (2) there are lexical items in the language without a concept representing the meaning of the lexical item in the ontology. The first problem is overcome by allowing in the lexicon also non-lexical (fully compositional) phrases to be represented. The second problem is solved by extension of the ontology. The lexicon items are then mapped to grammars, we call them concept annotation grammars. These grammars relate the lexicon to the text. This mapping is necessary as much as lexical items and phrases from the lexicons allow for multiple realizations in the text and require some additional linguistic knowledge in order to disambiguate between different meanings of some lexical item or phrase. The following figure depicts the elements of the model.



We have been using the relations between the different elements for the task of ontology-based search. The connection from ontology via lexicon to grammars is relied on for the concept annotation of the text. In this way we established a connection between the ontology and the texts. The relation between the lexicon and the ontology is used for definition of user queries with respect to the appropriate segments within the documents.

In order to implement this relation for Bulgarian we have reused the following components: (1) tokenization; (2) morphosyntactic annotation; (3) lemmatization; (4) chunk grammar. The lexical items from the ontology-based lexicons are lemmatized and then concept annotation grammars are constructed on the basis of the lemmatized lexicon. Concept annotation grammars are also connected to the general chunk grammar in order to recognize the same chunks where there are overlappings of the two grammars (the general and the domain ones). The disambiguation is done statistically on the basis of the local context of each ambiguous lexical item in the lexicon. Thus we reused all the elements of the BulTreeBank language infrastructure. In future work we are planning to use the general semantic lexicon and the frame lexicon as additional means supporting the disambiguation of the ambiguous terms in the ontology related lexicon.

These short examples demonstrate that the language infrastructure for Bulgarian, created within BulTreeBank Project, is ready to be used in real applications. The developers of these applications need to concentrate on the real new functionalities and language processing.

5. Conclusion and Outlook

The intensive work within a project for creation of a treebank for a less-spoken language pays off in several ways:

Practical

It triggers the creation of LRs which still lack for the certain language. Consequently, the created set of LRs minimizes the work during the actual annotation of sentences within the treebank and ensures a high quality result.

Another advantage is that the developed resources and processing modules have a natural environment for intensive testing and improvement. This guarantees their appropriateness and adaptability for other NLP applications.

Also these LRs are a reliable basis for further development of the resources and processing modules included in BLARK.

Theoretical

It raises the question about the ways of LRs creation. It turns out that there are two ways: (1) starting from basic tasks and after their completion pursuing next-level tasks of complexity or (2) having in mind some more complex task and compiling all the other basic resources in order to

adequately face it. We believe that the latter is an appropriate way for a less-spoken and less-processed language to come up with the state-of-the-art LRs in the natural languages.

The worked out methodology for the creation of basic language resources is implemented in the CLaRK system as reusable modules and can be parameterized to other languages as well. We have used these modules in real applications such as question answering and semantic annotation and search.

References

Bimnnenpoorte, Cucchiarini, D'Halleweyn, Sturm and De Vriend. (2002). *Towards a roadmap for Human Language Technologies: Dutch-Flemish experience*, LREC-2002.

Ivanova, K. and Doikoff, D. (2002).*Cascaded Regular Grammars and Constraints over Morphologically Annotated Data for Ambiguity Resolution.* In: Proc. of The Workshop on Treebanks and Linguistic Theories. Sozopol, Bulgaria.

Osenova, P. (2002). *Bulgarian Nominal Chunks and Mapping Strategies for Deeper Syntactic Analyses*. In: Proc. of the Workshop on Treebanks and Linguistic Theories. Sozopol, Bulgaria.

Osenova, P. and Simov, K. (2002). *Learning a token classification from a large corpus. (A case study in abbreviations).* In: Proc. of the ESSLLI Workshop on Machine Learning Approaches in Computational Linguistics, Trento, Italy.

Osenova, P. and Kolkovska, S. (2002). *Combining the named-entity recognition task and NP chunking strategy for robust pre-processing.* In: Proc. of The Workshop on Treebanks and Linguistic Theories. Sozopol, Bulgaria.

Osenova P., Simov, A., Simov, K., Tanev, H. and Kouylekov, M. (2005). *Bulgarian-English Question Answering: Adaptation of Language Resources.* Lecture Notes in Computer Science 3491. pp. 458-469.

Popov, D., Simov, K. and Vidinska, S. (1998). *A Dictionary of Writing, Pronunciation and Punctuation of Bulgarian Language*. Atlantis LK, Sofia, Bulgaria.

Simov, K., Peev, Z., Kouylekov, M., Simov, A., Dimitrov, M., Kiryakov, A. (2001). *CLaRK - an XML-based System for Corpora Development.* In: Proc. of the Corpus Linguistics 2001 Conference. pp 558-560.

Simov, K. and Osenova, P. (2001). *A Hybrid System for MorphoSyntactic Disambiguation in Bulgarian.* In: Proc. of the RANLP 2001, Tzigov chark, Bulgaria.

Simov, K. and Osenova, P. (2003). *Practical Annotation Scheme for an HPSG Treebank of Bulgarian*. Proceedings of the 4th International Workshop on Linguistically Interpreteted Corpora (LINC2003), Budapest, Hungary. pp.17-24.

Simov, K., Osenova, P., Simov, A. and Kouylekov, M. (2005). *Design and Implementation of the Bulgarian HPSG-based Treebank.* Journal of Research on Language and Computation - Special Issue 2:4.pp. 495-522.

Simov, K. and Osenova P. (2006). BulQA: *Bulgarian–Bulgarian Question Answering at CLEF 2005.* Lecture Notes in Computer Science 4022. pp. 517-526

Strik, Daelemans, Bimnnenpoorte, Sturm, De Vriend, Cucchiarini. (2002). *Dutch HLT Resources: From BLARK to Priority lists*. ICSLP-2002.

III. Specific and Universal Linguistic Phenomena: How to Treat them in Multilingual Environment

The Category of Predicatives in the Light of the Consistent Morphosyntactic Tagging of Slavic Languages[1]

Ivan Derzhanski, Natalia Kotsyba
Department for Mathematical Linguistics, Institute for Mathematics and Informatics,
Bulgarian Academy of Sciences, Sofia,
Institute of Slavic Studies, Polish Academy of Sciences, Warsaw
iad58g@gmail.com, gnatko@gmail.com

Abstract

This paper presents an overview of the history of the category of non-verb predicatives, its definition and coverage in grammars, dictionaries and corpora of four Slavic languages (Russian, Ukrainian, Polish and Bulgarian). It is shown that all accounts which single out predicatives as a separate part of speech are built upon inconsistent choices of criteria, and it is argued that predicatives should be treated as a class of adverbs, for the sake of greater consistency.

Key words*:* adverbs, classifiers, parts of speech

1. Parts of Speech: How Many?

Since the early days of digitalising Slavic language resources and attempts at organising grammatical information in them, scholars in different countries have been approaching the question of classifying lexical-grammatical categories in varying ways. Most adhere on the whole to the traditional division into Aristotle's ten parts of speech, predominant in academic and school grammars: 10 in the Czech National Corpus, about 14 (depending on the status of some subcategories) in the Ukrainian Lingua-Information Fund corpus of Ukrainian, 16 in the Russian National Corpus. In some cases much further granulation is provided: in the Czech National Corpus the 10 parts of speech are divided in turn into 75 sub-parts of speech, each with its own code. A different approach to classification was used in the IPIPAN corpus of Polish, where 29 so-called flexemes[2] are recognised (see also Kotsyba et al. 2008).

The only attempt known to us to create a common morphological tagset for numerous languages including several Slavic ones was made within the frame of the project MULTEXT-East (Multext-East 1998), where altogether 14 parts of speech were differentiated. This tagset does not, however, include East Slavic languages or Polish, which differ considerably in their present independent corpus presentations. As the relatively simple example of derogatory nouns in Polish (see fn. 1) shows, even when categories coincide in their names, they do not necessarily cover equivalent sets of words. It is clear that we need compatibility and coherence (within each language and between languages) in order to present grammatical information in a consistent way, both for theoretical comparative studies of Slavic languages and even more so for multi-language morphosyntactic tagging, corpus construction and lexicography.

[1] The study and preparation of these results have received funding from the EC's Seventh Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX.

[2] The category of flexeme was introduced by Janusz Bień in the 1960's as a class of words with uniform morphosyntactic characteristics. For example, a subclass of **derogatory nouns** (masculine nouns denoting human beings but pluralised as non-human nouns to express disparagement, e.g., derogatory *profesory* vs neutral *profesorowie* 'professors') was singled out as a separate flexeme different from the noun as characterised by a unique plurality patterns, different from common nouns. From the traditional grammar point of view, flexemes are often a purely technical class, which is however convenient for automated language processing, as those classes are comparatively strictly defined.

1.1. Predicatives: a Problematic Category

The so-called predicatives constitute one of the categories where descriptions of Slavic languages differ conspicuously. They appear in Polish, Russian and Ukrainian electronic resources. They can be recognised by a common set of core words, although they are called by different names from grammar to grammar and from language to language: **improper verbs** (*czasowniki niewłaściwe*) or **predicatives** (*predykatywy*) in Polish, **adverbialised words** (*адвербіалізовані слова*) or **predicative words** (*присудкові слова*) in Ukrainian, **category of state** (*категория состояния*), **impersonal predicative words** (*безлично-предикативные слова*), **non-verb predicatives** (*неглагольные предикативы*) or **predicative adverbs** (*предикативные наречия*) in Russian. Some Bulgarian and Czech grammars talk of predicative adverbs or of predicatives as a subset of adverbs. Contrariwise, corpora and treebanks of Bulgarian, Czech, Slovak and Slovenian recognise no such category at all.

1.2. Part-of-speech Status of Predicatives and Part-of-speech Hierarchy

As suggested by the various terms used for the category under scrutiny (we will provisionally call it **predicative** further on), its placement in the part-of-speech hierarchy also varies. It can be either listed among other parts of speech (as does Yury Maslov for Bulgarian and Vladimir Vinogradov for Russian), or considered a variety (a so-called syntactic derivative, a sub-part of speech) of the adverb (BgAcGr 1983), the noun and the adverb (RuAcGr 1970), the noun, the adverb and the participle (RuAcGr 1980), or the verb (Віхованець 1988), on its own or together with the copula (when used).

1.3. The Coverage of Predicatives

The unclear status of predicatives is probably the reason for which the coverage of this category differs so much from language to language and from one author to another. The IPIPAN corpus of Polish numbers a total of 26 predicatives, but in the Ukrainian Grammatical Dictionary there are 176, and in Yefremova 2006 and the Russian National Corpus (RNC) about 1200. In order to understand where the difference comes from, we need to see what words are in fact classified as predicatives in each case. To this end we compared evidence from grammars, dictionaries and corpora for four Slavic languages: Russian, Ukrainian, Polish and Bulgarian.

The relative[3] core of predicatives is constituted by modal words such as Pl *trzeba*, Ua *треба*, Ru *надо* ‘necessary’*,* Pl *można*, Ru *можно*, Ua *можна* ‘possible’, Ru *нельзя* ‘impossible’. Also commonly, though not universally, included are words for:

- physical properties and sensations and states of the environment: Bg *студено*, Ru, Ua *холодно* ‘cold’, Pl *zimno*;
- mental states and sensations: Ru, Ua, Bg *весело* ‘joyful’, Pl *wesoło*;
- evaluations of events or situations: Bg, Ua *добре*, Pl *dobrze*, Ru *хорошо* ‘well’.

Most such words can also be used as adverbs of quality (if listed in the dictionary as predicatives, they are said to be homonymous to adverbs of quality). Some of them can be considered as belonging to more than one of these semantic classes, sometimes with minor differences in syntax.

Some predicatives are derived from nouns expressing states and emotions going, according to many authors, through the stage of adverbs (e.g., Ua *жаль, шкода* ‘pity’*, гріх* ‘sin’*, сором, ганьба* ‘shame, disgrace’, *страх* ‘fear’*, охота* ‘wish’*, час, пора* ‘time’*, кінець* ‘end’). Many scholars include a group derived from infinitives (Ua *видати,* Ru *видать,* Pl *widać* ‘seen, one sees’*,* Ua *чути,* Ru *слыхать,* Pl *słychać* ‘heard, one hears’, etc.) and, in the case of Ukrainian, the corresponding impersonal participles in *-но/-то* (*видно* ‘visible’*, чутно* ‘audible’).

The remaining predicatives in each mentioned language include different groups of words, which we will try to analyse further. In various sources we can find mention of: short (predicative) forms of adjectives, participial predicatives, negative proforms with sentential value, comparative degree forms of adverbs and adjectives, figurative (deverbal or onomatopoeic) words with semelfactive semantics (Ua *бух* ‘bang!, thud!’*, цвірінь-цвірінь* ‘chirp-chirp’), diminutive verbs, performatives (‘thank you’, ‘yes’, ‘no’), invariable foreign words (Ua *каюк* ‘all up, over’, *негліже* ‘in undress, in dishabille’, *пас* ‘pass’), collocations, etc.

[3] This reservation is due to the fact that Bulgarian prefers **modal verbs**—mostly impersonal ones: *трябва* ‘must’, *може* ‘may’.

The table on the following page presents a summary of the coverage of the category in different accounts. A black circle means that lexical items of this group are by and large included into the principal predicative category, whatever its name and status are; a shaded cell means that the group does not exist in this language. Some special cases are explained in the notes following the table.

	Russian							Ukrainian		Polish		Bulgarian	
	Shcherba 1928	RuAcGr 1970	RuAcGr 1980	CtRuLg 1964	Maslov 1975	Yefremova 2006	RNC	Šerech 1951	UGD	PlAcGr 1998	IPIPAN corpus	Maslov 1981	BgAcGr 1983
1. modal term	●	●	●	●	●	●	●	●	●	●	●	●	●
2. other adverb-like term	●	●	***a***	●	●	●	●	***a***	●	●	●	●	●
3. negative adverb		●	***a***		●	●	●		●				
4. participial adverb		***p***	***p***			●	●	***p***					●
5. noun	●	●	●	●	●	●	●	●	●	●	●	●	
6. ex-noun	●	●	●	●	●	●	●					●	●
7. negative proform		●				***s***	●	●	●				
8. comparative degree							●						
9. figurative word									●				
10. foreign word									●				
11. performative									●				
12. copulative pronoun							●				●		
13. diminutive verb									●				
14. infinitive of perception									●	●	●		
15. *stać*											●		
16. personal adverb	●				●	●	●					●	
17. personal collocation	●					●	●						
18. closed-class adjective	●				●	●	●			●	***fl***		
19. open-class adjective	●												
20. instrumental noun	●												

Legend:

1. A word with a modal meaning (Ru *надо*, *надобно*, Ua *треба* ‘need’).
2. A word having the form of a deadjectival adverb, denoting a state of nature, a mental or physical state, etc.
3. A negative adverb (Ru *невдомёк* ‘unknown, unbeknown’, Ua *невтямки*, *невдогад* dto.).
4. An adverbial form of a participle.
5. A citation form of a word which is also in use as a noun.
6. A citation form of a noun no longer in use in the language.
7. A negative proform with predicative value (Ru *нечего* as in *нечего пить* ‘there is nothing to drink’).
8. A comparative degree form of an adjective or adverb.
9. A deverbal or onomatopoeic word with semelfactive semantics.
10. An invariable foreign word (Ua *нас* ‘pass’).
11. A performative expression, which may be a verb (Ua *дякую* ‘{I} thank [you]’), a particle (Ua *так* ‘yes’, *ні* ‘no’), etc.
12. The copulative pronoun Pl *to*, Ua *то*, *це*, Ru *это*.
13. A diminutive formed from a verb (Ua *спунькати, спатки* < *спати* ‘sleep’, *ходитоньки* < *ходити* ‘walk’).
14. An infinitive of a verb of perception (Ua *чути,* Pl *słychać* ‘heard, one hears’).
15. Pl *stać* (lit. ‘stand’) as in *stać mnie na to* ‘I can afford it’.
16. An adverb used as a predicate with a noun phrase subject (Ru *навеселе* ‘tipsy’, Bg *добре* ‘well’).
17. A prepositional phrase used as a predicate with a noun phrase subject (Ru *в состоянии* ‘in a position ‹to›’).
18. A short (predicative) form of an adjective from a closed class (Ru *рад*, Pl *rad* ‘glad’).
19. A short (predicative) form of an adjective from the open class (Ru *весел* ‘merry’).
20. A noun in the instrumental case, contrasting with the same noun in the nominative.

Notes:

a: predicative adverb (RuAcGr 1980), adverb which arguably is one no longer (Šerech 1951).
p: predicative form of participle (RuAcGr 1970), participial predicative (RuAcGr 1970), special form of verb (Šerech 1951).
s: only those with secondary meanings (Ru *нечего* as in *нечего пить* ‘no sense in drinking’: Yefremova 2006).
fl: separate flexeme (IPIPAN corpus).

1.4. Where Does the Difference Come from?

Some differences in treatment are due to languages' specific features, lexical items and constructions restricted to one or a few languages, such as the Bulgarian[4] expressions of feelings and sensations with a noun as a predicative and an accusative experiencer (e.g., *яд* *ме*$_{\text{Acc}}$ *е* 'I am vexed'). More often, however, the differences are due to the dissimilar approaches chosen by schools and scholars. For example, Polish *to* '(this) is' in *To książka* 'This ‹is a› book' is traditionally called a predicative, although its counterparts *то*, *це* in Ukrainian and *это* in Russian are usually treated as pronouns. In Russian forms of the comparative degree, common to adjectives and adverbs, are listed among predicatives, as well as numerous collocations.

There are also instances of discrepancies regarding predicatives and their place in the part-of-speech hierarchy within one author's work, or even within the same text, so that in different chapters the same or similar phenomena may be treated in different ways.

2. Predicatives in History

2.1. Russian

There is a general tendency throughout the history of Russian linguistics towards increasing the status of the class of predicatives.[5] Introduced very tentatively by Lev Shcherba and promoted by Vladimir Vinogradov, it has been constantly 'gaining in weight'. But this is not a strict tendency. In one of the most recent lexicographic works – Шаров 2002 – no such category is mentioned at all. Most of its instances are listed as adverbs, and some of them are listed within the category 'other' (the latter includes a total of 85 items, which is far fewer than the number in Ефремова 2006).

As CtRuLg 1964 reports, grammarians have manifested interest in these words since the early 19th century. Alexander Vostokov and Aleksey Shakhmatov classified these words with the verbs (*глагольные слова*). Konstantin Aksakov treated them as short (predicative) forms of adjectives.

Щерба 1928 focussed on such invariable words as *нельзя* 'may not', *можно* 'may', *жаль* 'pity', which function exclusively as predicates, and on adverbs such as *холодно* 'cold' in predicate position. He argued that these were not adverbs or adjectives, since they modified neither verbs (or adjectives, or adverbs) nor nouns. Considering that their stative meaning was their most distinctive property, he noted that there was, perhaps, a separate part of speech in the making here, one that could be described as a **category of state** (to contrast with adjectives and verbs, which can also denote states, but present them as qualities and actions). Yet he acknowledged that the range of words which can function as stative predicates (or rather as their complements) in Russian is much wider, including short forms of adjectives (*я весел* 'I am joyful') and even nouns in the instrumental case (*я был солдатом* 'I was a soldier').[6] For this reason he stopped short of stating that there was such a part of speech in the language already. Notwithstanding he thought that, if there was not one, words such as *пора* '(high) time', *холодно* 'cold', *навеселе* 'tipsy', as they could not be called adverbs, should remain part-of-speechless.

Later authors have argued for a more or less autonomous class with somewhat varying coverage. The general consensus is that modals (*можно* 'may', *надо* 'must', *нужно* 'need') constitute the nucleus of the class. Words of state having the form of adverbs of quality, for the most part doing double duty as short singular neuter forms of adjectives[7] (e.g., *весело* 'jolly'), are regularly included as well, although Мещанинов 1945 suggests that short forms of adjectives and participles are better candidates for being set off as a separate part of speech, since they are restricted to predicate position. The list is commonly extended to cover nominative singular forms of nouns (e.g., *грех* 'sin', *охота*

[4] Actually Balkan (shared with Serbo-Croat and Slovenian, as well as some non-Slavic languages of the region).

[5] In Russian secondary schools predicatives have been recently introduced as a part of speech into the syllabus in Russian language.

[6] As opposed to the full form of the adjective and the nominative case of the noun. One should note the contrast between *я весел* (a temporary state) and *я весёлый* (a permanent quality), cf. Spanish *estoy alegre* vs *soy alegre*, and between *я был солдатом* (a temporary state) *я был солдат* (an identity), cf. French *j'étais soldat* vs *j'étais un soldat* (Shcherba's example), Irish *bhí mé im shaighdiúir* (glossed 'was I in:my soldier') vs *ba saighdiúir mé* (glossed 'was soldier I' with a different copulative verb).

[7] Occasionally the two differ, as in *ранне* (adjective) but *рано* (adverb) 'early', or *больно* 'ill, painful' with stress on the stem when an adverb and on the ending when an adjective.

'wish, inclination'; also the word *жаль* 'pity', which has practically gone out of use as a noun) and some other items.

Most authors define the category so as to include only words that function as predicates in impersonal sentences, but Маслов 1975 relaxes the criteria to also cover predicates with noun subjects, as in Щерба 1928's proposal, though only such where the predicative word has no other use with the same meaning—that is, invariable words (*начеку* 'alert') and short forms of adjectives with no corresponding long forms, thus no possible attributive use, but inflecting for number and gender to agree with the subject (*рад* 'glad', *намерен* 'intent').

The status of the category and the associated terminology vary also. To RuAcGr 1970 and 1980 **predicatives** are a syntactic derivative within the categories of adverb and noun[8] (and participle, in the later work), that is, the part of speech to which such a word belongs depends on its morphological structure, although its syntactic behaviour does not. To CtRuLg 1964 the **impersonal predicative word** is a separate part of speech (although the words themselves may be homonymous to adverbs or to citation forms of nouns), as is the **non-verb predicative** to Маслов 1975 (a part of speech intermediate between the verb and the adverb, represented in English by such words as *asleep, awake, afloat, alive*).

Щерба 1928's expression **category of state** is also used—a troublesome term, as Маслов 1975 points out, because its structure ill fits into the set of names of parts of speech, among other reasons. It appears, however, that Shcherba himself did not intend it to be a name of a part of speech, but rather a description of one; he wrote that the category of state, 'for want of a better term, might be called a **predicative adverb**' (even though none of these words are adverbs in his view). The term **predicative adverb** is applied to all predicatives affiliated to adverbs in RuAcGr 1970; in RuAcGr 1980, as we shall see, it excludes words expressing necessity or possibility ('**predicatives** proper') but covers the rest (including some that have no adverbial use).

The choice of status assigned to these words has one more implication. If they are deemed to constitute a separate part of speech, this part of speech is claimed to possess the categories of tense and mood, so that *было тихо* 'it was quiet' is an analytic past tense form of *тихо* '(it is) quiet', and *стало тихо* 'it became quiet' must be one too (somewhat of a problem for this analysis). Otherwise these are phrases consisting of an inflecting copulative verb and its invariable complement.

CtRuLg 1964 constructs a detailed classification of all impersonal predicative words on the basis of their semantics, as follows:

- mental and physical states of living beings, states of nature and the environment:
  - mental states of a human being (*боязно* 'frightened', *грустно* 'sad'),
  - physical states of living beings (*больно* 'painful', *тошно* 'sickening'),
  - states of nature and the environment (*ветрено* 'windy', *уютно* 'cosy');
- modal states (necessity, possibility);
- evaluations of states or positions:
  - extent in time and space (*поздно* 'late', *время* 'time', *далеко* 'far'),
  - psychological or moral and ethical evaluations (*плохо* 'bad', *легко* 'easy', *позор* 'disgrace'),
  - sensory perception (*видно* 'visible', *слышно* 'audible').

A separate classification divides them into words with adverb and noun origin, and the former group into such as are adverbs and such as are not. Conversely, it is pointed out that circumstantial adverbs are more likely to give rise to impersonal predicative words than attributive adverbs are (*было рано* 'it was early', but not **было быстро* 'it was quick', **было длинно* 'it was long').

RuAcGr 1970 and 1980 merge the two classifications. RuAcGr 1970 recognises

- predicatives affiliated to the noun;
- predicatives affiliated to the adverb *alias* predicative adverbs:
  - words that are adverbs:
    - emotional states,
    - physical states,
    - states of the environment,

[8] Modal words and expressions, such as *правда* 'indeed', *словом* 'in a word', *в частности* 'in particular', *к сожалению* 'unfortunately', are another syntactic derivative of the noun.

    - words that are not adverbs (modals as well as internal states such as *стыдно* ‘ashamed’, *любо* ‘lief’);
- negative terms (*недосуг* ‘no leisure’, *некогда* ‘no time’, *невдомёк* ‘no idea’).

In RuAcGr 1980 predicatives affiliated to the adverb and predicative adverbs are separated, the former comprising modals and the latter everything else. Participial predicatives (short singular neuter forms of past passive participles: *закрыто* ‘closed’, *запрещено* ‘forbidden’) are also singled out. Negative terms, however, are no longer treated as impersonal predicative words. Thus in both works a semantic classification is only applied to predicative adverbs.

All semantic classifications place each word in one class only, although some words have different meanings depending on the construction (e.g., *мне плохо* ‘I am unwell’ expresses a physical state, whereas *воровать плохо* ‘it is bad to steal’ is a moral evaluation).

2.2. Ukrainian

In Ukrainian grammars predicatives were particularly ‘fashionable’ in the 1950-60’s. The only direct mention of them as a full-fledged part of speech, called **category of state** (*категорія стану*), can be found in Жовтобрюх & Кулик 1965.[9] The authors divide them into such as denote

- mental or physical states of human or any other beings,
- states of nature,
- states of the environment, their subjective assessment, their extent in time or space,
- modal states.

Кулик 1961 is more circumspect. He describes the **impersonal predicative word** (*безособово-предикативне слово*) in terms of its syntactic function and presents a classification, but refrains from specifying its part-of-speech status.

The prominent Western linguist of Ukrainian origin Jury Šerech (Shevelyov), whose works were not known in Soviet Ukraine, discusses **adverbial impersonal sentences** with stative semantics and a null or overt copula plus

- an adverb,
- a modal or a denominal, called a **predicative word**,
- a negative pronoun or adverb such as *нікому* ‘no one to … (to)’, *нічим* ‘nothing to … with’, *ніде* ‘nowhere to …’, etc.,

in predicate position. Regarding the status of these words the author’s attitude is tentative. Pointing out that adverbs ought to serve as adverbial modifiers of verbs, adjectives or other adverbs, he suggests that it would be more correct to have a special term to refer to them in those cases when they act as predicates, and mentions that the Russian linguist Vinogradov has proposed calling them **category of state**; however, whether he accepts this strategy is not completely clear (Шерех 1951:92). Later on he also uses the term **adverbialised word** along with Vinogradov’s **category of state**.

Леонова 1983 calls these words **predicative adverbs** as opposed to attributive adverbs. They include modals, states proper, and denominatives. Predicative adverbs are likened to verbs because they can govern cases.

In the 1990’s Ukrainian scholars preferred writing about syntactic functions. Consider for example the accounts of Ukrainian morphology by Безпояско et al. 1993 and of Ukrainian syntax by Віхованець 1993, published as a set. The first part of the grammar does not mention predicatives at all, and it is not clear which part of speech words like *треба, можна* (the most uncontroversial representatives of the class) belong to; *можливо* ‘(it is) possible’ is referred to as a modal adverb.

In the second part predicatives are described in terms of their syntactic role (Віхованець 1993:253-254):

> The core group of predicates of state is composed of invariable words that are referred to the so-called category of state or to predicative adverbs. These words have verbal forms of tense and mood and function as the principal member of impersonal sentences, corresponding to a predicate. Tense and mood are expressed analytically with the help of the analytic syntactic morpheme-copula *бути* ‘be’ or of analytic syntactic semi-morphemes such as *ставати, робитися* ‘become’. […] They are

[9] The very choice of terminology is evidence of the powerful influence of Russian upon Ukrainian linguistics.

usually univalent, or bivalent at most (*Дідусеві видно гори* ‘Grandpa can see the mountains’, *Дівчині жаль пташини* ‘The girl is sorry for the bird’).

The author mentions pairs of denominal predicatives that preserve the form of nouns and counterparts of the adverb variety derived from them: *гріх—грішно* ‘sin’, *досада—досадно* ‘annoyance’, *жаль—жалко* ‘pity’, *кривда—кривдно* ‘offence, injustice’, *сором—соромно* ‘shame, disgrace’. In this work he does not seem to adhere to any position on the status of the category, only saying that predicatives are a separate part of speech according to some scholars, while others refer them to a subclass of adverbs, namely, predicative adverbs. In his other, earlier work he is confident in referring to such words as verbs transposed from adverbs (Віхованець 1988: 119–122).

The *Ukrainian Grammatical Dictionary* (UGD) is modelled upon Andrey Zaliznyak’s *Grammatical Dictionary of the Russian Language*, and treats predicative words in the same way. As we can see from the table and the statistics, however, their coverage in UGD is much poorer.

2.3. Polish

The Polish grammatical tradition varies considerably in its approach to predicatives. The popular ‘rigorous’ description of Polish in Saloni & Świdziński 1985 mentions only the modals *trzeba* and *można*, calling them **improper verbs** with a ‘minimal verbal paradigm’: indicative present *można* ‘(it is) possible’, past *było można*; conditional present *można by*, past *byłoby można* (Saloni 1974:66).

The most comprehensive and consistent overview of the category is presented in PlAcGr 1998. The authors single out **predicatives** as a separate part of speech, dividing them into personal (deadjectival) and impersonal (deverbal, denominal, deadverbial) ones. This division is purely morphological, as in fact some impersonal predicatives express person by oblique arguments (see examples below). Personal predicatives (traditionally ‘short’, predicative adjectives) are opposed to full adjectives as they do not inflect for case. Such are *(po)winien* ‘indebted, owing’, *rad* ‘glad’, *kontent* ‘content’, *wart* ‘worthy’, which have no full forms, and also *ciekaw* ‘curious’, *godzien* ‘worthy’, *gotów* ‘ready’, *pełen* ‘full’, *pewien* ‘certain’, *łaskaw* ‘kind’, *syt* ‘satiated’, *świadom* ‘aware’, *wesół* ‘joyful’, *zdrów* ‘hale’.

The authors point out that most predicatives are impersonal. They include modals (*trzeba, można* etc.), words of perception (*widać, słychać* etc.), mental states. Many of these are homonymous to citation forms of nouns: *czas, pora* ‘(high) time’, *grzech* ‘sin’, *szkoda, żal* ‘pity’, *wstyd* ‘shame’ (PlAcGr 1998:129). It is said that they are treated differently because of their ‘morphological peculiarities’, as they only have analytic forms, whereas verbs proper have both synthetic and analytic forms (PlAcGr 1998:60). Lexemes that ‘are traditionally treated as adverbs’, but can function as predicates, are also included: *nudno mi* ‘I am bored’, *Dziecku zimno* ‘The child is cold’, *Ależ tu cicho!* ‘Oh but it’s quiet here!’, *Duszno dzisiaj* ‘It is stuffy today’. They are opposed to adverbs on the basis of their syntactic function of a predicate; they require a dative noun phrase and/or a locative expression: *Jemu w Krakowie jest zimno* ‘He is cold in Cracow’. And, as the authors point out, while the majority of adverbs formed from adjectives end in *-o* and a minority end in *-(i)e*, about a dozen adverbs which have forms with both endings, the first form being predicative and the second attributive: *gwarno* vs *gwarnie* ‘noisy’, *rojno* vs *rojnie* dto., *mglisto* vs *mgliście* ‘foggy, nebulous’, *nudno* vs *nudnie* ‘boring’; the words *pochmurno* vs *chmurnie* ‘gloomy’ are also usually counted, although they differ by a prefix as well (PlAcGr 1998:61). The term ‘predicative’ is sometimes used interchangeably with ‘impersonal predicators’ (*predykatory nieosobowe*) in this work.[10]

The IPIPAN corpus of Polish language adopts the approach of PlAcGr 1998 with some changes. All adverbs of state are classified simply as adverbs, with no distinction between those ending in *-(i)e* and in *-o*, even if they constitute pairs. Predicative adjectives are listed as a separate flexeme called *winien* by its only[11] actual representative in the corpus of 300 million tokens. The short adjectives listed in the grammar seem to be obsolete as they cannot be found in the corpus or are erroneously classified as nouns.[12]

[10] Other terms used for ‘predicatives’ are **non-inflecting verbs** (*czasowniki niefleksyjne*, Jodłowski 1971), **improper verbs** (Saloni 1974). Those accounts, however, include only modals.

[11] Together with its variant *powinien*.

[12] This is often the case with *rad* ‘glad’, recognised by the morphological analyser as the genitive plural form of the noun *rada* ‘advice’.

2.4. Bulgarian

Among the characteristic traits of Bulgarian which have a bearing on the constructions at hand are the obligatory use of the copula in all moods and tenses[13], the obligatory marking of focussed objects by cliticised pronouns, the lack of an opposition of short and long forms of adjectives, the use of the indefinite singular neuter form of most adjectives as an adverb[14] (as in Russian, but unlike Ukrainian and Polish), the prevalent use of verbs of necessity and possibility rather than adverbs, and the absence of an infinitive, which is replaced by analytic forms of the so-called conjunctive.

BgAcGr 1983 singles out under the name **predicative adverbs** a class of adverbs which occur as complements of a third person singular neuter form of the verb *съм*, *бъда* 'to be' or *стана*, *ставам* 'become' and denote states rather than properties of actions, qualities or entities. By far most of these have the form of adverbs derived from adjectives, though some differ from the corresponding adjectives in their meanings (e.g., *драго* 'pleasant' from *драг* 'dear'). There are a few, of various origin, which have no other use (*блазе* 'blessed, lucky', an obsolete adverb from the adjective *благ* 'gentle, mild, sweet'; *еня* 'concern, care', a negative polarity item, from the Greek noun *ἔγνοια*). Words in active use as nouns (such as *яд* 'anger; worry, trouble') are not included in this group. The grammar classifies all items on the basis of their meaning:

- states of nature and the environment (*тъмно* 'dark', *уютно* 'cosy');
- physical and mental states of a human being (*зле* 'bad', *опасно* 'dangerous');
- ethical and emotional estimates (*грешно* 'sinful, wrong', *мило* 'pleasant');
- intellectual or modal states (*нужно* 'needed', *лошо* 'bad');
- miscellaneous estimates of a state or situation (*късно* 'late', *далеч(е)* 'far').

То Маслов 1981:290–293 the **non-verb predicative** (alias **'category of state'**) is a separate part of speech comprising three syntactic subclasses:

- words of noun origin, most (though not all) active in the language as nouns, but set apart in this function by several circumstances:
  - they appear in an unchanging form (indefinite singular),
  - they can't be modified by adjectives, only by adverbs,
  - their gender no longer matters,
  - most subcategorise for an obligatory or optional experiencer expressed by a cliticised personal pronoun, chiefly accusative (*яд* *<u>ме</u>*$_{\text{Acc}}$ *е* '<u>I</u> am vexed'), less frequently dative (*жал* *<u>ми</u>*$_{\text{Dat}}$ *е* '<u>I</u> am sorry');
- words of adverb origin, mostly usable as adverbs, but set apart in this function by appearing in a position atypical for adverbs and by the fact that many subcategorise for an obligatory or optional experiencer expressed by a dative clitic (*добре е* 'it is well', *добре ми е* 'I am fine');
- a sparse group of **non-impersonal non-verb predicates** (*добре съм* 'I am well').

Non-verb predicatives inflect for tense and mood; that is, the verb *съм*, *бъда* 'to be' or *стана*, *ставам* 'become' (the latter only with words with adverb origin, e.g., *стана тъмно* 'it became dark') is part of the analytic word form. This implies that *<u>вечер е</u>* 'it is evening' and *<u>късно е</u>* 'it is late' are entirely distinct constructions (a noun plus a verb and a non-verb predicative, respectively), which is rather unexpected. The relation between the two is similar to the one between *<u>лъжец е</u>* 'he is a liar' and *<u>лъгал е</u>* 'he has lied' (a verb in the perfect tense), but is more complex in view of the opaque status of *става късно* 'it is getting late', which has no counterpart in the analytic part of verb paradigms.

[13] There is a small number of adverb-based constructions that don't involve a copula (*Добре че дойде* 'Good that you came', *Чудно как е избягал* 'One wonders how he escaped', *Жалко за парите* 'Shame for the money', *Край на болестта* 'The malady is over', *Блазе му* 'Happy he'), which BgAcGr 1983 brings up as showing that the adverb is the bearer of the semantics of the predicate. However, the paucity of adverbs that admit such use, the syntactic deficiency of the constructions (they require certain types of complements, allow neither negation nor interrogation, etc.) and their specific semantics all argue that this is not a case of a copula being omitted in the present tense, but a different kind of construction, one that might be called an **attitudinal**, anchored to the present situation and insensitive to the category of tense.

[14] Among the exceptions are *добре* 'well', *зле* 'ill, badly', *рано* 'early' (adverbs), but *добро*, *зло*, *ранно* (singular neuter forms of the corresponding adjectives).

3. Criteria for Singling out Predicatives. Discussion

Considering the similarity in the syntactic behaviour of the words under consideration regardless of their origin, a unified treatment is certainly desirable.

CtRuLg 1964 and Маслов 1981 stress the fact that such words, even when they are nouns in appearance, seem to lose their gender, because the copula is always in the neuter (when it is in a tense in which its form expresses gender). But agreement ought to take place if the predicative word were a subject. This, however, is not so (cf. Ru *Не время бунтовать* 'It is not time to revolt': normally the negative particle precedes the subject only if the negation has narrow scope, as it were, 'It is not time, but something else, to revolt'). Also, predicatives affiliated to adverbs have the same distribution, and adverbs are even less likely subjects than they are complements of copulative verbs.

On the other hand, in Russian the non-zero copula in the past or future tense can have an instrumental complement (as said above, in stative sentences rather than statements of identity), but **Охотой было идти* '(One) felt like going' (instead of *Охота было идти*) is ungrammatical. This can be explained if *охота* here is not a noun but an adverb, and by virtue of this invariable.

What prevents these words from being counted as adverbs is a more complicated question. Those authors who consider them a separate part of speech say that many of them are homonymous to adverbs (though the sameness of form, so common across languages, is certainly not fortuitous). They are said to be confined to the predicate, a position inaccessible to their homonyms. Yet this is effectively equivalent to saying that the same words may appear both in the predicate and outside it, though frequently with an accompanying difference in meaning.

This leads us to draw an analogy between adverbs and adjectives. Adjectives can be modifiers of nouns or complements of copulative verbs. By far most English adjectives have both uses, but some are exclusively attributive (*main reason*, *undue pressure*, *future lawyer*) or exclusively predicative (*glad*, *afraid*). As we saw, in Russian too there are exclusively predicative adjectives, morphologically marked by having no long (attributive) forms (*рад* 'glad') or semantically marked by having recognisably different meanings depending on the form (*должен* [short form] 'obliged, owing', cf. *должный* [long form] 'due, owed'). It is more difficult to find exclusively attributive adjectives in Russian, but there are adjectives whose long forms must be used even in predicative position. Compare:

1) a. **Эта причина главна.*
   b. *Эта причина — главная.*
2) a. **This reason is main.*
   b. *This reason is the main one.*

Notwithstanding adjectives are treated as a single part of speech in nearly all accounts.

Adverbs modify verbs, adjectives, other adverbs or, less commonly, nouns; this is their attributive function. But many of them (e.g., locative or temporal ones) can be complements of copulative verbs (*The building is here*, *The debate was yesterday*). Let us admit that adverbs of manner can be complements of copulative verbs too. It is more common for adverbs to be restricted to one of the two functions than it is for adjectives, but this is a difference of quantity, not of quality.

3.1. Semantic criteria

The semantic criterion, namely, the stative semantics of predicatives, has been a major part of the case for singling them out as a class since the outset. But already Щерба 1928 encountered the problems associated with it. Denoting states is not a prerogative of predicatives: there are stative verbs and verb forms (even if the verb is generally regarded as a category of action), as well as stative sentences with adjectives or nouns as the complement of the copula.

A general problem with all attempts to define a part of speech on the basis of semantic criteria is that the same meaning can often be expressed by various means. For example, CtRuLg 1964 states that 'an impersonal predicative word is a stranger to the meaning of a feature (a feature of an entity is an adjective; a feature of an action or a feature is an adverb)'. This seems to miss the fact that *Жить хорошо* and *Жизнь хороша* 'Life is good' mean the same thing, as do *Мне весело* and *Я весел* 'I am joyful', and if one sentence expresses a feature, so should the other. Nor is the difference between a state and a feature made clear in most accounts.

3.2. Morphological criteria

We mentioned the morphologically marked opposition of predicative forms in *-o* and adverbial forms in *-(i)e* that characterises a few lexical items in Polish (PlAcGr 1998). However, if we look up these items in the corpus, we can see that the speakers' intuition does not support this difference, or at any rate it is not followed strictly. Corpus examples show predicative use in 11 cases out of 17 for *gwarnie*:

> *W miniony weekend było* ***gwarnie*** *i kolorowo* 'The last weekend it was noisy and colourful'
> *wiedziałem jedynie, że jest wesoło,* ***gwarnie*** 'I only knew it was merry, noisy'

Since *kolorowo* and *wesoło* end in *-o*, the choice of the form in *-(i)e* is not motivated by analogy. The same can be seen in the case of *rojnie*, 2 out of whose 3 uses are predicative.[15] Neither does the choice of one of the two endings by those items that only have one form correlate with their use in any way. The former paronyms seem to have become unstable variants.

The argument that non-verb predicatives possess the categories of tense and mood seems an unnecessary artefact of the analysis. The verb phrase does, being headed by a copulative verb (perhaps a zero one). This is no different from any other verb phrase consisting of a copula and a complement.

3.3. Syntactic criteria

The syntactic criterion is another important argument for singling our predicatives. The fact that they can appear in predicate position (and some are restricted to it) is often brought up as a claim for part-of-speech status. However, this function is not unique to them: adjectives and nouns can be complements of copulative verbs too, so can some classes of adverbs, and of course the verb is the predicative word *par excellence*. Meanwhile the adverb is far from being syntactically uniform: for instance, those adverbs that can modify nouns (Ru *яйца всмятку* 'eggs soft-boiled'*, совсем ребенок* 'altogether a child'; examples from Маслов 1975) have just as good a reason for forming a subclass.

The employment (or omission) of the copula is sometimes brought up as an argument for classifying predicatives as verbs (Віхованець 1988). In fact the use of the copula varies across languages (and grammatical contexts) from Russian, where it is regularly left out in the present indicative, to Bulgarian, where it is obligatory in all tenses and moods.[16] The other languages fall in between. For example, Ukrainian, unlike Russian, employs the copula in negated present-tense sentences: *Читати цю книгу не є цікаво* 'To read this book is not interesting'. Again, the omission of the copula is not restricted to sentences with predicatives.

This said, the syntactic part of the 'identity' of predicatives is a highly interesting matter that requires further investigation.

4. Conclusions

We have seen that the words classified as predicatives vary considerably in their morphology, etymology and syntactic behaviour. There is no criterion that would keep them together in one group and oppose them to all remaining parts of speech, which suggests that they should rather be treated as a subclass of one of the existing parts of speech. Classifying most of them as adverbs doesn't make the adverb more diverse than it already is, or than the adjective is in most accounts. It is commonly taken for granted that adjectives can function as both attributes and predicates, whereas adverbs can only be attributes (Маслов 1975:208–215). But given that there are attributive-only adjectives and predicative-only adjectives, why should it not be the same with adverbs?

This refers to the core of the category. What about the periphery? Polish *to* (as in *To książka* 'This ‹is a› book') can be treated as a pronoun, as Ukrainian *то* or *це* and Russian *это*. The shared comparative degree forms of adjectives and adverbs in Russian ought to be classified as adjectives or adverbs, disambiguating them depending on the function they perform. The Ukrainian and Russian

[15] The statistics of the use of these adverbs is quite interesting. One can clearly see a preference for using the predicative form of some of them and the attributive form of the others, which correlates with their semantics: *nudno* vs *nudnie* 146::13, *mglisto* vs *mgliście* 28::78 (*mgliście wiem* 'I have a dim idea', lit. 'I know nebulously'), *gwarno* vs *gwarnie* 78::17, *rojno* vs *rojnie* 18::3.

[16] It also varies in history; it was used with greater regularity in Old Russian writing: *Се ми* <u>*есть*</u> *ближе, к тому поиду переже* 'This one <u>is</u> closer to me; I will go to that one first' (Hypatian Chronicle 1908:400).

diminutive verbs are verbs even if they have a super-defective paradigm only including an infinitive. Semelfactives can be classified as interjections (or perhaps as invariable verb forms).

There may be an advantage in introducing an attribute of predicativity for various parts of speech with values 'predicative', 'non-predicative' and 'either' for each lexeme, but this issue needs additional scrutiny.

Further research

The notion of predicatives and the whole ongoing discussion around it brings up at least three topics for further research:

1. Whether we dispose of predicatives as part of speech altogether dissolving them among parts of speech where they came from or leave them as subclasses of each, we are still left with the phenomenon of the exclusive predicativeness of certain words. This feature certainly brings important information for the theory of language in general and its applications for automatic language generation, and should be explored in depth.
2. It seems important to describe the syntactic features of 'predicatives' in detail, in order to clarify and justify their status from the syntactic point of view.
3. One more eminently worthwhile task is to align other categories in present tagsets of Slavic languages and work on a common pattern for them.

References

BgAcGr (1983): Стоян Стоянов (гл. ред.), «Граматика на съвременния български книжовен език (т. 2: Морфология)». София: Издателство на Българската академия на науките.

CtRuLg (1964): Валгина Н. С., Д. Э. Розенталь, М. И. Фомина, В. В. Цапукевич, «Современный русский язык» (2-е изд.). Москва: «Высшая школа».

Hypatian Chronicle (1908): «Полное собрание русских летописей, изданное по высочайшему повелению Императорскою археографическою комиссиею», т. 2: Ипатьевская летопись. Санкт-Петербург: М.А. Александров.

Jodłowski, S. (1971): *Studia nad częściami mowy*. Warszawa: Państwowe Wydawnictwo Naukowe (PWN).

Kotsyba N, Shypnivska O., Turska, M. (2008). 'Principles of Organizing a Common Tagset for PolUKR (Polish-Ukrainian Parallel Corpus)'. // Mieczysław Kłopotek et al. (ed.), *Intelligent Informatic Systems XVI. Proceedings of the International IIS'08 Conference held In Zakopane, Poland, June 16–18, 2008* Academic Publishing House EXIT, Warsaw, 2008, 471–480.

Multext-East (1998): L. Dimitrova, T. Erjavec, N. Ide, H.-J. Kaalep, V. Petkevič, D. Tufiş, 'Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages'. // *Proceedings of the COLING-ACL'98*, Montréal, Québec, Canada, pp. 315–319.

PlAcGr (1998): R. Grzegorczykowa, R. Laskowski, H. Wróbel (red.), *Gramatyka współczesnego języka polskiego (t. 2: Morfologia)*. Warszawa: PWN.

RuAcGr (1970): Шведова Н. Ю. (отв. ред.), «Грамматика современного русского литературного языка». Москва: Наука.

RuAcGr (1980): Шведова Н. Ю. (отв. ред.), «Русская грамматика». Москва: Наука. Online edition.

Saloni, Z. & Świdziński M. (1985): *Składnia współczesnego języka polskiego*. Warszawa: PWN.

Saloni, Z. (1974): „Klasyfikacja gramatyczna leksemów polskich". *Język Polski* LIV: pp.3–13, pp.93–101.

Безпояско О.К., Городенська К.Г., Русанівський В.М. (1993). «Граматика української мови. Морфологія». Київ: «Либідь».

Вихованець І.Р., (1988). «Частини мови в семантико-граматичному аспекті». Київ, «Наукова думка».

Вихованець І.Р., (1993). «Граматика української мови. Синтаксис». Київ: «Либідь».

Ефремова Т.Ф., (2006). «Современный толковый словарь русского языка». Москва: Астрель.

Жовтобрюх М. А., Кулик Б.М., (1965). «Курс сучасної української літературної мови» (частина І). Київ: «Радянська школа».

Кулик Б. М., (1961). «Курс сучасної української літературної мови (частина ІІ: Синтаксис)». Київ: «Радянська школа».

Леонова М.В., (1983). «Сучасна українська літературна мова. Морфологія». Київ: «Вища школа».

Маслов Ю. С, (1975). «Введение в языкознание». Москва: «Высшая школа».

Маслов Ю. С. (1981). «Грамматика болгарского языка». Москва: «Высшая школа».

Мещанинов И. И., (1945). «Члены предложения и части речи». Москва—Ленинград: АН СССР.

Шаров С. А., (2002). «Частотный словарь русского языка». Online edition.

Шерех (Шевельов) Ю.В., (1951). «Нарис сучасної української мови». Мюнхен, «Молоде життя».

Щерба Л. В., (1928). «О частях речи в русском языке». // Л.В. Щерба, «Языковая система и речевая деятельность», Москва, 1974, pp. 77–110.

Remarks on Classification of Parts of Speech and Classifiers in an Electronic Dictionary[1]

Violetta Koseska-Toszewa, Roman Roszko
Institute of Slavic Studies, Polish Academy of Sciences,Warsaw
amaz1312@gmail.com, roman.roszko@ispan.waw.pl

Abstract
The paper presents various classifications of parts of speech for Polish, developed on the base of homogenous or mixed criteria: syntactic, semantic and grammatical (morphological) ones. In the authors' opinion, in electronic dictionaries one should endeavour to increase the numbers of both parts of speech and classifiers. Such an operation would facilitate machine translations. The authors propose a new classifier – a scope quantifier, presented in the paper on the example of scope quantification of time. The subject is much broader, and can also cover quantitative quantification and scope quantification of nouns. The increase in the number of classifiers in electronic dictionaries can become their advantage, distinguishing them from the traditional dictionaries.

Key words: lexical category / word class, classifiers, multilingual dictionaries, existential and universal quantifiers of time, Bulgarian, Polish

1.

The discussion on parts of speech, on the classification criteria for distinguishing them, on establishing borders between parts of speech and on the way of defining them still continues, and will continue in the future. In Maciej Grochowski's opinion, ,,no grammatical class is an isolated being, nor a being completely abstracted away from the context – from the situation of its use. There are few units this could be unequivocally classified into a specific grammatical category" (Grochowski 2005: 7).

2.

As a rule, the division into parts of speech is carried out based on the following types of features:

- morphological (e.g. flexion-related) features,
- syntactic features,
- semantic features, which are to be represented by the object of division – i.e. lexemes.

The classification can be carried out either based on one of the criteria distinguished above, or based on an arbitrary combination of those criteria – consistent or not for the consecutive levels of division.

2.1.

The Polish traditional (school) grammar assumes the division into 10 parts of speech, inherited from the over 2000-year old school of stoicism, Aristotle and the Latin tradition. Hence parts of speech for Polish are: adjectives, adverbs, conjunctions, exclamations, nouns, numerals, particles, prepositions, pronouns, verbs (in alphabetical order). In a school grammar of Polish reissued for the thirteenth time, Piotr Bąk (1977/2007) tries to explain the existing division into parts of speech by referring to the character and structure of the extra-language reality. However, Bąk does not set forth any criteria for the classification of parts of speech he presents.

A traditional approach to distinguishing parts of speech, with small modifications, is also presented in a new work by Renata Przybylska (2004). The author uses a three-stage procedure for distinguishing parts of speech based on mixed criteria, namely, a semantically-syntactic, syntactic, semantic or morphological criterion, depending on the division level.

Zofia Zaron (2003) proposes classification of the Polish lexis based on a mixed morphosyntactic criterion, with emphasis on either connotation or accommodation. For Zaron, the main element of the

[1] The study and preparation of these results have received funding from the EC's Seventh Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX.

morphological criterion is flexion. At the beginning, the connotation allows the author to distinguish exclamations (in this case, in fact the absence of connotation). Following that, syntactic and accommodation characteristics with respect to the morphological category of the person allow her to distinguish verbs and conjunctions standing in opposition to them, and with respect to the morphological category of case – nouns in opposition to personal pronouns (both parts of speech accommodate grammatical gender) and numerals (which accommodate number). Later on, the author distinguishes adjectives and adverbs. Other classification elements are only uninflected parts of speech. The author distinguishes prepositions due to their potential for expressing accommodation. Modalizers, localizers (a part of speech which did not use to be distinguished in Polish literature, e.g. *wczoraj*, *wieczorem*, *kiedyś*, *tutaj*, *tam* etc.) standing in opposition to particles close the structure of Polish lexis division into parts of speech. It is also worth noting that among personal pronouns we can find both the *ja*, *on* and *ktoś*, *coś* forms – carrying different quantification values. Similarly, among localizers (e.g. *kiedyś* and *wczoraj*), forms with different quantification values can also be placed together.

2.2.

„On the level of morphological processing of a natural language", writes Adam Przepiórkowski (2008), "the individual words in the text are assigned their grammatical interpretations. In the simplest case, those interpretations represent information on the so-called parts of speech, and hence about the fact if a given world in a given context is a verb, a noun, a preposition, or a form belonging to another part of speech." Przepiórkowski refers to the classes of lexemes separating off the traditional parts of speech as grammatical categories. The values of those categories are case, gender, and number. In his opinion, the relationship between the morphological and syntactic levels consists also in the fact that syntactic processing systems are based on specific sets of possible morphosyntactic interpretations – so-called tagsets – and such sets are often designed with syntactic processing in mind. This relationship is strongly stressed in Przepiórkowski's work, where we learn– which is worth emphasising – that Polish is one of the few natural languages in the world for which an attempt to extract valency information from text automatically has been undertaken (Przepiórkowski 2008).

2.3.

Other proposals for classification of Polish lexis avoid combining morphological, syntactic and semantic (ontological) criteria, see e.g. works by Saloni (1974) and Laskowski (in GWJP).

2.3.1. Zygmunt Saloni (1974) adopts morphological criteria as the basis for classification. He stresses the distinction between inflected and uninflected lexemes. Inflected lexemes have been subjected to flexion tests (person, gender, case, number) – which was the way to distinguish parts of speech like verb, noun, adjective, etc. Saloni's classification is considered as a morphological one, but uninflected lexemes are distinguished on non-morphological (e.g. syntactic) grounds.

2.3.2. The academic Polish grammar, developed – at least with respect to the issue we are interested in – as a result of many years of discussions, sees the criterion for division into parts of speech in syntactic assumptions (GWJP 1984). Roman Laskowski (a co-author of the mentioned grammar) distinguishes, on the base of syntactic criteria only, 12 parts of speech: verb, add-on, numeral, modalizer, particle, preposition, adjective, adverb, relator, noun, connective, exclamation. Since there is no doubt that there are syntactic relationships between all the words, a division carried out on the base of syntactic functions covers the whole language material. As one can notice, there is no room for pronouns in Laskowski's classification. Other parts of speech known from the traditional classification (i.e., 9 of them) have been preserved, though their scope has sometimes considerably changed. New classes have appeared: (1) relators (which cover the fragment of traditional pronouns restricted to relative pronouns (see Wróbel 1996)), (2) add-ons (e.g. *tak*, *jasne*) and (3) moralizers (e.g. *właśnie*). The adopted assumptions of syntactic classification of the parts of speech have resulted in a change in the set and character of the parts of speech distinguished earlier. For example, the numerals have been restricted to the so-called cardinal numbers. Participles, which used to be interpreted in various ways up to that time, are now placed either in the adjective group (e.g. *czytający*) or in the adverb group (e.g. *czytając*). Laskowski distinguishes 6 criteria for division of Polish lexis: (1) speech independence, (2) syntactic independence, (3) superiority (related to the verb) /

inferiority (related to the noun) in the sentence, (4) accommodation with emphasis on the accommodated element, (5) superiority (related to the noun) / inferiority (rejected in case of the noun) restricted to the name group, and again (6) accommodation, but with emphasis on the accommodating element.

2.3.3. In turn, Maciej Grochowski (2003) proposes a division of parts of speech limited to a fragment of lexis known as synsyntagmatics. Though his field of interest covers only a fragment of the Polish lexis, the criteria are noteworthy. One of them is the criterion of word order: fixed or variable one. Another criterion is the number of positions opened. Still another is the influence scope: local (sentence component) or the whole sentence.

In his doctoral thesis *Wykładniki przybliżoności adnumeratywnej w języku polskim i rosyjskim* (Exponents of adnumerative approximability in Polish and Russian) (2008), Maksym Duszkin draws attention to the fact that Grochowski sections off adnumerative operators into a separate part of speech. Duszkin considers the fact that approximate expressions belong to different parts of speech as a secondary issue. The author focuses on the syntactic properties which in his opinion reflect this fact. Duszkin does not analyze the reasons why exponents of approximability belong to different parts of speech. He only mentions that Mel'czuk speaks „of propositional and adverbial (narečija) exponents of approximability", and that in turn in Polish the same exponents are in general classified to propositions and particles. In Duszkin's opinion, the criteria for distinguishing the individual parts of speech are very complicated, especially that their classification is different for different linguistic schools.

2.3.4. Jadwiga Wajszczuk (1997, 2000) proposes a division of the lexis also based on the syntax (see Laskowski, Grochowski), but with the crucial points positioned differently. First, she is speaking of a dependency-based syntax (syntactems) and of connectivity depending upon relations from other levels (quasi-tactems = paratactems). Among the syntactems, she distinguishes autosyntagmatics (which constitute syntagms) and meta-predicative operators.

2.3.5. A conception for distinguishing parts of speech which is isolated due to moving the emphasis from syntax to semantics is proposed by Andrzej Bogusławski, who presents the outlines of syntactic and semantic descriptions of adverbs correlated with adjectives such as *niespokojnie* in *Jaś śpi niespokojnie*. However, Bogusławski does not deal with epistemic, temporal, or multiplicity adverbs (Bogusławski 2005: 15-44).

3.

We think that, for the completeness of our description, it is interesting to quote the comments on classification of parts of speech in the light of categorial grammar developed by Kazimierz Ajdukiewicz – the father of mathematical linguistics, a Polish logician and philosopher who died in 1963 (Pawlak 1965). Ajdukiewicz's conceptions – which constitute the foundations of contemporary mathematical linguistics – were taken up and developed further only after World War II, in connection with the application of computers to translation. This was done by Jehoshua Bar-Hillel (1953). Ajdukiewicz's and Bar-Hillel's conception concerns the so-called categorial grammar and allows us to resolve whether a given sequence of symbols constitutes a sentence in a natural language or not. This conception was used to obtain the first translation from English to Russian executed by the computer in the USA. In a categorial grammar, the basic notion is that of a syntactic category. A slightly different approach to mathematical linguistics was proposed by Noam Chomsky (Pawlak 1965).

3.1.

In structural research, we are not interested in the meaning-related aspect of the language. Such an approach is known as „syntactic" one, and covers phonetic, morphological and syntactic phenomena. Semantics, or the meaning-related aspect of a natural language, is not taken into consideration here. From the viewpoint of that grammar, all words which can be inserted in place of any words in a correctly built sentence so that the sentence still remains correct perform the same grammatical function, i.e. possess the same syntactic category (Pawlak 1965: 36-37). Example: *Ania (Piotr) lubi (słucha) muzykę (klasykę)*.

3.1.1. So what are the syntactic categories for the major parts of speech in the light of that conception? In the categorial grammar, two categories have been adopted as the basic ones: the sentence category, denoted by the symbol *z,* and the name category, denoted by the symbol ***n*.** Other categories are created on their basis.

The rule establishing the grammatical categories of other parts of speech is as follows:

> If part of speech *C0* constitutes a sentence with parts of speech *C1*, *C2*, ... *Cn*, than *C0* possesses the syntactic category *z/k1. k2*, ... *Cn*, where *k1*, *k2*, ... *kn* are the categories of parts of speech *C1*, *C2*, ... *Cn*.

In the sentence *Ania lubi muzykę*, the verb *lubi* will have the category *z/nn*, since it forms a sentence with two nouns, while in the sentence *Deszcz pada* the verb *pada* has the category *z/n*. Z. Pawlak stresses that a double syntactic category is related to the fact that a verb can perform two functions in a sequence: it can either denote relations between two objects or determine the state of one object. The roles of the verb in these two cases are quite different, and, obviously, the two types of verbs are not interchangeable.

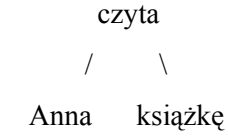
An adverb has the syntactic category:

$$\frac{z}{\frac{z}{n}}$$

since it forms a sentence out of a name and a verb of the *z/n* type. For example, the sentence *Deszcz bardzo pada* can be represented in this way.

In turn, conjunctions like: *i, lub, albo* have the category *z/z*, for they are capable of forming a sentence out of two sentences, etc.

The schema of the sentence *Anna czyta książkę* is the notation: n z/nn n. which can also be represented in the form of a tree:



or using parentheses, as: ((Anna) czyta (książkę)).

This implies that ,,Ajdukiewicz's syntactic category" depends on the connectivity of the verb in the sentence, and that each time it can have a different schema. And one more thing – all this is a sentential form, a verbal form, etc. – but not the meaning of those forms.

4.

The conceptions regarding classification of parts of speech presented above give rise to the question which of them is more useful in bi- and multilingual dictionaries. Is it a generalized classification which has a smaller number of classifiers and allows for placing lexemes representing various grammatical categories within one part of speech, or is it a more detailed, but also more precise classification, capturing also the meanings of language forms rather than the language forms alone? In our opinion, the detailed classification would be more helpful in machine translation. It would allow the dictionary to present various classification variants of words more precisely, as well as to show the meanings of the words, e.g. temporal ones, see Antoni Mazurkiewicz (in this volume). Due to this, some mistakes in machine translations could be avoided, see Ludmila Dimitrowa, Violetta Koseska (in this volume). All the researchers agree that word forms taken out of context usually have several interpretations. Hence an electronic dictionary should give as many interpretations as possible. The classification of words should make the classifiers more detailed by taking into consideration the semantics, whose role has been negligible up to now but without which one can hardly imagine a correct machine translation.

4.1.

On the example of expressions which quantify time and aspect in a natural language, we will show how one can classify words taking into consideration their quantification meanings. *Słownik języka polskiego PWN* [The PWN Dictionary of Polish] (2004) makes us realise how important it is to be able to identify temporal or quantification meanings and to distinguish them from the forms. Indeed, the above dictionary does not make a sufficient distinction between a verbal form and its temporal meaning, see a fragment of the *czas* (time; tense) entry: „Δ *jęz.* Czas przeszły «gramatyczna *forma* czasownika wyrażająca to, że *czas* dokonywania się czynności jest wcześniejszy od chwili mówienia o niej»". ("Δ *lang.* Past tense «grammatical *form* of a verb expressing the fact that the *time* when the action took place is earlier than the moment of speaking about it »"). As we can see, the past tense *form* has been equalled here with its *temporal meaning* (**time** is the **meaning** of the form and not the **form** itself**)**.

4.2.

On the sentence level, language expressions quantify time and aspect uniquely, existentially or universally. Many authors erroneously associate scope quantification with aspect only. The Bulgarian aorist form, independently of the information on the aspect, reserves the place for the **uniqueness** quantifier (operator) only. Quantification uniqueness is expressed using aorist forms of both perfective and imperfective verbs. And this proves that scope quantification refers to time rather than to aspect only; see the sentences below, incorrect in Bulgarian, where each of the aorist forms is in conflict with an existential or universal quantification expression:

* Той ход′и (aorist of an imperfective verb) там *понякога / винаги.*

* Той *понякога / винаги* замина (aorist of a perfective verb) за София.

* Той *понякога / винаги* се лекув′а (aorist of an imperfective verb).

4.3.

The distribution of the Bulgarian aorist of perfective and imperfective verbs is dictated by the language context where the uniqueness quantifier occurs. On the other hand, there are no limitations of this type on the Bulgarian imperfect and present – they are encountered in contexts with existential, universal and unique quantifiers. See

Той *понякога* ходеше до майка си в неделя. – He *sometime* visited his mother on Sunday. (existential quantifier)

Той *винаги* пътуваше за София с автобус. – He *always* went to Sofia by bus. (universal quantifier)

Точно в този момент той я обичаше още. – *Exactly at this moment* he still loved her. (unique quantifier)

4.4.

The typical lexical means for expressing existential quantification are Bulgarian words and word combinations: *понякога*, *някога*, *невинаги*, *отвреме-навреме*, *сегиз-тогиз*, *час по час* and their Polish analogues: *czasami*, *czasem*, *kiedyś*, *nie zawsze*, *od czasu do czasu*, *co jakiś czas*, *po jakimś czasie* and others. Some of these expressions are ambiguous with regard to the quantifier strength. They can have both **strong** and **weak** existential meanings, following from either the **external** or **internal** position of the quantifier in the semantic structure of the sentence (Koseska, Gargov 1990). In its strong quantification meaning, Bulg. *понякога* means 'in some cases' (*в някои случаи*, *има случаи да*, *случва се да*). Its Polish analogues: *czasami, nieraz, czasem* should be translated as `it happens that', `it can be that', e.g.:

Неведнъж е пял хубаво. – He often sang well.

Понякога пееше хубаво. – He sometimes sang well.

In the weak existential meaning, the Bulg. *понякога* corresponds to the Polish expression *od czasu do czasu* or *czasami*$_2$ (in the meaning *od czasu do czasu*):

По време на работа Мария понякога се оглеждаше разсеяно. – Od czasu do czasu słychać było grzmoty.

5.

General quantification of time can in turn be expressed either independently or by cooperation of the following means characteristic for both languages: verb aspects, verbal forms, lexical means helping to determine the quantification type, e.g. adverbs and various lexical constructions.

Typical language means for expressing universal quantification of time are both Bulgarian: *винаги*, *всеки път*, *навсякъде*, *никога не*, *никога*, *никъде*, *през цялото време* and their Polish analogues: *zawsze*, *za każdym razem*, *wszędzie*, *nigdy nie*, *nigdy*, *nigdzie*, *przez cały czas*, and others. Some of these expressions (e.g. *винаги* / *zawsze*) are ambiguous with respect to the quantifier strength, i.e. they can have both strong and weak universal meanings. More exactly, Bulgarian: *всеки път*, *навсякъде*, *никога не*, *никога*, *никъде* and Polish: *za każdym razem*, *wszędzie*, *nigdy nie*, *nigdy*, *nigdzie* have only **strong** meanings of universal quantification of time. On the other hand, Bulgarian *през цялото време* and Polish *przez cały czas* are only assigned **weak** meanings of universal quantification of time. The ambigiuity of *винаги* / *zawsze* as to the strength of quantifier can be shown on the following examples:

> Напишеше ли писмо, Иван винаги$_1$ го изпращаше по пощата. – Za każdym razem/zawsze$_1$, kiedy Jan kończył pisać list, wysyłał od razu go pocztą.
>
> Иван е бил винаги$_2$ мъж на Мария. – Jan zawsze$_2$ był mężem Marii.

In the first sentence, *винаги$_1$* / *zawsze$_1$* with the paraphrase ‘each time when …’ are exponents of strong meanings of universal quantification. In the second example, *винаги$_2$* / *zawsze$_2$* possess a different paraphrase: ‘without a break / all the time’ and express universal quantification within a unique state, here „being Maria’s husband”. The interchangeability of *винаги* / *zawsze* with, respectively, *непрекъснато* and *ciągle* / *bez przerwy* show that here we have to do with the second meaning of *винаги$_2$* / *zawsze$_2$*.

6.

The distribution of the expressions presented in Points 4-5 shows that one cannot talk about the meanings of selected lexemes without analysing the entire semantic structure of the sentence. This fact might be trivial, but it is nevertheless worth stressing, since in linguistics the semantics is traditionally still understood as lexical semantics only.

Let us note that the expressions described in Points 4-5 perform an identical semantically-syntactic function in the sentence – they quantify time (i.e. states and events) and refer to the predicate. From the logical viewpoint, they transform the predicate into a logical sentence. Due to the separate semantic and syntactic character of the quantifier (quantification expression) we propose that it be placed in an electronic dictionary. However, we would like to stress here that the subject is broader and can cover both quantitative quantification and quantification of nouns, which will be the subject of research in our future papers.

7.

Below we present the modifications of some entries existing in *Słownik Języka Polskiego PWN* (The PWN Dictionary of Polish) in an abbreviated form:

czasami$_1$ egzystencjalny kwantyfikator czasu «czasem, nieraz, niekiedy, w niektórych przypadkach, zdarza się, bywa»
(**czasami$_1$** existential quantifier of time «sometimes, more than once, on some occasions, in some cases, it happens that, it can be that»)

czasami$_2$ egzystencjalny kwantyfikator czasu «od czasu do czasu, co jakiś/pewien czas, zdarza się, bywa»
(**czasami$_2$** existential quantifier of time «from time to time, once in a while, it happens that, it can be that»)

czasem$_1$ egzystencjalny kwantyfikator czasu «czasami$_1$, nieraz, niekiedy, zdarza się, bywa»
(**czasem$_1$** existential quantifier of time «sometimes$_1$, occasionally**,** more than once, on some occasions, it happens that, it can be that»)

czasem$_2$ «przypadkiem, może»
(**czasem$_2$** «accidentally, occasionally, might»)

niekiedy egzystencjalny kwantyfikator czasu «co pewien czas, powtarzając się z przerwami, zdarzając się wielokrotnie (choć niezbyt często), w niektórych wypadkach; czasem$_1$; czasami$_1$»
(**niekiedy** existential quantifier of time «from time to time, repeating with breaks (though not too often), in some cases; occasionally$_1$; sometimes$_1$**)**

zawsze$_1$ ogólny kwantyfikator czasu «za każdym razem»
(**zawsze$_1$** universal quantifier of time «each time »)

zawsze$_2$ ogólny kwantyfikator czasu «przez cały czas, który się ogarnia myślą; wciąż, stale, nieustannie»
(**zawsze$_2$** universal quantifier of time «all the time that we can cover with our thoughts; at all times, constantly, continuously)

zawsze$_3$ «partykuła ekspresywna towarzysząca zdaniu lub jego części; bądź co bądź, pomimo wszystko, w każdym razie, jednak»
(**zawsze$_3$** «expressive particle accompanying a sentence or its part; nevertheless, despite all, anyway, yet»)

nigdy$_1$ ogólny kwantyfikator czasu «za każdym razie nie; w żadnej sytuacji, pod żadnym warunkiem, wcale, zupełnie»
(**nigdy$_1$** universal quantifier of time «each time not; in no situation, on no condition, not at all, absolutely not)

nigdy$_2$ ogólny kwantyfikator czasu «zawsze$_1$ nie, w żadnym przedziale czasu; wcale, zupełnie»
(**nigdy$_2$** universal quantifier of time «always$_1$ not; in no time interval; not at all, absolutely not)

Proposals for selected Bulgarian entries:
понякога$_1$ квантор на екзистенциалност за време «в някои случаи, има случаи да, случва се да»
(**понякога$_1$** existential quantifier of time «in some cases, there are cases such that, it happens that»)

понякога$_2$ квантор на екзистенциалност за време «често, редовно, няколко пъти»
(**понякога$_2$** existential quantifier of time «often, regularly, a number of times»)

винаги$_1$ квантор на всеобщност за време «всеки път, когато...»
(винаги$_1$ universal quantifier of time «each time, each time when... »)

винаги$_2$ квантор на всеобщност за време «през цялото време»
(винаги$_2$ universal quantifier of time «all the time»)

Proposals for selected entries in a bilingual Polish-Bulgarian dictionary:
czasami$_1$ egzystencjalny kwantyfikator czasu «czasem, nieraz, niekiedy, w niektórych przypadkach, zdarza się, bywa» (**czasami$_1$** existential quantifier of time «sometimes, more than once, on some occasions, in some cases, it happens that, it can be that»)
– **понякога$_1$** квантор на екзистенциалност за време «в някои случаи, има случаи да, случва се да» (**понякога$_1$** existential quantifier of time «in some cases, there are cases such that, it happens that»)

czasami$_2$ egzystencjalny kwantyfikator czasu «od czasu do czasu, co jakiś/pewien czas, zdarza się, bywa» (**czasami$_2$** existential quantifier of time «from time to time, once in a while, it happens that, it can be that»)
– **понякога$_2$** квантор на екзистенциалност за време «често, редовно, няколко пъти» (**понякога$_2$** existential quantifier of time «often, regularly, a number of times»)

zawsze$_1$ ogólny kwantyfikator czasu «za każdym razem» (**zawsze**$_1$ universal quantifier of time «each time »)
– **винаги$_1$** квантор на всеобщност за време «всеки път, всеки път, когато... »
(винаги$_1$ universal quantifier of time «each time, each time when.»)

zawsze$_2$ ogólny kwantyfikator czasu «przez cały czas, który się ogarnia myślą; wciąż, stale, nieustannie» (**zawsze**$_2$ universal quantifier of time «all the time that we can cover with our thoughts; at all times, constantly, continuously)
– **винаги$_2$** квантор на всеобщност за време «през цялото време»
(винаги$_2$ universal quantifier of time «all the time»)

zawsze$_3$ «partykuła ekspresywna towarzysząca zdaniu lub jego części; bądź co bądź, pomimo wszystko, w każdym razie, jednak» (**zawsze**$_3$ «expressive particle accompanying a sentence or its part; nevertheless, despite all, anyway, yet»)
– **все пак**

References

Ajdukiewicz K. (1960). *Język i poznanie* [Language and Cognition], PWN (Polish Scientific Publishers), Warsaw.

Bąk P. (1977/2007). *Gramatyka języka polskiego* [Grammar of the Polish Language], Wiedza Powszechna, Warsaw.

Bar-Hillel Y. (1953). *A quasi-arithmetical notation for syntactic description.* Language, 29: pp. 47 – 58.

Bogusławski A. (2005). *O operacjach przysłówkowych* [On adverbial operations], [in:] M. Grochowski (ed.), *Przysłówki i przyimki. Studia ze składni i semantyki języka polskiego* [Adverbs and Prepositions. Studies in Syntax and Semantics of Contemporary Polish], Toruń: UMK [Nicolaus Copernicus University] Publications: pp. 15-44.

GWJP / *Gramatyka współczesnego języka polskiego* [Contemporary Polish Grammar] (1984/1998). R. Grzegorczyk, R. Laskowski, H. Wróbel eds., Warszawa.

Duszkin M. (2008). *Wykładniki przybliżoności adnumeratywnej w języku polskim i rosyjskim* [Exponents of Adnumerative Approximability in Polish and Russian]. Doctoral Thesis Supervised by Prof. E. Janus.

Grochowski M. (1986). *Polskie partykuły. Składnia, semantyka, leksykografia* [Polish Particles. Syntax, semantics, lexicography.], Prace Instytutu Języka Polskiego PAN [Polish Language Institute PAS Reports], Wrocław: Ossolineum.

Grochowski M. (1997). *Wyrażenia funkcyjne. Studium semantyczne*. [Functional Expressions. Semantic study.] Kraków: PAN IJP [Polish Language Institute PAS].

Grochowski M. (2003). *Szyk jednostek syntagmatycznych w języku polskim (główne problemy metodologiczne)* [Order of Syntagmatic Units in Polish (main methodological problems)], Polonica 22-23: pp. 203-223.

Grochowski M. (ed.) (2005). *Przysłówki i przyimki. Studia ze składni i semantyki języka polskiego* [Adverbs and Prepositions. Studies in Syntax and Semantics of Contemporary Polish], Toruń: UMK [Nicolaus Copernicus University] Publications.

Koseska V., Gargov G.(1990). *Българско-полска съпоставителна граматика*, том 2. *Семантичната категория определеност-неопределеност*, София [Bulgarian-Polish Contrastive Grammar, vol. 2. Special Definiteness-Indefiniteness category, Sofia].

Pawlak Z. (1965). Gramatyka i matematyka, [Grammar and Mathematics] PWN, Warsaw: 5-44.

Przepiórkowski A. (2008). *Powierzchniowe przetwarzanie języka polskiego* [Surface Processing of Polish], Warsaw, s. 307.

Przybylska R. (2004). *Wstęp do nauki o języku polskim* [Introduction to Studies of Polish].

Saloni Z. (1974). *Klasyfikacja gramatyczna leksemów polskich* [Grammatical Classification of Polish Lexems], Język Polski 54.

Słownik języka polskiego PWN [The PWN Dictionary of Polish] (2004), version 1.0.

Wróbel H. (1995). *Problemy dyskusyjne w syntaktycznej klasyfikacji polskich leksemów* [Discussive Problems in Syntactic Classification of Polish Lexems]. [in:] *Studia gramatyczne XI*. Kraków: PAN IJP [Polish Language Institute PAS].

Wróbel H. (1996). *Nowa propozycja klasyfikacji syntaktycznej polskich leksemów*. [A New Proposal for Syntactic Classification of Polish Lexems] (in:) H. Wróbel (ed.) *Studia z leksykologii i gramatyki języków słowiańskich* [Studies in Lexicology and Grammar of Slavica Languages]. Kraków: PAN IJP [Polish Language Institute PAS].

Wróbel H. (2001). *Gramatyka języka polskiego* [Grammar of the Polish Language], Kraków.

Wajszczuk J. (1997). *System znaczeń w obszarze spójników polskich. Wprowadzenie do opisu* [The System of Meanings in the Area of Polish Conjunctions. Introduction to a Description], Warsaw.

Wajszczuk J. (2000). *A Polish conjunction on its way of paradox,* Linguistica Silesiana 21: pp. 35-41.

Zaron Z. (2003). *Funkcjonalna klasyfikacja leksemów polskich* [Functional Classification of Polish Lexems]. [in:] M. Gębka-Wolak, I. Kaproń-Charzyńska, M.Urban (red.) *O języku polskim nie tylko formalnie. Studia lingwistyczne dedykowane prof. Marii Szupryczyńskiej* [On Polish Not Only Formally. Linguistic Studies Dedicated to Prof. Maria Szupryczyńska], Toruń.

The Significance of Entry Classifiers in Digital Dictionaries[1]

Ludmila Dimitrova, Violetta Koseska-Toszewa
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences, Sofia,
Institute of Slavic Studies
Polish Academy of Sciences, Warsaw
ludmila@cc.bas.bg, amaz1312@gmail.com

Abstract
The paper discusses some problems related to entry classifiers in digital dictionaries. Information technologies offer great possibilities to linguists and lexicographers for the development of various dictionaries, especially for bi- and multilingual digital dictionaries. We use our experience from the development of a Bulgarian-Polish Digital Dictionary. We briefly present lexical specifications for Bulgarian in the EC international project MULTEXT-East, developed on the basis of a semantic and grammatical classification of Bulgarian wordforms.

Keywords: digital dictionaries, entry classifier, morphosyntactic description, state, event, corpus, lexicon, Bulgarian, Polish

Introduction: Basic Advantages of the Digital vs. Paper Dictionary

Information technologies offer great possibilities to linguists and lexicographers for the development of various dictionaries, especially for bi- and multilingual digital dictionaries.
First let us mention briefly the basic advantages of the digital vs paper dictionary. The preparation of the paper dictionary is a continuous process (it takes several months or even years) and the dictionary remains unchangeable after publication, i.e. the paper dictionary is a static collection of dictionary entries. The creation of a digital dictionary is also a continuous process in time, but the collection of words can be continuously expanded. New dictionary entries can be added or their content can be enriched by addition of supplementary information (grammatical, etymological) about the headword, of examples (for clarification of usage), of phrases and combinations, etc. The digital dictionary is a dynamic collection of dictionary entries, which provides a dynamical structure of the dictionary entry per se. This characteristic admits:

5. a relatively easy adaptation of the lexical database, which the collection of words in a dictionary actually is, to a new (improved) model of dictionary entry and its enrichment with new information, for example the addition of the word-forming group of the headword, etc.;
6. perfection of the system of classifiers, used for structuring the dictionary entry in order to describe optimally the headword;
7. use of the digitally-presented information for the creation of a new (or different type of) digital dictionary, for example two monolingual digital dictionaries (explanatory or terminological) in two different languages can be used to produce a new bilingual dictionary, although in practice that is non-trivial;
8. last but not least – correction of various mistakes if necessary.

Problems and Challenges

One of the main problems of the development of digital dictionaries is the *choice of classifiers* of the dictionary entry. Whenever the development of a system of bilingual digital dictionaries, serving as a basis for a system of multilingual dictionaries in perspective, is concerned, there arises an issue of *unification of the classifiers* in the dictionary entry. This is an *issue of harmonisation of the classifiers for various languages*, and the solution to this problem has to present a *unified selection of classifiers and a standard form of their presentation*. In a broader sense the issue of unification of classifiers in the dictionary entry *approaches the issue of a new part-of-speech classification* keeping in mind the specifications of a digital dictionary.

[1] The study and preparation of these results have received funding from the EC's Seventh Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX.

Classifiers

It is accepted that classifiers carry different morphosyntactic and/or semantic characteristics of the words (in particular, the dictionary entry). They split the set of words according to their properties.

Most often the classifier connects the word with its respective part of speech, depending on the class that the word belongs to. But the classifier can also show specific features of the word, such as gender, number, tense, etc. Tense is a meaning of the form, but has not been fully defined, see the examples about aorist (*аорист* in Bulgarian) and imperfectum (*имперфект* in Bulgarian).

At the current stage of research the part-of-speech classification in a natural language continues to be under discussion because it is not consecutive. It is based on different criteria (morphological, syntactic or "narrow" semantic) which are reduced only to the separation of grammatical categories. Thus the part-of-speech classification is different not only depending on language but is also significantly different in certain languages (See Koseska, Roszko 2008). This fact made us consider the unification of the part-of-speech classification at least in MONDILEX's six Slavic languages – Bulgarian, Polish, Russian, Slovak, Slovene and Ukrainian. In order to accept a common solution for the six languages, i.e. a standard type of part-of-speech classification, we start a discussion on these issues in this article. At the same time we contribute new arguments to this issue on Bulgarian and Polish material using F. Slawski's *Bulgarian-Polish Dictionary* (Sławski 1987) as well as the examples of machine translation from English to Polish and from English to Bulgarian.

So far the meaning of the forms has been the Achilles' heel of the description, dictionaries and corpora, both mono- and bilingual. That is why we shall focus our attention on some entries in the Bulgarian-Polish Dictionary depending on the form's meaning and its differentiation from a given meaning.

Examples

Let us have a look at the following examples of dictionary entries which do not explain anything in the dictionary. It is not clear whether they concern form or meaning. Neither is it clear what the meaning of this form is.

Example 1) Entry with headword "aorist"

аòрист, -и *m gram.* aoryst *m*

This entry with a headword 'the verbal form "aorist"' does not make clear what kind of aorist is meant. In Bulgarian aorist can be formed from perfective and imperfective verbs, for instance, ***написа*** and ***писа***. In the sentence "Той написа интересна книга" ('He has written an interesting book') the form "**написа"** is a perfective aorist. But the form "***писа"*** in "Той писа тази книга 5 години" ('He has been writing this book for 5 years') is an imperfective aorist.

Perfective aorist determines an event that has happened before the state of speaking and reserves a place for a unique quantifier in the semantic structure of the sentence. (See Koseska, Mazurkiewicz 1998; Koseska 2006).

Imperfective aorist means a configuration of states and events that have happened before the state of speaking and reserves a place only for a unique quantifier in the semantic structure of the sentence. (See Koseska, Mazurkiewicz 1998; Koseska 2006; Koseska, Roszko 2008 in this volume).

In order to describe these two different meanings of "aorist" we suggest the following two new dictionary entries ***Entry 1*** and ***Entry 2***.

Entry 1:

аòрист от свършен вид, -и *m gram.* – единично събитие настъпило преди състоянието на изказването. (In English: *A unique event that has happened before the state of speaking.*).

This meaning is conveyed by Polish perfective praeteritum (See Koseska 2006). For example:

Той боледува от грип.

On chorował na grypę.

Entry 2:

аòрист от несвършен вид, -и *m gram.* – единично квантифицирана конфигурация от състояния и събития, извършваща се преди състоянието на изказването (In English: *A unique-quantified configuration of states and events that have happened before the state of speaking*.).

This Bulgarian meaning is conveyed by Polish imperfective praeteritum. (See Koseska 2006). For example:

В четвъртък ходих пеша до центъра на града.
W czwartek chodziłam pieszo do centrum miasta.

Example 2) Entry with headword "imperfect":

Имперфект *m* gram. Imperfectum *n*

Just as in the case of **aorist**, we have no information that in Bulgarian this form (if form is meant here) is formed from imperfective as well as perfective verbs. We have no information about the difference in the meaning of the two. The imperfective imperfect serves to determine configurations of states and events that have happened and lasted before the state of speaking. The form here in contrast to the imperfective aorist (which is connected with a unique quantifier), reserves a place for all quantifiers (existential, universal, although rare, unique). (See Mazurkiewicz 2008 in this volume).

In this case our suggestion about the new entry with headword ***imperfective imperfect*** is the following:

Имперфект от несвършен вид, -и*, m gram*. Многозначно квантифицирана конфигурация от състояния и събития, настъпили и траещи преди състоянието на изказването – по значение съответства полската форма praeterium от несвършен вид (In English: *Multiply-quantified configuration of states and events that have happened and lasted before the state of speaking – by meaning corresponds to Polish imperfective praeterium.*). (See Koseska, Roszko 2008.)

For example:

Той понякога намираше време за разходка.
On od czasu do czasu znajdował czas na spacer.

Той понякога боледуваше от грип.
On czasem chorował na grypę.

Concerning the alternative "**Имперфект от свършен вид"** (***perfective imperfect***) we must note that it occurs very rarely and only in special modal, conditional contexts, such as:

Пийнеше ли (perfective imperfect), вдигаше (imperfective imperfect) много шум около себе си.

Example 3)

Let us consider the entry:

минал *part. adi* przezły, zeszły, ubiegły; **миналата година** dva lata temu; **-о време** *gram*. Czas przesły.

Here we have another type of problems. There are three Polish forms "**przezły**", "**zeszły**", "**ubiegły**" that correspond to the Bulgarian form "**минал" (*past*).** As in the case of "aorist" and "imperfect" it is not clear what is meant – meaning or form of past tense.

If a meaning is meant, it is not clear what past tense is meant. If however a form is meant, it must be mentioned that this is a form with multiple meanings.

We already mentioned (Dimitrova, Koseska 2008) that a single form can have multiple meanings and they naturally vary in number across the various languages. A solution to this problem would allow creation of a new $\mathbf{L_2}$-$\mathbf{L_1}$ dictionary from a $\mathbf{L_1}$- $\mathbf{L_2}$ dictionary. How do we invert a Bulgarian-Polish dictionary entry so that it represents a Polish-Bulgarian dictionary entry? It is obvious that the elimination of shortcomings among the entries of a given $\mathbf{L_1}$- $\mathbf{L_2}$ bilingual dictionary, eliminating the impossibility of a new ordering of information with the scope of obtaining an inverted $\mathbf{L_2}$-$\mathbf{L_1}$ bilingual dictionary, requires a reconsideration of the representation of the relation "form-meaning" in the dictionary.

An automated inversion of the dictionary is possible and easy to implement only when the relation "form-form" is considered. But then the inverted dictionary is quite poor and its cognitive value quite weak.

In order to keep all different meanings we suggest for discussion the option where each meaning is shown with the same form but enumerated, for example:

минал (1) – przezły
минал (2) – zeszły
минал (3) – ubiegły

In other words the form is indexed and appears in the list as many times as its different meanings.

Another example from the Bulgarian-Polish Dictionary – the dictionary entry for headword "**май"**:

май (1) *m* maj ; първи май – pierwszy maja;
май (2) *adv*. chyba, prawie, zdaje sie, prawdopodobnie.

Maybe in this case it is necessary to list this form a third time so that its third Polish meaning "prawie" corresponding to Bulgarian "почти" (*almost*) is listed as well:

май (3) *adv*. prawie.

Other examples:

независимост *f*, (1) – niepodległość *f*
независимост *f,* (2) – niezależność *f*

превежда/м, -ш *vi* (1) przeprowadzać
превежда/м, -ш *vi* (2) przekładać
превежда/м, -ш *vi* **пари** (3) przelewać (pieniądze)

кърп/а, - и *f* (1) ręcznik
кърп/а, - и *f* (2) ścierka
кърп/а, - и *f* (3) chustka

A short look at the Explanatory Dictionary of Bulgarian (reference) shows us the following two ways to describe homonymy:

(1) when the forms are different parts of speech, the difference in meaning is shown by indexing the different meanings

малко[1] *нарч.* ...в ограничено или недостатъчно количество...
малко[2] *ср.* Наскоро родено или излюпено същество...

or it is implied by listing the respective part of speech

май *м.* Петият месец на годината...
май *част.* За изразяване на предположение....;

(2) when the forms belong to the same class, the different meanings are indexed

мина[1] *ж.* ... рудник
мина[2] *ж.* ... снаряд
мина[3] *ж.* ... израз на лицето.

The usage of indexing for each meaning of a form (as in (2)) would allow the Bulgarian-Polish dictionary to be "inverted" and thus to obtain automatically a Polish-Bulgarian digital dictionary.

Whenever a bilingual digital dictionary is being complied, in the beginning the most common wordforms (parts of speech) are selected in a given digital corpus of $\mathbf{L_1}$ language. Then this frequency dictionary is completed with the translated equivalents from $\mathbf{L_2}$ language. We must mention here that besides frequency the forms may be selected according to a certain topic which contains them and which they describe. In other words, the dictionary may be compiled according to topics (thus combining topic and frequency).

Suggestions

Our suggestions can be grouped around the mode of form classification and the mode of writing the meanings of verb tense forms (two types with exact definition that can be "translated" in a formal language, for example, Petri nets).

We take a step back so to say from the "form-meaning" principle and limit ourselves to the "form-form" principle in bilingual dictionaries.

We suggest the headword form in the dictionary entry of the digital dictionary to be indexed according to the number of meanings, and each different meaning to 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 it is obvious that the greater number of classifiers provides a more adequate translation correspondence.

Bulgarian Experience

Traditional grammatical classifications for Bulgarian

Traditional Bulgarian grammar recognizes three main grammatical classifications:

- Semantic-grammatical – depending on the most general common meaning and on the grammatical properties words are ordered in classes, called parts of speech:
  - Nouns (a general terminological meaning of objects with common grammatical categories – "gender", "number", "definiteness"/"indefiniteness"),
  - Adjectives (have something in common in their lexical meaning, which is "indication, property, quality" of an object,
  - Verbs (common lexical meaning is "action or state" of a person or objects with common grammatical categories "tense", "person", "number", "mood", "voice"),
  - Numerals,
  - Pronouns,
  - Prepositions,
  - Conjunctions,
  - Interjections,
  - Particles.
- Morphological classification – according to the criterion "Open-class words or closed-class words":
  - Open-class words are nouns, adjectives, numerals, pronouns and verbs,
  - Closed-class words are adverbs, prepositions, conjunctions, interjections and particles.
- Syntactic (functional) classification – depending on whether the word functions in the sentence independently or not:
  - Nouns, adjectives, numerals, pronouns, verbs, and adverbs are independent,
  - Prepositions, conjunctions, and particles are dependent. The interjections are excluded.

Lexical specifications for Bulgarian in MULTEXT-East

The semantic-grammatical classification of the Bulgarian wordforms was used during the development of lexical specifications for the Bulgarian language in the EC project MULTEXT-East (Dimitrova 1998, Dimitrova et al. 1998).

In the MULTEXT-East project multilingual parallel (the translation of G. Orwell's "1984") and comparable (fiction and newspapers) corpora for six East-European languages – Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene – were developed and a lexicon was complied for each corpus and language.

The lexicons have been prepared in the form of lexical lists where each line contains one entry in the following form:

wordform <tab> lemma <tab> morphosyntactic description

Morphosyntactic description (**MSD**) contains encoding lexical specifications of the corresponding wordform ("wordform" represents an inflected form of the lemma). When the wordform (inflected form) coincides with its main form (lemma), then the entry "lemma" is replaced by "=".

The MULTEXT-East project has provided harmonised lexical specifications for the six East-European MTE languages and English. The specifications are presented as sets of attribute-values, with their corresponding codes used to mark them in the lexicons. The core features were determined (these features are shared by the most of the languages) and this provided the comparability of the information encoded in the lexicons across the MULTEXT-East languages. Except these "general properties" the so-called language-specific features were defined, which describe language-specific morphosyntactic phenomena.

Bulgarian MSD

Here we shall briefly present morphosyntactic description of the Bulgarian wordforms because these can provide useful information for digital bilingual Bulgarian-lang2 (digital bilingual dictionaries with Bulgarian language) as possible classifiers in the dictionary entry with regard to applications of digital dictionaries in machine translation systems, e-learning, etc.

MSD is defined as a linear string of symbols, representing the morphosyntactic descriptions, the positions of a string are numbered 0, 1, 2, etc. in the following way:

the symbol at position 0 encodes part of speech;

each symbol at position 1, 2, n, encodes the value of one attribute (person, gender, number, etc.);

if an attribute does not apply, the position in the string contains a hyphen "-".

Some examples of Bulgarian **MSDs**:

барабан	=	Ncms-n	(Noun, common, masculine, singular, no-definiteness)
барабани	барабан	Ncmp-n	(Noun, common, masculine, plural, no-definiteness)
барабани	барабаня	Vmia2s	(Verb, main, indicative, aorist, 2nd person, singular)
барабани	барабаня	Vmia3s	(Verb, main, indicative, aorist, 3rd person, singular)
барабани	барабаня	Vmip3s	(Verb, main, indicative, present, 3rd person, singul)
барабани	барабаня	Vmm-2s	(Verb, main, imperative, 2nd person, singular)

май	=	Ncms-n	(Noun, common, masculine, singular, no-definiteness)
май	=	Qgs	(Particle, general, simple)
май	мая	Vmm-2s	(Verb, main, imperative, 2nd person, singular)

мина	=	Ncfs-n	(Noun, common, feminine, singular, no-definiteness)
мина	=	Ncft	(Noun, common, feminine, count)

малки	малко	Ncnp-n	Noun, common, neutral, plural, no-definiteness)
малки	малък	A---p-n	(Adjective, plural, no-definiteness)
малките	малко	Ncnp-y	(Noun, common, neutral, plural, yes full_article)
малките	малък	A---p-y	(Adjective, plural, yes full_article)

Examples of Machine Translation

Let us have a look at some examples of machine translation, randomly picked from a web-page with an original text in the English language, which offers translation to Bulgarian, Polish and other languages.

The lack of adequate classifiers (or any classifiers) in the database (or in the digital dictionaries), used in the machine translation system, leads to the following translation mismatches:

First example

Original English text:

His play/direct partnership with the Scottish Chamber Orchestra has been particularly fruitful, and as well as touring extensively with the orchestra he has recorded a disc featuring Mozart's G major and D minor piano concertos.

Machine translation in Bulgarian:

Неговата игра/преки партньорство с шотландски камерен ансамбъл е било особено ползотворно, а както и още по обстойно с оркестър той е записано диск, с участието на Моцарт G големи и малки D пиано concertos.

Comment:

(For the sake of comparison – English translation (as far as it is possible) of the Bulgarian text: *His game/direct partnership with a Scottish Chamber Orchestra has been particularly beneficial, and as well as more extensively with an orchestra he was recorded a disc with the participation of Mozart G major and minor D piano concertos*.)

The errors in the machine translation of the sentences in the examples can be grouped as follows:

- first, the choice of lexical meaning for the translation is wrong:

play = изпълнение <-> игра= game
direct = ръководи, дирижира <-> пряк = direct, straight; immediate
fruitful = плодотворно <-> ползотворно = beneficial
featuring = включвайки <-> участието = the participation

- second, there is a lack of concordance between pronoun (as subject) and the verb form in the sentence:

he той (pronoun, masculine)
recorded записано (participle, neutral).

Machine translation in Polish:
Jego grać / bezpośrednej współpracy ze Scottish Chamber Orkiestra była szczególnie owocna, jak również szerokie tournee z orkiestrą ma zapisane dysk zawierający Mozarta G- dur i d – moll koncerty fortepianowe.
Comment:
The errors in this sentence are:

- "play" is translated as a verb infinitive due to lack of classifiers, in this case the English "play" is a noun, not a verb.
- the participle "zapisane" is neutrum and is not in accordance with the masculine noun "dysk".

Second example
Original English text:
Piotr Anderszewski was born in Warsaw to Polish-Hungarian parents.

Machine translation in Bulgarian:
Пьотр Anderszewski е роден във Варшава с полския-унгарски родители.
(For the sake of comparison – English translation (as far as it is possible) of the Bulgarian text: *Piotr Anderszewski was born in Warsaw with the Polish-Hungarian parents*.)
Comment:
There is a lack of concordance between qualifier and word that it qualify in the translation of "Polish-Hungarian parents" "с полския-унгарски родители".

Machine translation in Polish:
Anderszewski urodził się w Warszawa – Węgier do Polski rodziców.
Comment:
The error here is triggered by the preposition "to", to which only one meaning is given (from... Hungary to Poland). The English phrase "Polish – Hungarian parents" is not quite logical. Rather it should say "parts of Polish and Hungarian origin" or "Hungarian mother and Polish father".
Furthermore, Warszawa instead of Warszawie – lack of the form of locative case.

Third example
Original English text:
An exclusive artist with Virgin Classics since 2000, Anderszewski's first disc on the Virgin label was Beethoven's Diabeli Variations, a work which had already fascinated him for a decade.

Machine translation in Bulgarian:
Един изключителен артист с Вирджински класика от 2000 г. насам, Anderszewski първия диск на Богородица етикет е на Бетовен Diabeli варианти за работа, която вече е очарован му за едно десетилетие.
Comment:
(For the sake of comparison – English translation (as far as it is possible) of the Bulgarian text: *One exceptional artist with Virginia classic since 2000, Anderszewski first disc of Virgin Mary label is of Beethoven Diabeli work versions, which is already fascinated to him for a decade*.)

Machine translation in Polish:

Artysta na wyłączność z Virgin Classics od 2000 roku, Anderszewski pierwszy dysk na etykiecie Dziewicy było Beethovena Diabellego wariacje na pracę, która fascynowała go już od dekady.
Comment:
The errors in this sentence are:

- genitive case for Anderszewski in the sentence is missing.
- the participle "było" is neutrum and is not in accordance with the masculine noun "dysk".
- the noun "work" is translated as a "pracę", while the right translation is "tvorba".
- in the phrase "wariacje na pracę, która fascynowała go" the predicate is missing, it should be: "jest to dzieło, albo jest to utwór, który go fascynował...".

Fourth example
Original English text:
The 2008-09 season will see Anderszewski giving recitals at Carnegie Hall, Chicago's Symphony Center, the Walt Disney Concert Hall in Los Angeles and the Royal Festival Hall, London.

Machine translation in Bulgarian:
В сезон 2008-09 ще видите Anderszewski като точки в (*as points at*) Карнеги Хол, Чикаго на Симфония център (*Chicago of the Symphony Center*), Уолт Дисни Концертната зала в Лос Анджелис и Роял Фестивал Хол, Лондон.
Comment:
For the sake of comparison – English translation (as far as it is possible) of the Bulgarian text: *During the 2008-09 season you will see Anderszewski as points at Carnegie Hall, Chicago of the Symphony Center, the Walt Disney Concert Hall in Los Angeles and the Royal Festival Hall, London*.

Machine translation in Polish:
W sezonie 2009 – 09 będzie zobaczyc Anderszewski podając motywów w Carnegie Hall, Chicago of the Symphony Center, the Walt Disney Concert Hall in Los Angeles and the Royal Festival Hall, London.
Comment:
In the Polish translation "można" is missing, it should be "można będzie zobaczyć". Also the word "motyv" is not correctly translated.

Fifth example
Original English text:
Currently he lives in Paris and Lisbon.

Machine translation in Bulgarian:
В момента той живее в Париж и Лисабон.
Comment:
The translation was performed successfully.

Machine translation in Polish:
Aktualnie mieszka w Paryzu i Lisbona (correct "w Lisbonie").
Comment:
The locative case for "Lisbon" in the sentence is missing:

In Polish we observe the following mistakes:
-wrong gender,
-lack of cases,
-incorrect translation of tenses – see above the lack of "można",
-incorrectly translated prepositions,
-incorrect translation of lexical meanings (see above the example of "motyv").

There is not a single correctly translated sentence in the Polish text, in contrast to Bulgarian, but that is due to the analytical character of English and Bulgarian, whereas the Polish cases pose an additional difficulty to the translation software.

Concluding Remarks

In conclusion we want to emphasise that the unification of the classifiers of the dictionary entry will make electronic dictionaries more attractive. A dictionary with more classifiers will be significantly more useful to the user. The increase of the number of classifiers of the headwords in the entry will make machine translation more adequate and enrich electronic dictionaries. We believe that it is necessary to establish a possibility to obtain the inverse dictionary automatically. With traditional bilingual dictionaries this is impossible because of the polysemy of forms. Using the contemporary process theory (Petri nets theory) we suggest that dictionary entries related to time in a natural language render the content as well as the form. The content must reflect the main elements of time: the event, the state and the configuration of events and states (see above *Example 1 and 2*; and Mazurkiewicz 2008).

References

Dimitrova L., Koseska-Toszewa V. (2008) Some Problems in Multilingual Digital Dictionaries, *In International Journal Études Cognitives*, Vol. 8, SOW, Warszawa, pp. 237 - 254.

Dimitrova L. (1998). Lexical Resource Standards and Bulgarian Language. *In International Journal Information Theories & Applications,* Vol. 5, No. 1, 1998. pp. 27 – 34.

Dimitrova, L., Erjavec T., Ide N., Kaalep H-J., Petkevic V., and Tufis D. (1998). Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. *Proceedings of COLING-ACL '98*. Montréal, Québec, Canada, pp. 315-319.

Koseska-Toszewa V., Roszko R. (2008). Remarks on Classification of Parts of Speech and Classifiers in an Electronic Dictionary (to appear).

Koseska-Toszewa V. (2006). Bułgarsko–polska gramatyka konfrontatywna, t. VII. Semanticzna kategorija czasu. SAW, Warszawa, Polska.

Koseska-Toszewa, V., Mazurkiewicz A. (1988). *Net representation of sentences in natural languages,* Advances in Petri Nets, LNCS 340, Springer Verlag, pp. 249–259.

Mazurkiewicz A. (2008). A Formal Description of Temporality (Petri net approach) (to appear).

Sławski F. (1987). Podręczny słownik Bułgarsko-Polski z suplementem. Warszawa, Polska.

A Formal Description of Temporality (Petri Net Approach)[1]

Antoni Mazurkiewicz
ICS PAS Warsaw
antoni.mazurkiewicz@ipipan.waw.pl

Abstract
The aim of this paper is to present a Petri net formalism and to show how it can be used for defining temporal situations. Reichenbach's schemes have been thought to be used for the same purpose. It is clear that the net formalism covers the formalism of Reichenbach, treating points on a number line as events and intervals as states; however, Reichenbach's formalism does not cover independency of events, uncertainty of sequencing events and states, and various aspects of modality. Thus, the net formalism can be viewed as an essential extension of Reichenbach's one. The scope of this paper is limited to a presentation of the net formalism; it was not intended to analyze temporal situations from the linguistic point of view. Presentation of using the net formalism in connection with some specific linguistic phenomena is expected in forthcoming parts of this study.

Keywords: temporality, grammatical tenses, Reichenbach schemes, Petri Nets

1. Temporality Description Issues

The main difficulty in proper translation of temporal and modal phrases (expressions) consists in an imprecise description of situations described by these phrases. Additional difficulty is caused by the fact that different languages exhibit a variety of different means used for the same situations and there is a great variety of temporal situations expressible directly in one language and not expressible in another. Clearly, a faithful translation of temporal phrases is of a primary importance, hence there is an urgent need for a background of strict and reliable temporality description. The main issue discussed in this paper is how to describe temporality dependencies and which are necessary means to grasp and express given temporal situations. There are several possible approaches to these questions.

a. Explaining temporal situations formulated in a language by their detailed descriptions expressed in the same language (self-explaining);
b. Expressing temporality by equivalence of phrases in different languages (the question: "what does it mean" is replaced by "how it is expressed in another language"). Then temporality is an abstraction induced by all temporarily equivalent phrases;
c. Expressing temporality by an *inter-language* (an "in between" language), where the meaning of temporality is supposed to be known;
d. Creating some formal models of temporal situations and then comparing how they are described in different languages.

In the present paper the last approach is discussed. The aim of the paper is to present two formal models of temporality, creating background of understanding the temporal situations and meeting the following requirements.

- **Directedness**. The scope of the description possibilities should be limited to temporal situations only. The intention is to get rid of unnecessary linguistic phenomena that can obscure the essence of temporality features.
- **Completeness**. The required model should cover all possible temporal (and modal) situations, leaving no room for imprecise and intuitive interpretations or for a necessity of relying on some hidden assumptions.
- **Independency.** The model description language should not use the linguistic temporal means specific for different languages; instead, it should have its own formalism of description, not relying on specifications introduced by existing natural languages.
- **Simplicity**. The model structure should be simple enough to guarantee its proper understanding by languages users.

[1] The study and preparation of these results have received funding from the EC's Seventh Framework Programme [FP7/2007-2013] under grant agreement 211938 MONDILEX.

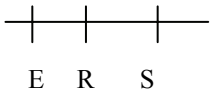
- **Applicability**. The model should be applicable for possibly large number of situations that can be described in natural languages. Some temporal situations that can be distinguished in some natural languages may be not such in other languages; therefore, the required models should be capable to cover all of them.

Below, two such formal models are presented and compared. Both of them use graphical representations of temporality. The first one is so-called *Reichenbach's model*, formulated in his book *Elements of Symbolic Logic* (Reichenbach 1944); the second one is so-called *net model*, or *Petri net model*, formulated by Carl A. Petri (1962) and described in (Koseska-Toszewa, Mazurkiewicz 1988).

2. Reichenbach's Representation of Temporality

Reichenbach's model of temporality phenomena was the first graphical representation of temporal relationships between a speaking subject and described objects occurring in sentences in natural languages. The basis for this model is a straight line, representing time scale, running from the left to the right, and some points on it, representing moments occurring in the reality described by the analyzed sentence. Among them, three points are distinguished: the point of utterance (corresponding to the moment of speech), the point of event (corresponding to the moment the statement is referring to), and the point of reference (the moment to which all other moments of the described situation are referred to).

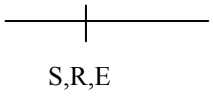
The following Reichenbach's schemata can (taken literally from his paper) serve as examples of using his model for explaining some temporal dependencies. Let consider simple variations based on phrase "to see John", expressed in different moments and referring to different moments. According to (Reichenbach 1944) , we have the following descriptions:



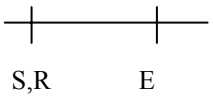
I had seen John

I saw John

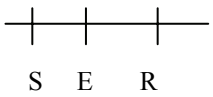
I have seen John



I see John

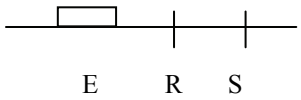


I shall see John

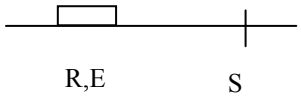


I shall have seen John

For completeness, Reichenbach introduces an additional graphical symbol to indicate time duration of some events, as is shown below:

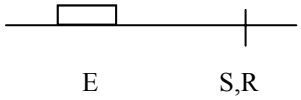


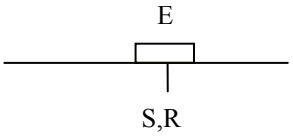
I had been seeing John



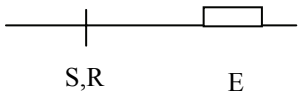
I was seeing John

I have been seeing John

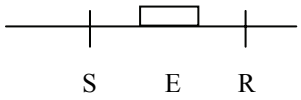




I am seeing John



I shall be seeing John



I shall have been seeing John

Actually, the diagrams shown above are not radically different from those given in the previous scheme; the only difference is that point of event (E) is not a moment anymore, but it is a period of time. In such a way it introduces a continuity of events, expressed by continuous tenses of English.

The Reichenbach's model can be summarized by the following table (remembering that events may be points as well as segments of a time line):

3.

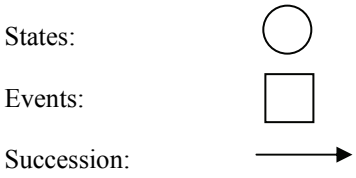
Time ordering	**Tense**
$\{e, r, s\}$	Present simple
$e \rightarrow \{r, s\}$	Present perfect
$\{s, r\} \rightarrow e$	Future simple
$s \rightarrow \{r, e\}$	
$s \rightarrow e \rightarrow r$	
$\{s, e\} \rightarrow r$	Future perfect
$e \rightarrow s \rightarrow r$	
$\{e, r\} \rightarrow s$	Past simple
$e \rightarrow r \rightarrow s$	Past perfect

Petri Net Approach

Reichenbach's temporal scheme is perfectly suited for English grammatical tenses description. For more general purposes, as e.g. for comparing grammatical means in different languages, more general framework is needed. Such a method for describing real **temporal** schemes on a very basic level, precisely defined in a formal (natural language independent) way has been formulated by C.A Petri (1962). Since then, it is widely used for many purposes and interpreted in a number of different ways. From the name of its author, models using formalism introduced by Petri, are called Petri Nets. For linguistic purposes, the model presented below seems to fulfill requirements of temporality description. It is built from the following basic notions:

states, situations expressed in sentences, directly or indirectly
events, initiating or terminating states,
succession, a relation binding events with states that are initiated or terminated by them,

represented graphically by circles, boxes, and arrows:



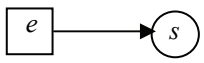
States, describing situations we are talking about in everyday language, are properties of **objects** (or collections of objects). Examples of states are: "*the door is open*" or "*the door is closed*". The characteristic feature of states is their extension in time – a state is a property holding during some amount of time. States can be permanent, without beginning or ending (then lasting infinitely long), or temporary, holding a finite amount of time.

Events are changes of situations; as such, they are momentary, taking no time – they can only occur in some moments. Example of an event is a change the state of the door from "open" to "closed" The characteristic feature of any event is that it happens either in the past or in the future with respect to any chosen moment (saying "*e* is happening" we have in mind a collection of events, not a single one).

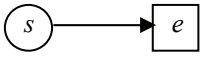
Succession is a relation between states and events establishing which events initiate (or terminate) which states. This relation determines a flow of time in the model in such a way that the beginning of any state always precedes (in time) its ending. The succession can be also treated as a causal relationship between elements described by nets.

States and events are fundamental concepts of the Petri nets theory; their causal (or temporal) ordering is the main issue discussed in terms of nets. In general, to describe real situations one needs a number of states and events, bound together by succession relation. Graphical representation of such a temporal scheme is a finite directed graph, called a *net*, with circles and boxes (as its nodes) joined with arrows (directed arcs). This graph is bipartite, i.e. any arrow leads either from a box to a circle or from a circle to a box; neither two boxes nor two circles are joined by an arrow. For the present purposes the graph is additionally assumed to be finite and containing no cycles (without elements joined with themselves by a sequence of arrows).

The basic constructs used in Petri nets are:



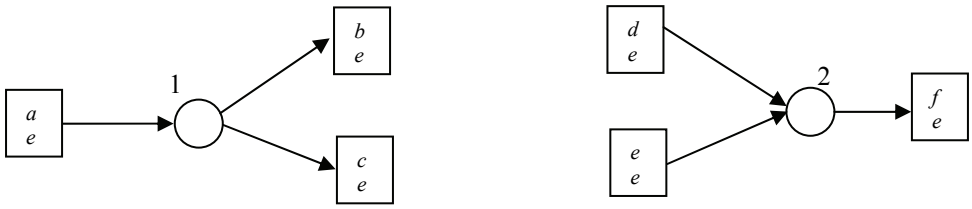
State *s* is initiated by event *e*



State *s* is terminated by event *e*

Nets arise by arbitrary combinations of the above constructs. Basic (and simple) combinations of them are:

Both diagrams describe actions *e* and *c*, the first action terminates state 1 and initiates two (coexistent) states 2 and 3, the second action terminates two (coexistent) states 4 and 5, and initiates state 6. The diagrams below describe other situations:



The first diagram represents state 1 initiated by event *a* and terminated by exactly one of mutually excluded events, namely event *b* or event *c*; the second diagram represents state 2 initiated by exactly one of mutually excluded events *d* and *e*, and terminated by event *f*. In this way nets admit a possibility to deal with some alternative actions. There is a similarity of states, events, and their mutual relationship to intervals, points, and their relationship on the number line. Namely, any event begins a state in the same way as a point starts an interval. Points begin some intervals which in turn are ending with some points. Any interval of the number line is terminated or initiated by a single point; on the other hand, any point can begin or end many intervals of the line.

4. Histories

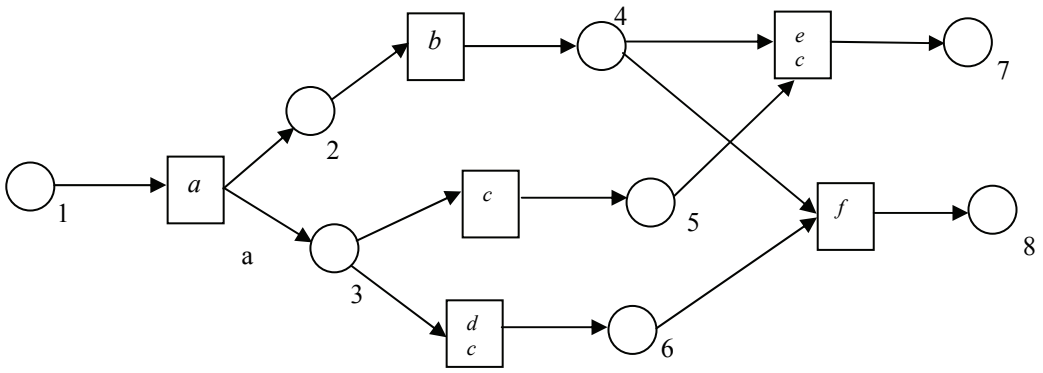
Nets describe temporal structures of pieces of reality, limited to states, events, and their mutual relationship. They can describe a single course of actions as well as a number of such courses, depending on different possibilities or conditions. Any specific course of actions defined within a net is a single *history* supported by the net. There can be one or more different histories supported by the same temporal scheme; histories can engage only a part of the scheme. In any case, any history must respect the succession relation between events and states defined by the net. For the time being, as it has been mentioned above, the temporal schemes without repetitions are considered; nets with repetitions, i.e. containing cycles, will be discussed in a separate paper. For the needs of the present aims the following definition is sufficient: a history is a connected part of temporal scheme such that:

1. Any state of the history is initiated or terminated by at most one event of this history,
2. All states initiated or terminated by an event in the history belong to this history.

However, in a history a single event can initiate or terminate a number of states (as a single point on a number line can start or end a number of intervals).

In general, temporal structures can contain a number of different histories, representing various possibilities of the course of actions. It is reflected in the net model of such structures by presence of states ending (or beginning) by a number of events excluding each other. In a history singled out from such a structure, the termination or initiation of all its states is determined. In any history, two events can either precede each other, or can be independent –happening independently of each other; similarly, two states can either precede each other (i.e. the ending of one of them precedes beginning

of the other), or can be coexistent – both of them exist at some interval of time. The best way to explain these relationships is to discuss an example of a complex temporal scheme. The diagram given below represents a net containing more than one history. It consists of 8 states, 6 events, and 15 arrows representing the succession relation.



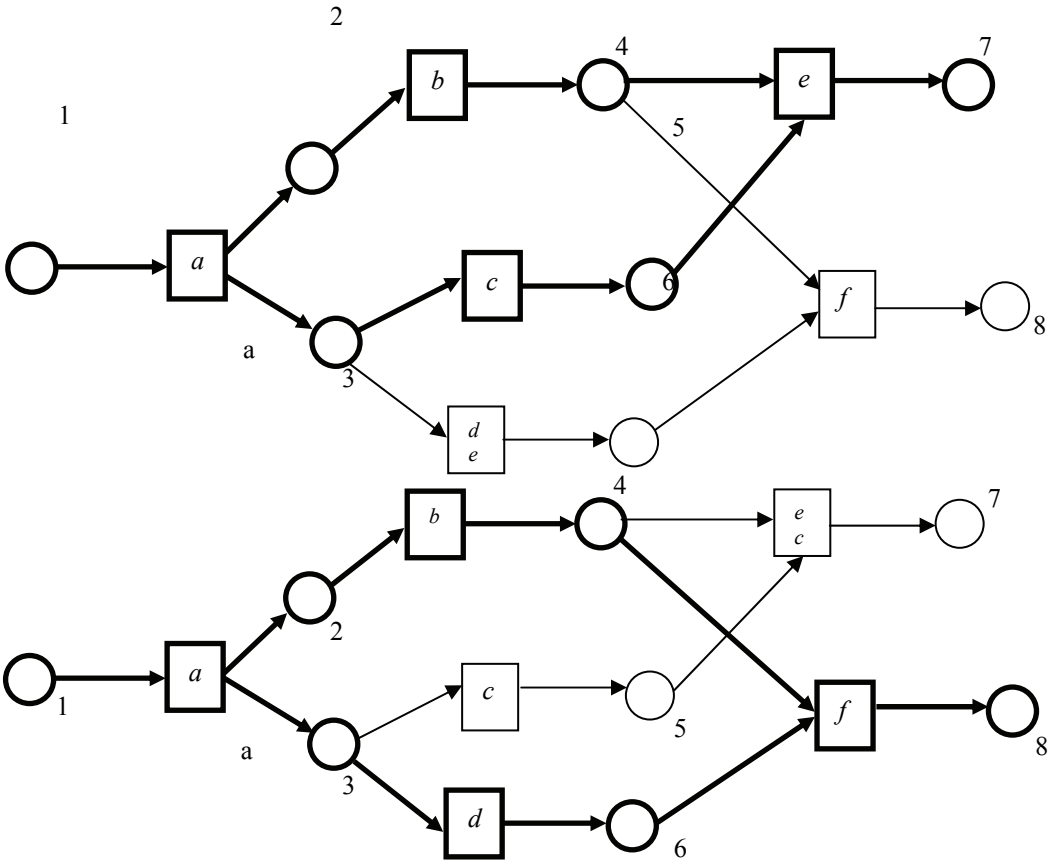
Event *a* terminates state 1 and starts two states, 2 and 3. States 2 and 3 are coexistent, as started with the same event. State 3 can be ended by two events *d* and *e* (mutually excluding each other). State 2 is terminated by event *b* initiating in turn state 4. If state 3 is terminated by *c*, state 5 begins to exist, as initiated by *c*; otherwise (if state 3 is terminated by *d*) state 6 begins to exist. States 4 and 5 are coexistent and terminated by event *e.* Similarly, states 4 and 6 are coexistent and terminated by common event *f*. Events *c* and *d* exclude each other; hence, states 5 and 6 are not coexistent, as initiated by mutually excluded events and then also exclude each other. However, state 4 is coexistent with 5 as well as with 6. Coexistent states 4 and 5 are closed by event *e*, and coexistent states 4 and 6 are terminated by event *f*. Events e and *f* are excluding each other, since state 4 can be terminated by exactly one event, either *e* or *f*, but not by both of them. Consequently, states 7 and 8 are not coexistent.

This is an abstract explanation of the above net structure. To be more specific, assign the following meaning to states and event of the presented net. Namely, interpret it as a (fragment of) a real life reviewing procedure of a paper submitted for a publication, with states and events explained given in the tables below.

STATES	
1	Preparing paper
2	Paper is ready
3	Preparing evaluation
4	Waiting for opinion
5	Opinion is positive
6	Opinion is negative
7	Expecting publication
8	Thinking about corrections

EVENTS	
a	End of preparing paper
b	Start waiting for opinion
c	Taking positive decision
d	Taking negative decision
e	Sending paper to publisher
f	Rejecting the paper

Two diagrams below represent two possible histories contained in the above scheme. The first one is “optimistic”, and in the end the paper is accepted, the second one is “pessimistic”, and in the end the paper is rejected.



Analysis of the above histories in terms of the proposed interpretation is left for the interested reader.

Such an annotated net is a description of a temporal situation in a way independent of linguistic properties as well as of peculiarities of different languages.

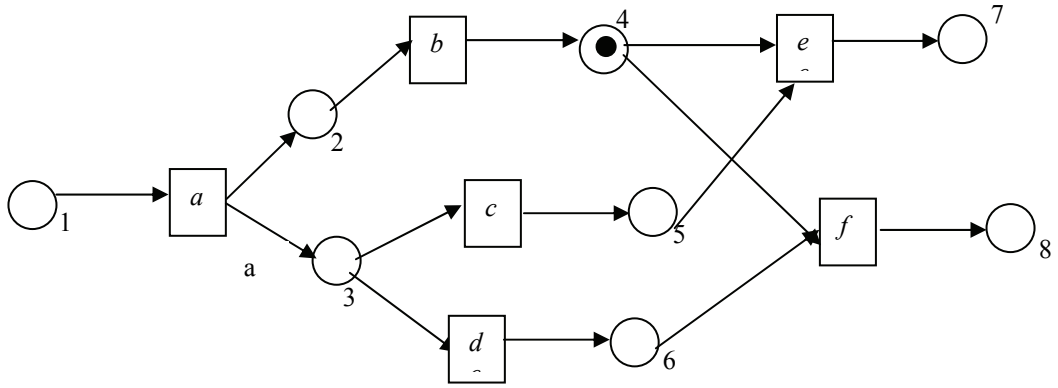
5. State of Utterance

The main objective of net schemes presented here is a description of temporal properties of phrases in a similar way to Reichenbach’s line sketched above. To explain a phrase expressing a temporal situation, one has to know objects (states and events) the phrase is referring to, and a state of an utterance subject. Their mutual combination determines linguistic means adequate to the described situation. The proper understanding of the phrase describing a given situation depends on proper choice of its net representation. Once the situation is characterized by a net with a given state of utterance, phrases of different languages expressing this situation can be compared and analyzed, using the net scheme as a bridge joining different formal means specific for compared languages.

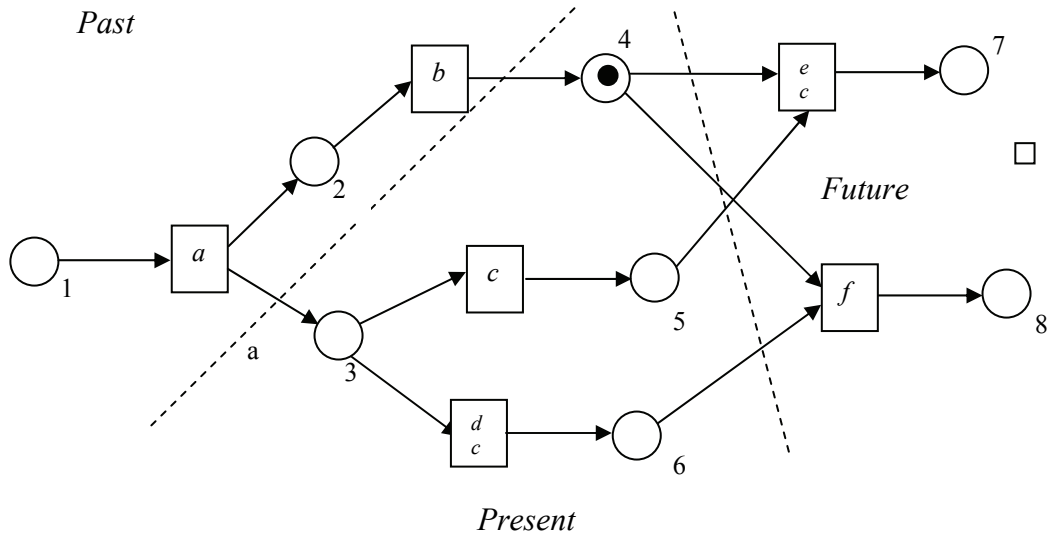
To this end, one has to construct a net comprising objects of utterance, i.e. to assign objects to states or events of the net, and to chose a state of utterance. Then the temporal meaning of the analyzed phrase is completely defined. Placing in a net scheme the state of utterance (i.e. choosing a state of the scheme where the speaker is situated) has an essential influence on the grammatical form of the analyzed sentence, similarly as it has been done earlier using Reichenbach’s schemata. In the net scheme, placing the state of utterance can distinguish the actual history from any other which is impossible in the accepted course of action. Namely, by introducing the state of utterance, the net scheme is split into:

(1) the present, past, and future of the history (histories), and
(2) the possible and impossible situations of distinguished histories.

In graphical representation used here the place representing the state of utterance is marked by a dot. Some examples of the net representation of some temporal situations are given below. Consider the net discussed above with 4 as the state of utterance.

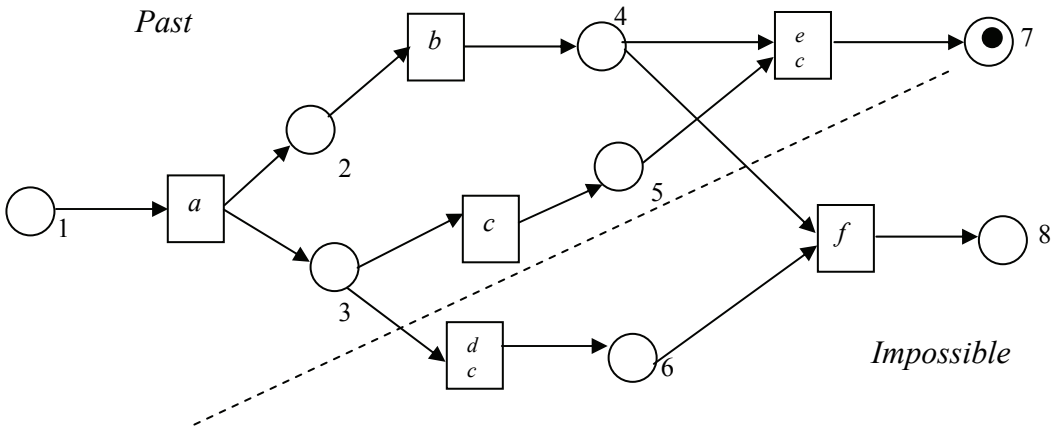


Then the whole net is partitioned into three parts:



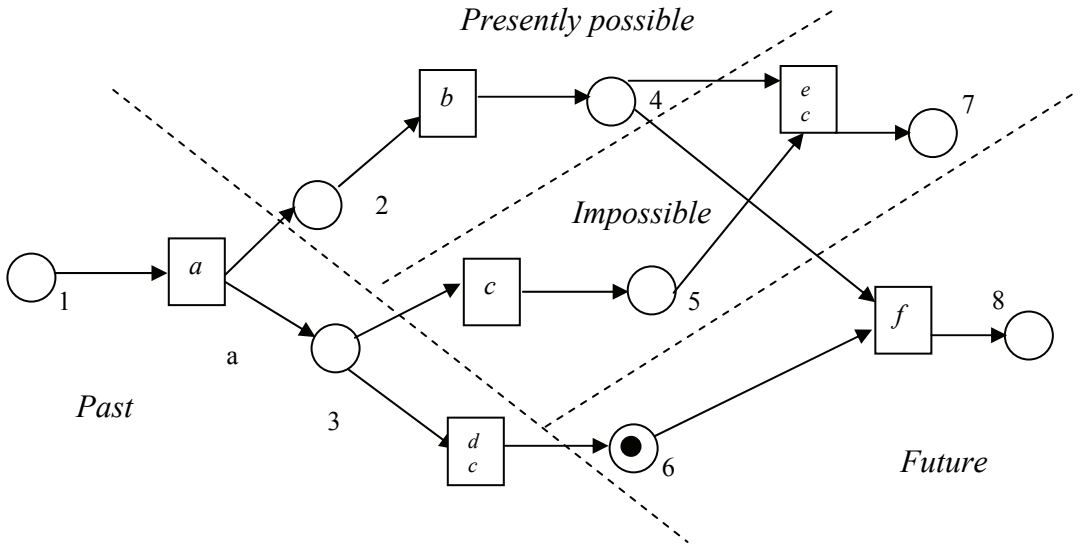
From the point of view of state 4 (the state of utterance) states 1, 2, and events *a*, *b* are in the past. Events *e* and *f* are in the future; the future is uncertain, since only one of the two can happen. It depends on the present, which is also uncertain; the speaker does not know whether state 3 or one of states 5, 6 is holding now. Moreover, the speaker does not know which one of *c*, *d* will happen or has already happened. In other words, at state 4 the speaker does not know which history is going on.

The next diagram shows the same structure, but with the state of utterance placed in state 7.



Then states 1, 2, 3, 4, 5 and events *a, b, c, e* are in the past, while states 6 and 8 will never take place, although they would be possible if *d* rather than *c* had happened in the past.

Placing the state of utterance at place 6 we have the partition of the considered scheme into four parts: past, future, impossible, and possible at present:

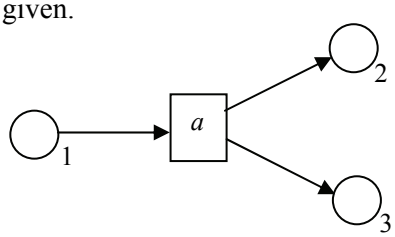


Term "presently possible" means here that, from the point of view of state 4 (of utterance), the speaker does not know whether it is now 2 or 4, but certainly one of them. As for event *b,* the speaker can be sure that either it has happened (and then state 4 is now) or it has not happened yet (and, consequently, it is still state 2).

6. Enhancing Nets

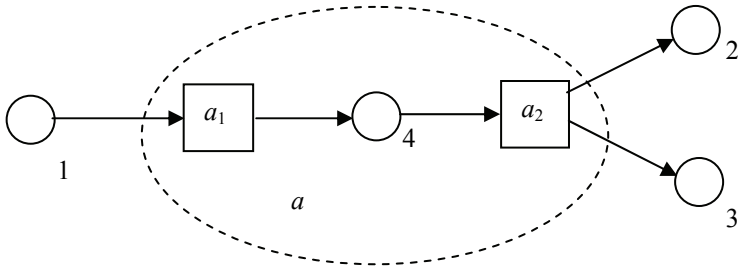
Two properties of net descriptions are worth to be mentioned. First, in order to a faithful description of situations, some additional event and states, not mentioned explicitly in their descriptions, should be inserted into the net. Inserting them into the net serves to proper sequencing of remaining states and events. Secondly, the net description can be made more or less precise, depending on the description purposes. Sometimes an event should be refined to more detailed

structure, as it is shown in the following simple example. The net describes a very simple situation of a person (for example, John) who leaves his home and goes to his office. The following states of John are taken into account: 1. John is at home; 2. John is outside of his home; 3. John is on his way to his office. The single event binding the above states is leaving home (event *a*). Below the net describing the above situation is given.



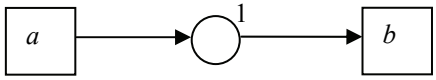
It says that the state "staying at home" (1) is terminated by action of leaving home (*a*) and two new states are initiated: "to be outside home" (2) and "to be on the way to John's office"(3).

One can contest the qualification "leaving home" as a momentary event without any duration. Then one can refine the above scheme to the following:

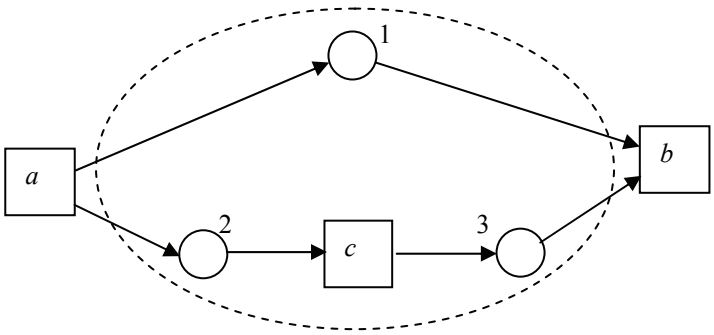


where event *a* ("leaving home") has been split into two, more specific events, a_1 ("begin of leaving home") and a_2 ("end of leaving home") and a new state 4 ("action of leaving home") has been introduced. In this way, "leaving home" lost attribute of being an event and became a state. It is a general situation: in order to be more specific, states and events can be refined, enhancing corresponding net. Thus, net descriptions can be enhanced by adding some new objects and by refining existent objects.

A similar possibility we have in case of states. Consider the scheme



State 1 can be viewed as e.g. "to be on holidays", event *a* is the beginning of holidays, *b* is their end. To be more specific, one could state explicitly that the first part of holidays are to be spend at the Baltic sea (state 2), while the rest at home (state 3). Then enhanced net would look like that:



Event *c* in the above diagram separates states 2 and 3 and the nature of *c* is left unspecified. In effect of this transformation, the scheme has been enhanced by adding new elements.

Both transformations given above make nets more specific; they are called net *refinements*. There are possible transformations in the opposite direction, making nets less specific, to avoid some unnecessary details. Such transformations are called net *abstractions*. Both of them allow us to tailor net schemes exactly to the description needs.

References

Koseska-Toszewa, V., Mazurkiewicz A. (1988) *Net representation of sentences in natural languages,* Advances in Petri Nets, LNCS 340, Springer Verlag, pp 249–259.

Petri, C.A. (1962) *Fundamentals of the Theory of Asynchronous Information Flow,* Proc. of IFIP Congress 62, Amsterdam: North Holland Publ. Comp., pp. 386-390.

Reichenbach, H. (1944). *Elements of Symbolic Logic*, New York, McMillan Publ. Comp.

Reisig, W. (1985). *Petri Nets – An Introduction,* New York, Springer Verlag.

Authors

Igor Boguslavsky: head of Laboratory of Computational Linguistics; Kharkevich Institute of Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia; professor; Madrid Polytechnic University, Madrid, Spain.

Ivan Derzhanski: senior researcher, Department for Mathematical Linguistics, Institute for Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria.

Vyacheslav Dikonov: researcher of Laboratory of Computational Linguistics; Kharkevich Institute of Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia.

Ludmila Dimitrova: associate professor, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria.

Tomaž Erjavec: researcher at the Department of Knowledge Technologies at the Jožef Stefan Institute, Ljubljana, Slovenia.

Radovan Garabík: researcher, Slovak National Corpus department Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia.

Leonid Iomdin: acting head of Laboratory of Computational Linguistics and leading researcher; Kharkevich Institute of Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia.

Jan Jona Javoršek: researcher at the Department of Experimental Particle Physics at the Jožef Stefan Institute, Ljubljana, Slovenia.

Violetta Koseska-Toszewa: professor, head of Department of Semantics, Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland.

Natalia Kotsyba: assistant professor, researcher; Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland.

Antoni Mazurkiewicz: professor, Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland.

Petya Osenova: researcher of Linguistic Modelling Laboratory, Institute for Parallel Processing, Bulgarian Academy of Sciences and assistant professor in Linguistics, the Sofia University, Sofia, Bulgaria.

Radoslav Pavlov: associate professor, PhD, head of Mathematical Linguistics Dept., deputy director of Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria.

Roman Roszko: researcher of Department of Semantics, Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland.

Volodymir Shyrokov: director of Ukrainian Lingua-Information Fund, National Academy of Sciences of Ukraine, Kiev, Ukraine.

Kiril Simov: senior researcher of Linguistic Modelling Laboratory, Institute for Parallel Processing, Bulgarian Academy of Sciences, Bulgaria, Sofia.

Victor Sizov: researcher of Laboratory of Computational Linguistics; Kharkevich Institute of Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia.