National Academy of Sciences of Ukraine
Ukrainian Lingua-Information Fund

Organization and Development of Digital Lexical Resources

MONDILEX Second Open Workshop
Kyiv, Ukraine, 2–4 February, 2009

Proceedings

Volodymyr Shyrokov, 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

Kyiv 2009

Organization and Development of Digital Lexical Resources.
Kyiv, ULIF NAS, 2009, – 128 p.

The volume contains contributions presented at the Second Open Workshop "Organization and Development of Digital Lexical Resources", held in Kyiv, Ukraine, on 2–4 February, 2009. 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 program.

Editor of the volume: Dovira Publishing House, Kyiv, Ukraine
Computer design: Maxim Krygin

ISBN 978-966-507-252-2

Contents

FOREWORD

This volume contains contributions presented at the MONDILEX project second open workshop "Organisation and development of digital lexical resources", held in Kyiv in February 2009. 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), a project developed under EC's Seventh Framework Programme.

The purpose of this Workshop was to study the state of the art in mono-, bi- and multilingual Slavic digital (local and on-line) lexical resources (corpora and dictionaries) and requirements for their integration, and to formulate expert recommendations for the standardisation and integration of multilingual Slavic lexical resources and their opening to research, education, business and the public. The papers discuss current trends and achievements in the field of digital lexicography, especially for Slavic languages. We do hope the volume will be of interest to both lexicographers and computer scientists.

The first part of the volume "Information Technologies for Supporting Research Activities in Digital Lexicography", discusses achievements of the information technologies and their applications for supporting the language technologies. The paper by V. Shyrokov presents the National Dictionary Base of Ukraine in the context of the MONDILEX project. A number of digital lexicographical systems and so-called lexicographic virtual laboratories have been created that can provide professional interaction of remotely situated scientists via Internet when developing joint lexicographic projects. The paper by E. Martynov is dedicated to the Ukrainian Academic Grid, a powerful computing resource for fundamental and applied scientific research. The concepts of virtual lexicographic system are considered in the paper by O. Rabulets.

The second part of the volume is "Digital Lexicographic Resources (Corpora and Dictionaries) and their Application". The paper by S. Krek and T. Erjavec presents a proposal for lexical encoding concentrating on morphological properties of words, with special emphasis given on the rich inflectional properties of Slavic languages. The paper by L. Dimitrova and P. Rashkov also deals with morphological properties and their descriptors for Bulgarian, namely proposing new attributes in the morphosyntactic description of participles. The paper by L. Dimitrova, R. Garabík, D. Majchráková analyses the differences between the morphology specifications for MULTEXT-East of Bulgarian and Slovak languages. All parts of speech are described in detail with emphasis on the analysis of tagset differences. The paper by L. Dimitrova, V. Koseska, I. Derzhanski, and R. Roszko describes a comparison of the morphosyntactic characteristics of the words of the first Bulgarian-Polish parallel corpus from the point of view of a prospective unification. The current state of work on the Polish-Ukrainian Parallel Corpus POLUKR is shown in the paper by N. Kotsyba. The problems of creation of a Polish grammatical dictionary are discussed in the paper by I. Shevchenko. The digital etymology illustrated by the example of the Etymological Dictionary of Ukrainian language is described in the paper by I. Ostapova. The paper by T. Lyubchenko describes the grammar dictionary modelling of Russian. A frequency dictionary of Finnish word building is briefly described by K.Tyschenko and B. Rudyj.

The third part of the volume is dedicated to linguistic and mathematical foundations that support language resources. A multi-volume Contrastive Grammar of Bulgarian and Polish is presented in the paper by V. Koseska. The paper by V. Koseska and A. Mazurkiewicz discusses net-based description of modality in natural language (on the example of conditional modality). Statistical methods used for comparison and analysis of texts are presented in M. Krygin's paper.

The fourth part of the volume is "Problems of Etymology". V. Luchyk's paper is dedicated to the creation of etymological dictionary of suffixes of Ukrainian language. The current trends in the reconstruction of common Slavonic lexis are discussed in T. Chernysh's paper. The problems of irregular phonetic phenomena in languages (so called delabialization *l'u- › *li-) are presented by V. Shulhach.

The workshop in Kyiv has been highly useful and efficient. The editors hope that the presented contributions will be of interest to both lexicographers and computer scientists.

Volodymyr Shyrokov, Ludmila Dimitrova

I. Information Technologies for Supporting Research Activities in Digital Lexicography

Experience in Creating a National Dictionary Depositary of Ukraine and its Use in Conceptual Modelling of Networking of Centres for High-quality Research in Slavic Lexicography and their Digital Resources[1]

Volodymyr Shyrokov
Ukrainian Lingua-Information Fund
National Academy of Science of Ukraine
vshirokov48@mail.ru

Abstract
The National Dictionary Depositary of Ukraine is under description. Organizing, linguistic and technological aspects of this Base are demonstrated. The main platforms of National Dictionary Depositary are described.

Key words: *national dictionary corpora, computer lexicography*

1. In the world of applied linguistics we can recently observe a real boom in the development of intelligent analysis of texts. The principal, «crucial» directions in this area are conceptography and multilingualism. Obviously, the creation of workable software products in these areas involve the use of relatively broad and complete set of lexicographic tools, to put it bluntly – a variety of digital dictionaries to present fully enough a lexicographic description of the involved parts of lingual systems.

At the same time, the creation of completely universal multilingual lexicographic means encounters some difficulties. Even now, it is not quite clear on what conceptual basis we should build such facilities. Some examples of language platforms that have initially claimed for the role of a conceptual framework (we mean the WordNet, the FrameNet or UNL – Universal Networking Language in particular) do not give cause for optimism so far. Moreover, there is currently no convincing example of a universal dictionary that provides a complete lexicographic description, at least for one language. At the same time, it is clear that a lexicographic description of the language alone is not enough to create effective means of processing and a grammatical description is required to be exploited (by itself not sufficient either). In this regard, the idea of an integral description of language and languages to bring together properties of both grammatical and lexicographic descriptions in one conceptual model and in the spirit of the principle of complementarity as set out in due time by N. Bohr are gaining ever growing popularity. A review of the ideas was presented in our work "Integral Slavonic lexicography in the linguotechnological context" published in the proceedings of the Moscow Open Workshop "Lexicographic Tools and Techniques" in the framework of the MONDILEX project that was held on October 3-4, 2008.

However, starting to work in these areas, researchers are often faced with a lack of evidence, unavailability (to the necessary extent) of the phenomenological framework for the studied languages and the lack of effective professional interaction to ensure proper interaction and speed of research, time for which is steadily declining.

It seems that a set of the above factors has been a principal motive when formulating the goals and objectives of the MONDILEX project associated with the establishment of a highly efficient environment for creative interaction between linguists-researchers.

However, when trying to create this environment, a number of "technological" problems arise, failure of which may call into question the very achievement of the project's objectives as a whole.

In this connection note that the simulation of a network of centres of Slavic lexicography what is closely related to the task of creating digital lexicographic resources to acquire the All-Slavic character have a real prospect of solutions based on the experience of the Ukrainian Linguo-Information Fund (ULIF),

[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

National Academy of Sciences of Ukraine in the field of compiling the National Dictionary Depositary of Ukraine. By now, more than 70 lexicographic works have been issued to constitute a series of «Dictionaries of Ukraine». Those dictionaries were created largely by the development of computer-aided technologies of lexicographicalization developed in the ULIF. Not only traditional paper dictionaries, but also a digital dictionary system has been created here. Many of them have a very significant volume as well as an advanced and lexicographic structure. Primarily, this relates to an integrated lexicographic system "Dictionaries of Ukraine", which is accessible on the site of the Ukrainian linguistic portal (http://www.ulif.org.ua, http://www.ulif.mon.gov.ua). In the ULIF a number of tools of so-called lexicographic virtual laboratories have been developed that allow geographically remote scientists for professional interaction over the Internet in the development of joint lexicographic projects. This technology has been tested in the development of the fundamental explanatory dictionary of Ukrainian language, which is in its paper version scheduled for publication in 20 volumes. The technology has demonstrated its high efficiency in carrying out this project. There are plans to integrate in the future the system of virtual laboratories in lexicography into a GRID system, now under development in Ukraine.

2. What is the National Dictionary Depositary of Ukraine at the moment? As noted above, it consists of a series «Dictionaries of Ukraine», published under the auspices of Ukrainian Linguo-Information Fund from 1994 to 2008 (about 70 dictionaries).

Furthermore, it includes electronic lexicographical systems as the user, as well as instrumental ones. They are listed in the table below.

OBJECTS of the National Dictionary Depositary of Ukraine

№		NUMBER
1.	Paper dictionaries of the series „Dictionaries of Ukraine"	62
2.	Digital CD ROM dictionaries	8
3.	Computer files of the lexicographical proceedings	20
4.	Lexicographical data bases and instrumental lexicographical systems	19
5	On-line lexicographical systems	5
6	Computer systems for language processing	3
7	Paper and digital dictionaries	>800

Lexicographical data bases and instrumental lexicographical systems

1. Ukrainian Linguistic Corpus (more than 70 mil. entries)

Instrumental lexicographical systems:

2. „Fundamental Academy Dictionary of Ukrainian Language"
3. „Grammar Dictionary of Ukrainian Language "
4. „ Grammar Dictionary of Russian Language "
5. „ Grammar Dictionary of English Language "
6. „Grammar Dictionary of Spanish Language "
7. „Grammar Dictionary of German Language "
8. „Grammar Dictionary of Turkish Language "
9. „Grammar Dictionary of Georgian Language "
10. „Ukrainian-Russian Dictionary "
11. „Russian-Ukrainian Dictionary"
12. „ Russian-Turkish Dictionary"
13. „ Ukrainian Spelling Dictionary "
14. "Etymological Dictionary of Ukrainian Language"
15. „ Ukrainian Synonyms"
16. „ Dictionary of Russian Language "
17. Research computer system „The Verb"
18. Research computer system „The Noun"
19. Research computer system „The Homonymy"

On-line lexicographical systems (http://ulif.org.ua/)

1. Ukrainian Linguistic Portal.
2. Digital Library.
3. „Dictionaries of Ukraine on-line".
4. „Dictionary of Russian Inflection on-line"
5. Journal of Linguistics on-line".

Computer systems for language processing

1. The automatic morphological analysis.
2. Research computer system „Linguostatistics".
3. Research computer system of the closeness analysis for Ukrainian texts

The National Dictionary Depositary of Ukraine includes also the Ukrainian language corps containing more than 70 million word entries with morphological marking and the implementation of the search for a range of grammatical and bibliographical parameters.

By now the volume and quality of the lexicographical objects of the National Dictionary Depositary allow Ukrainian Government to include it to the state register of the scientific objects which are the national property of Ukraine[1].

2. Theoretical foundation of all the systems to constitute the National Dictionary Depositary of Ukraine is our lexicographic theory. Since it was never presented in English in a regular form, its principal provisions will be set out in the next issue of the MONDILEX Proceedings. Here we are to explain just some points of this theory.

Generally, the theory of lexicographic systems provides a far-reaching generalization of the concept of the dictionary. According to this theory any dictionary structure reflects the interaction «subject-object» that takes place in any system including a lingual. These interactions (very diverse and multi-aspect, of course) result in the induction of a class of discrete, relatively stable entities caused with some phenomenological principle we call a lexicographic effect in information systems. This class is qualified as a class of elementary information units relative to a certain lexicographic effect. Then, due to expanding relations «form – content», in the system of elementary information units a lot of descriptions of these units with a certain structure take shape. Depending on a linguistic system any items, objects, relations of language can make classes of elementary information units, while the set of their descriptions generates in particular a set of lexicographic objects, i.e. dictionaries that are just designed to describe units, objects, relations of language because any linguistic phenomenon can undergo the lexicographing. The above set of descriptions of the elementary information units is identified as the dictionary articles, and the units themselves as the register words of the respective entries.

Thus, the macrostructure of any dictionary is generated by a number of dictionary entries. The microstructure is formed as an internal structure of its entries. However, because of the abovementioned «form – content» relationship, in the structure of each entry a certain opposition characteristic for this dictionary is formed to become a specific expression of this relationship, and be defined in the theory and practice of lexicography as a pair: "left-hand and right-hand sides of the entry".

The structure of any L-system embraces a set of register words W = (X), representing the elements of a class of elementary information units and serving at the same time as identifiers for the respective word entries V(X). In the structure of each entry V(X) the "left-hand side" – L (X) stands out to describe «formal» components of semantics (as a rule, it corresponds to the grammatical semantics of the register word of X), and «the right-hand side» – P (X), which provides the "substantial" component of semantics (it is usually consistent with the lexical semantics of X). Besides, the operator H: $L(X) \rightarrow P(X)$ is defined to secure the entry's integrity and the link between formal and substantive components of the description, the bearer of which is the X unit, as well as a number of other elements (some of them defined implicitly), that reflect different elements of the lexicographic description of the lexical system.

[1] Распоряжение Кабинета Министров Украины от 11.02.2004 г. № 73-р.

Noteworthy, the consistent implementation of the scheme if correlated with the information principles on which the Kolmogorov's information measure is built[1], leads to the construction of a scheme having features of quite a universal data model, a model of knowledge and a logical-linguistic calculus. Thus, the theory of lexicography provides a wide range of constructs for lexicographic modelling a wide range of linguistic phenomena.

The use of the lexicographical systems theory as a conceptual base allows us to obtain essential unification of the programming realization. In fact, three program platforms have been developed which present the ground for all main digital lexicographical systems of the National Dictionary Depositary have been built. Respectively, they are so called T-Platform, G-platform and L-platform.

Systems based on T-Platform:

„Fundamental Academical Dictionary of Ukrainian Language"
„Ukrainian-Russian Dictionary "
„Russian-Ukrainian Dictionary"
„ Russian-Turkish Dictionary"
„ Ukrainian Synonims"
„ Russian Synonims"
„ Dictionary of Russian Language "
„ Dictionary of Turkish Language "
"Etymological Dictionary of Ukrainian Language"

Systems based on the G-platform:

„Grammar Dictionary of Ukrainian Language"
„Grammar Dictionary of Russian Language"
„Grammar Dictionary of English Language"
„Grammar Dictionary of Spanish Language"
„Grammar Dictionary of German Language"
„Grammar Dictionary of Turkish Language"
„Grammar Dictionary of Georgian Language"

On the base of the L-platform such systems have been developed: Ukrainian Linguistic Corpus and such applied systems as Digital Archive of the documents of Presidium of National Academy of sciences of Ukraine, Ukrainian Biography Archive, The legislation of Ukraine; under development are Digital Encyclopaedia "Taras Shevchenko" and some others lingua-information systems.

The feature of the technical realization of the platforms above is their realization by the architecture of the virtual lexicographical laboratories. This allows one to realise mutual work in fulfilling of the large lexicographical projects by many linguists from different institutions and countries. In more details T-platform is expound in the O. Rabuletz article in this Proceedings. The example of the use of T-platform is expound by I. Ostapova (the paper "Digital Etymology" of this Proceedings). The examples for G-platform are presented in the articles of T. Lyubchenko and I. Shevchenko of this Proceedings. We are convinced that the methodology of the platforms above may be adapted to the tasks of MONDILEX and in some perspective will make a basis for Linguistic GRID.

[1] Колмогоров А.Н. Три подхода к определению понятия "количество информации". // "Теория информации и теория алгоритмов". – М.: Наука, 1987. – С. 213-223.

Ukrainian Academic Grid: State and Prospects

Eugene Martynov
Bogolyubov Institute for Theoretical Physics of NAS of Ukraine
Metrologichna str. 14b, Kyiv, UA-03680, Ukraine
martynov@bitp.kiev.ua

Abstract
Ukrainian Academic Grid (UAG) is presented as a powerful computing resource for fundamental and applied scientific research which are carried out at the NAS of Ukraine (NASU). Information on activity, structure, computational power of UAG follows the short historical outlook. Examples of a cooperation of NASU institutes with international Grid projects and organizations are presented. Prospects of UAG activity in developing various Grid applications within various scientific areas including humanitarian ones are shown.
Keywords: *computing, grid, middleware, grid-technology, grid-infrastructure, Ukrainian academic grid*.

Introduction

An area of information-computing technologies has been fundamentally modified afterwards the basic idea of the spatially distributed computing appeared in the eighties of the last century. Grid technologies occupied an important unless the most quickly progressing place in this area. Their principles have been logically and adequately formulated by I. Foster and C. Kesselman [1]. It is the authors's opinion that "A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities". In other words the Grid looks to users like one powerful computer with unlimited resources (processors, main storage, footprint etc.).

The impetuous development of Grid technologies in the world is caused by the increasing complexity of computational problems in the various human activities, on the one hand, and by the fast progress of computer material resources and the perfect element resources of modern computers, on the other one. An increase of the computer capacity and the Internet links getting cheaper rate play almost a key role in this process inasmuch as Grid uses the Internet as data communications medium.

From the outset Grid technologies have been mainly applied for high energy physics. Now they have already penetrated in the various fundamental and applied sciences (from physics and astrophysics to Earth sciences and biomedical ones) and are steadily advancing in industry, economics and social life. It is timely to refer to the analysis done by the INSIGHT Research Corporation in 2006 [2]. According to published thesis "Grid Computing: A Vertical Market Perspective 2006-2011" the investment to the Grid technologies will increase from 1.84 billion in 2006 to 24.52 billion in 2011 (see Diagram 1).

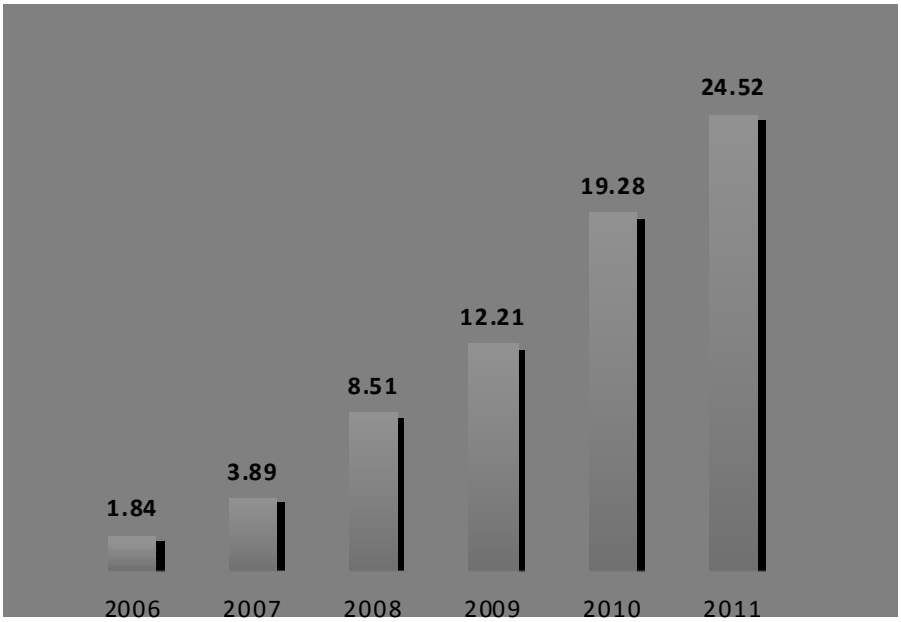


Diagram 1. Funds for developing Grid in the world (in billion dollars)

Developing the Grid technologies, building the national and international Grid infrastructures is naturally fitting in the objective globalization processes in economics, sciences and culture. The international associations and projects are playing now an increasingly appreciable role in sciences and economics. Moreover, it is a challenge for any country to realize independently some research projects such as program of studying a microcosm using the most powerful collider of elementary particles and nuclei - LHC (Large Hadron Collider) at CERN. These projects require quite serious data-processing support which Grid can satisfy.

Nowadays there is none of highly developed countries which reject an idea to build up a proper Grid network. We can state with assurance a country that has not Grid infrastructure and is not drawn into the world Grid community could not even pretend to be a developed one. Thus, a development and application of Grid technologies for the daily social life becomes a strategic trend for each state which is anxious about scientific, economic and social progress.

Grid technologies and grid infrastructure

From the practical view point the Grid is a service for sharing spatially distributed computer systems into one vast computing resource to provide more efficient use of computing resources available and to solve the problems which need the serious computing power. Functionally the grid infrastructure has to ensure:

- effective intercommunication of heterogeneous computers or computer systems,
- HPC - high-performance computing. Grid allows to share resources for performing great computations (for example, data processing from LHC at CERN), which can not be done using an individual computer system,
- processing of huge data array, which can be stored in different remote memory. These calculations must usually occur with extra load conditions in computing and communication resources,
- support and operational compatibility of different virtual organizations (VOs). Grid has to support a cooperative activity of the real Vos members and to provide, as required, the intercommunication of different Vos through the middleware.

To realize the functions listed above Grid as a data processing infrastructure must have three principal components:

1) clusters or single computers (proper computing and operating resources),
2) Internet high-speed channels,
3) middleware.

The necessity to use the Grid technologies in the National Academy of Sciences of Ukraine (NASU) mainly evolves from recent research areas, appearance of new computational problems with unexampled high demands to computation capacity and data processing capacity, computation speed as well. These problems arise, for example, in high energy physics and astrophysics, biophysics and biology, science of materials, Earth sciences and many others, both fundamental and applied studies. As it was proved Grid technologies can be installed (with minimal financial costs) into a daily practice of institute regardless of its geographical location in such a way that any researcher gains the access to the computational resources even though his institute does not have such resources.

Ukrainian Academic Grid

The first Grid site appeared in Ukraine in 2004. It was created by the group of physicists from National Scientific Center "Kharkov Institute of Physics and Technology". The computer cluster was built in the framework of collaboration with Joint Institute of Nuclear Researches (JINR, Dubna, Russia) and participation in CMS (Compact Muon Solenoid) experiment at LHC which is one of the large experiments planned to run at CERN (Geneva, Switzerland). We would like to emphasize slightly later but very soon the computer clusters have been built and put to operation at the Glushkov Institute of Cybernetics, NASU (Kiev), Institute of Condensed Systems, NASU (Lviv).

During several last years the scientists of the Bogolyubov Institute for Theoretical Physics of NASU were discussing the problems of constructing the powerful computing capacities necessary for an analysis of experimental data in high energy physics at the Ukrainian academic Institutes and Universities. Eventually at the end of 2005 the consensus has been reached that the Ukrainian community will try to contribute at CERN not only to physics but to the computing support of the future LHC experiments. Such a collaboration will provide, first of all, a direct access to the new data necessary for the theorists to develop and verify new physical ideas and, secondly, will help to the Ukrainian physicists "to be in the centre of events" due to the Grid possibilities maintaining the close professional contacts with CERN which is a world leading research organization.

With an incentive support from director of the institute academician A. Zagorodny the physicists of BITP, professor G. Zinovjev, leading researcher E. Martynov, senior researchers S. Svistunov and V. Shadura, have worked out the first program of implementing and developing the Grid activities in BITP and the NASU Institutes. It has been underlined already in this first document that the Grid technology application is an extremely advanced method not only to develop high energy physics, but to solve various fundamental and practical high computational problems in many branches of fundamental research. It opens the new horizons in the research process as well as activates the international cooperation in different human activities. That was a strong motivation for the initiators to be targeted on creating the Grid infrastructure which could be used by the NASU scientists and specialists from other institutions operating with high computational problems. Building the first Grid site at BITP has been announced and started to involve many people in high and enthusiastic activity in different scientific institutions.

At that time there was not high speed channel in the Institute and the prototype grid site of 2 servers has been created in Computer Center of Taras Shevchenko Kiev National University (CC KNU) by very professional activity of Dr. S. Svistunov and Dr. A. Sudakov who is a senior researcher of CC KNU. Due to a tight collaboration of BITP and ALICE Collaboration (A Large Ion Collider Experiment) [3] the Grid site has been certified by AliEn-grid (Alice-Enviroment-grid) [4] (the Grid organization for providing the computing resource for the ALICE experiment at CERN).

In June 2005 the session of Coordinating Board for Informatics of NASU was held at BITP. The project to create the Grid computer cluster at BITP has been comprehensively discussed and approved. The practical work was trigged up with quite clear prospects.

At the end of 2005 the computer cluster of 10 nods (double processors) has been built and the Grid middleware has been installed. A. Alkin, V. Savchenko, M. Zynovyev (at that time students of KNU and National Technical University (NTUU KPI) "Kiev Polytechnic Institute") were heavily involved in the practical work and kept well trained under the guidance of S.Ya. Svistunov. In the direct access mode and dialog with experts from CERN the Grid system has been configured and specific settings for AliEn-grid

have been installed. Several events which took place on April 2006 have provided the progress in the Grid development at BITP and NASU.

On April 25 the National Academy of Sciences of Ukraine has been officially affiliated to the WLCG (Worldwide LHC Computing Grid [5]) organization which has to coordinate the computer (Grid) support for LHC experiments. In July 2006 the Prime Vice-president of the NASU academician A. Shpak who represents the National Academy in WLCG has signed Memorandum of Understanding between NASU and WLCG.

By April 2006 Conception of the Grid infrastructure development in Ukraine has been worked out and approved by the NASU Presidium. At the end of 2006 new Grid clusters have been built in five institutes of NASU and in December the first prototype Grid system of 2 clusters (BITP and KNU) has been put into operation to process Grid network using NorduGrid middleware. Basing on this gained experience it became possible already in April 2007 to combine clusters of six NASU Institutes into the first Ukrainian grid segment.

Ukrainian Academic Grid (UAG) is a Grid infrastructure to share the computer resources of the NASU institutes and universities as well. The principal task of the UAG is to develop the distributed computings and grid technologies to advance computationally intensive fundamental and applied studies NASU. Besides, UAG has to ensure a participation of the Ukrainian scientists in various topical international Grid projects.

At present the UAG shares resources (more than 2000 CPU and 200 TB of disk memory) of the following organizations (see Fig. 1):

Fig. 1. Ukrainian Academic Grid

The idea to enlarge the Grid infrastructure user base via the so-called access Grid platforms, i.e. the Grid clusters with "minimal configuration", has been conceived realizing the NASU Program. Such a cluster holds the control server with installed middleware and two computer servers. Working permanently in the conditions of limited funding a system of the access grid platforms proposed and developed in the NASU grid infrastructure allows the specialists of Institutes without operable cluster to use the academic grid full profile. A local resource broker using available network resources distributes the tasks among the Grid infrastructure clusters as the computing requests (small problems require a hands-on operation and large-scale ones are directed to more powerful clusters).

With the funds available any mini cluster can be easily extended to the full scale cluster. This policy makes it possible to train system administrators for the work with more power clusters. The high-speed and reliable access channel to Internet networking is one of the necessary conditions at the Grid infrastructure building. Ukrainian Academic and Research Network (UARNET) company has built the infrastructure capable to unify the academic institutions with fiber-optic channels (capacity 100 Mb/s) by connecting to the long-distance backbone Lviv-Kiev-Kharkiv (capacity 2,5 Gb/s) with the next output to Slovakia and Poland. It is necessary to emphasize the traffic between the academic Grid clusters is free.

In the Institute for Theoretical Physics which is responsible for the academic Grid program running the UAG web site (http://uag.bitp.grid.ua) has been designed whereat the information on Grid technologies and Grid development over the world, news and publications can be found. The site is permanently refreshed and updated.

Today UAG is operating under the ARC NorduGrid middleware. A choice of this middleware was done because this package was supplied with good documentation, had simplier installation and entry-level configuration though did not possess some necessary properties to exploit more Grid possibilities (for example, automatic task allocation). Now an installation and testing middleware gLite at the sites of NASU are in the progress.

First of all, we consider the NASU Grid infrastructure both as a supporting equipment for solving the fundamental and applied problems and as a ground wherein the Grid technology methods are proved in accomplishment of various tasks. These methods could and should be used then in the national Grid infrastructure whose function is far wider than the research data computing. Nevertheless, below it is an appropriate list of the scientific areas where new UAG facilities have been already used.

High energy physics.

LHC (CERN) experimental data processing, their analysis and comparison to the theoretical results and phenomenological models aiming the full scale participation of the Ukrainian institutes in the ALICE experiments (BITP, KNU, IC, KIPT, ISMA, NTU KPI) and CMS ones (KIPT).

Astrophysics and astronomy.

- Dynamical computing of an evolution of the star concentration in the Galaxy external field. The hydrodynamic modeling of collision and fragmentation of the molecular clouds. Analysis of N-body algorithm and parallel computing on the GRAPE clusters. Cooperation with AstroGrid-D (MAO).
- Theoretical analysis and the observation processing of primary, roentgen and gamma radiation data which are obtained from the satellite telescopes INTEGRAL, SWIFT and others (BITP, KNU, MAO).
- Creation and formation of VIRGO – VIRtual Gamma and Roentgen Observatory (BITP, KNU).
- Development of nuclei activity models of Galaxy and star concentrations. Testing the dark matter and dark energy models. Collaboration with Lausanne and Geneva universities (BITP, GAO).

Biophysics and biology.

Computing of thermodynamic characteristics, infrared and electron spectra of sputter DNA fragments. Study of bionanohybrid system structures composed by DNA and RNA of different sequence (ILTPE, IC).

Molecular dynamic computing of Fts-Z-protein systems with the low-molecular associations (ICBGE).

Computer simulation of the spatial structure and molecular dynamics of cytokine-tyrosine-RNA synthetase (IMBG).

Nanotechnologies.

Computing of nanostructure oxides which seem to be perspective high-temperature superconductors, as well as physical characteristics of the DNA fragment with transition metal ions which could be good nano-conductors.

Computing of structures and interaction energy of bio-nano-hybrids on basis of the single-shell carbon nano-tubes with the various bio-objects (ILTPE, IMP, IC).

Environment monitoring.

- Weather forecast parameters on the Ukrainian terrain based on the computer simulation and satellite data. Estimate of biodiversity as ecologic parameter on Ukrainian terrain (ISR, IC).
- Development of GEO-UA information infrastructure (ISR).

The future development of Grid technologies in NASU should be focused on their application in the specific research. There are three directions for creating and using the Grid technologies which have to be applied in the daily research work in the nearest future:

work out new packages for computing on one multiprocessor cluster and on several various distributed clusters as well;

adapt the relevant middleware for parallel processing;

use the developed and free distributed license middleware which has been already tested in domestic and foreign institutions.

It is a well-known fact that there are a lot of highly-qualified specialists in Ukraine and in NASU, in particular, they are able to resolve the problems of this type. There is a need to manage and provide them with the financial and material resources as well as to set the cooperation between the computer specialists and physicists, chemists, biologists, engineers and others who are interested in the Grid applications and to make of fruitful.

The strategic emphasis in development of UAG should be primarily placed on the creation of such a system which is based on the power supercomputer centres allowing the distributed parallel computing with using ten or even hundred processors. Then any academic institution or research group in Ukraine could have a required time for computing and the facilities of these centers could be optimally used with the Grid technology advantages. The world experience, in particular, demonstrates a prospect of such an approach. At the same time there is a need to increase a quantity of the Grid sites in the NASU institutes in order to more and more scientists could have an access to the large computing resources.

This activity results from the fact that the largest NASU institutes which carry out the highly computational investigations already have or will have soon very powerful clusters. Certainly, the computer systems are very quickly progressing, therefore it is desirable to update the resources in proper time. At this stage with the operational Grid network it is necessary mainly to create the access platforms in the small institutions and to supply them with the high speed Inetrnet.

It is essential to realize that the Grid infrastructure with a lot of Grid sites can not be merely global computational resources with minimal management. Now Ukrainian Academic Grid is composed of 22 powerful Grid clusters and the access Grid network platforms. In the world practice examples the Grid infrastructure organizations abound within the big projects which consolidate the institutes and laboratories of many countries as well as within the national projects for the countries working separately.

In accordance with, for example, the WLCG scheme the infrastructure of UAG can be build as many-level system:

1. Basic Coordinating Centre which governs UAG mainly through the regional centres.

2. Regional Operating Centres which coordinate the activity of Grid sites (a Grid site is a Grid cluster or an access platform) in regions.

3. Separate grid sites (institutes) or minimal Grid network access platforms which belong, as a rule, to any virtual organization (VO). VO temporarily joins institutes (not necessarily from the same region) of common scientific interests to solve a problem.

The Glushkov Institute of Cybernetics (IC) with it's the most powerful computer system has to occupy the special place and role in the Grid infrastructure of NASU. It is necessary to organize the collaboration between the IC and UAG in such a way that the more resource-intensive tasks should be sent to this computer system. Therefore, first of all, its stable and reliable work as a pan-academic Grid resource should be provided and, secondly, the licensed application program packages for supporting research calculations of users from other institutes have to be installed on the IC cluster.

Each VO should have its own centre – Resource Centre of VO (RCVO). The institute which takes the front line in the appropriate research and disposes of the adequate computer resources can organize the collaboration with the other VO participants and should discharge the RCVO obligations.

Thus, in respect to an organization and management of the total computing resources UAG has three levels. It is ensured that due to the large performance of Grid infrastructure optimized, its stability and reliability are controlled, and the program of some of its elements is governed. Basic operation-resource centre of NASU consists of teams which provide such a performance.

Cooperation with Ukrainian Institutes behind NASU

At the very beginning the NASU researchers are working in full cooperation with the Universities and Institutes behind the NASU structure. Today Higher Schools of the Ministry of Education and Sciences have the good possibilities to teach the IT-specialists, some of them possess the considerable computational resources. They can and must make an essential contribution to the Grid technology development in Ukraine and creation of national wide branched Grid infrastructure.

It should be emphasized that the active work is now underway on the Grid technologies application in the medical institutions. The program of collaboration between the NASU and Academy of Medical Sciences of Ukraine and some big medical centers is successfully working. The NASU is ready to aid and to grant the computational resources for using the Grid technologies in medical practice in the framework of pilot projects.

Scientific and technological project "Creation of National Grid Infrastructure for Research Support" is fulfilled today in Ukraine along with grid project of NASU (BITP is a basic organization with 20 participants). This project was approved by the Ministry of Education and Sciences (MES) to be accomplished in 2007-2008 as a part of the State Program "Information and Communication Technologies in Education and Science". The program provides the formation of the national Grid infrastructure, the creation of Certificate Centre, the maintenance of Ukrainian international data center. NTU KPI is a leading organization in this project and 7 universities and institutions of MES are the contracting parties. It should be mentioned that NASU becomes the contracting party as well.

Taking into account that both projects have similar purposes the MES and NASU started collaboration. BITP and NTU KPI made an agreement to create the Ukrainian Grid Association which, in accordance to the EGI recommendations, was termed as Ukrainian National Grid Initiative (UNGI). The structure of UNGI was strictly outlined and documented. The immediate and advanced plans which should be realized after accepting and approval of the State Program were discussed and fixed. The formal procedure of association registration is very close to be completed now.

Bogolyubov Institute for Theoretical Physics of NASU propounded the initiative to develop the State scientific-technical Program of Grid technology development in Ukraine. Such a Program was oriented under name of NASU and presented to the Cabinet of Ministers of Ukraine. It is assumed that Program will be financed mainly from state budget, and that it will be performed during five years starting in 2009.

The modern state of Grid technologies in Ukraine is represented in the Program in the following way. "The main problem is that for now there exists in Ukraine the critical need of using modern information-communication technologies (first of all, Grid technologies) for processing the super large arrays in the interests of science, industry, and social sphere, but the necessary national Grid structure is not available. Presence of such infrastructure separate elements does not meet the modern level of Grid technologies development, and does not assist to solving the whole line of actual scientific, scientific-technical and some other problems, holds back European integration processes in these fields".

In Program the reasons of such a state are analyzed, necessity of building National Grid is proved, fields of possible use of Grid technologies are enumerated, economical and social gains which can be received as the result of Program execution, are accentuated.

"Goal of the program is national Grid infrastructure creation, and wide implementation of Grid technologies into all the spheres of social-economical life in Ukraine".

Program priority tasks are:

1 creation of the system, taking into account the information safety providing, integration of necessary elements of one national Grid: computing, communication and program resources,

2 grid technologies adaptation and application in Ukrainian multi-processor computational systems,

3 grid technologies implementation and application in scientific research, integration of Ukrainian scientific establishments into the world scientific space, drawing of Ukrainian scientists to participation in modern unique experiments and computer processing of their results, to participation in virtual scientific forums;

4. implementation of new methods of population medical service (creation of distributed diagnostic data bases, consultation with the use of telecommunication means, including large scale computer data analysis);

5 providing with efficient, real-time processing of the results of geophysical, meteorological, and space observations;

6 creating conditions for grid technologies implementation in economics, industry, financial activity, and social sphere;

7 creating the system of training specialists for the work with grid technologies.

Computer resources, different at technical realization and at the type of construction with the purpose of providing user with the aggregate computer resources, will be united into the single system. User will get service from grid infrastructure as from the system in whole, independently on where and what computer stores or processes his information, what transmission lines work at this. Creating such systems will radically increase the efficiency of using the aggregate computer resources of the country, will give it the main new possibilities to solve the complex scientific-technical and practical tasks and problems.

International cooperation of UAG

Grid may be considered as a new reinforced instrument for scientific and technological international cooperation. Grid becomes one of the principal factors and locomotives of the globalization process. Science has always an international nature but at the end of previous century because of the fight with the background of economic globalization the proper attention to developing the cooperation principles in the science management was not given in necessary extent. Nevertheless, due to Internet and new scientific projects (for example, Space exploration, the largest Colliders in CERN and USA, European project of thermonuclear reactor ITER etc.) the tasks of world science integration have been brought to the forefront. New international project called the World Grid could be realized by creating the national and big international Grid projects (WLCG, EGEE, GLORIAD, TERAGRID and others). Realizing a stable character of unification tendency and availability of this process in the country the organizers of UAG make the special efforts to integrate and consolidate the Ukrainian Grid into the international Grid community.

Today the Ukrainian experts are ready to learn the great experience their foreign colleagues, however, they have a lot of achievements as well which could be very practical for the international community of the Grid users. Below we list new trends and activities which as we believe are the quite promising for future cooperation and will provide both sides with a steady progress of the Grid technologies.

Ukraine was registered as a non-contracting participant in the EGEE project in 2007. It was planned that the Ukrainian specialists should work out the subject of Grid applications (e.g. high energy physics, astrophysics, life science, earth science) and focus their efforts on education and knowledge propagation of Grid and distribution of the Grid technologies into medicine and industry.

In May 2007 Ukraine signed the Memorandum of Understanding about the participation in EGI. The spectrum of tasks which are of interest for the Ukrainian experts corresponds quite well to the activity which was declared in EGEE.

Since April 2006 NASU is a member of the WLCG collaboration. Several academic institutes put their computational resources for common work in WLCG. The researchers of BITP and KIPT accomplish the theoretical and phenomenological tasks as well as the technical preparations for the future work in the LHC experiments.

The researchers of the Institute of Space Researches of NASU and Space Agency of Ukraine (SAU) in cooperation with their Chinese colleagues are fruitfully developing and use the Grid technologies in the satellite monitoring of the Earth surface and water.

The agreement about the joint investigations and cluster computing with GRAPE-cards in the framework of German AstroGrid-D project has been signed by MAO. In the Institutes of biological investigations the negotiations with colleagues from the Western countries about the common projects within the specific VOs are carrying on.

There are no doubts that the international relations of UAG, its collaboration in the Grid projects with experts of many countries will be considerably extended and intensified.

Prospects of UAG and UNGI

Creating the UAG basic infrastructure we gained the practical experience both in development of clusters and in management of cooperation with the various NASU institutes in the Grid activity. Now we understand better what should be done to guarantee an appreciable progress of computing constituent not only in NASU but in Ukraine, too. We believe that Grid will be used not only for various scientific and technical computations, but also in humanitarian sciences. Grid can help to deal with huge data bases, electronic libraries, quantitative and qualitative analysis of various texts and to solve many other tasks and problems (see, for example, the project TextGrid, http://www.textgrid.de) .

Despite all difficulties and problems in developing of grid technologies in NASU the background of the widest application of grid technologies in Ukraine has been provided. There is a good reason to believe that grid exists and operates in Ukraine, the collaboration with international grid community is intensified and Ukrainian National Grid will be built with joint efforts and occupy a fitting place in the world grid infrastructure.

References

1. Ian Foster, Carl Kesselman. The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann Publishers. ISBN 1-55860-475-8, 1999.

2. http://www.insight-corp.com/pr/06_22_06.asp
3. http://aliceinfo.cern.ch/Public/Welcome.html
4. http://pcalimonitor.cern.ch/map.jsp
5. http://lcg.web.cern.ch/LCG/
6. http://www.eu-egee.org/
7. http://web.eu-egi.eu/

Systems Engineering Principles of Virtual Linguistic Laboratories[1]

Alexander Rabulets
Ukrainian Lingua-Information Fund
National Academy of Science of Ukraine
ferry@bigmir.net

Abstract
In the article the concept of virtual lexicographic system (VLS) have been considered. Actually, it is to provide linguists who work in different institutions, different towns and even in different countries, the possibility of access to the computer systems, on which people could collectively carry out large linguistic projects.

Key words: *virtual lexicographic system (VLS), Virtual Lexicographic Laboratory (VLL), Grid.*

The concept of virtual lexicographic system (VLS) was first introduced by V. A. Shyrokov in the monograph "Information Theory of Lexicographic Systems".

Actually, it is to provide linguists who work in different institutions, different towns and even in different countries, the possibility of access to the computer systems, on which people could collectively carry out large linguistic projects.

Let's consider how this problem can be solved.

We will make it on an example of the project of creating the explanatory "Dictionary of the Ukrainian Language" (DUL) in 20 volumes. Despite the fact that the project to some extent can be considered completed – the first volume of the dictionary has been passed for publication – but this project should be continued. In fact, new tasks constantly appear, new linguistic facts and knowledge, which should be transferred to a dictionary form, are accumulated.

After all, it is interesting to have a universal lexicographic system, which develops in real time, demonstrating the development of the language itself and our views on it.

Moreover, the modern trends of developing the computer communications and, above all, the Internet lead to activization and dialogization of the lexicographic processes in the network. (Particularly prof. Martynov reported this in his report on the academic GRID).

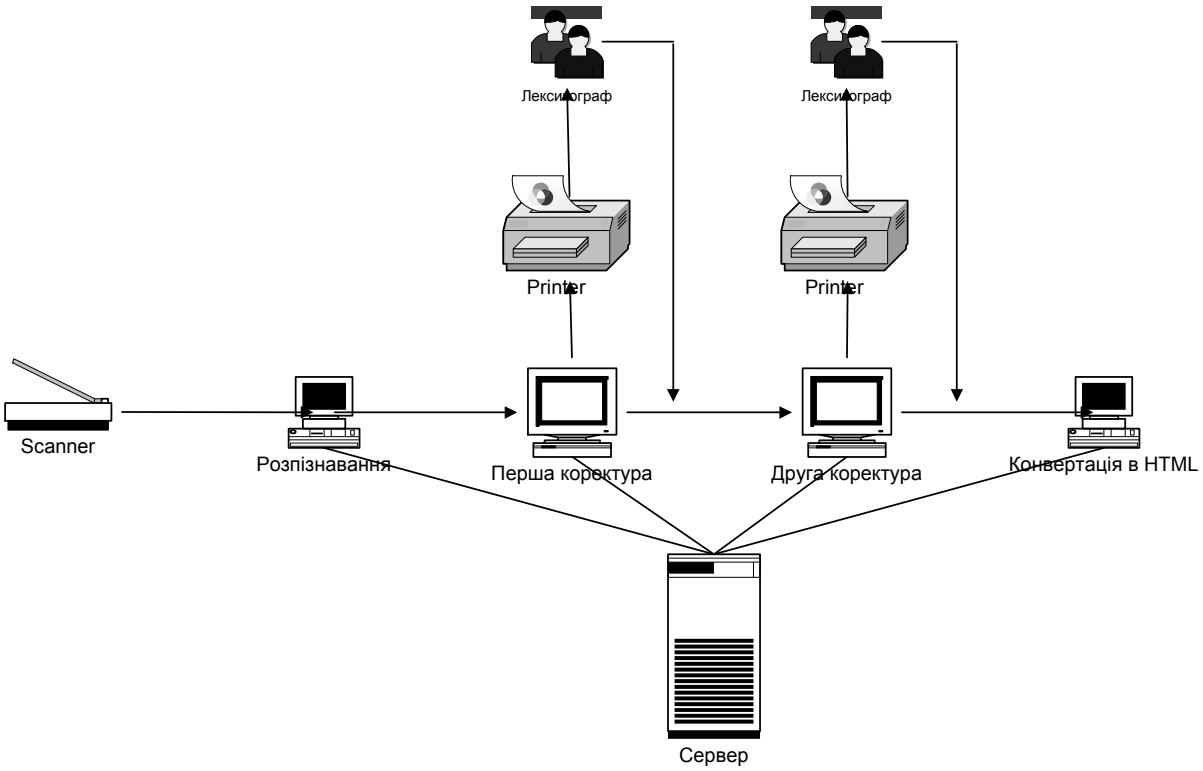
Finally, there is a task of creating a united linguistic space of Ukraine. When formulating the principles of this task, we assumed that for this work we should involve all linguistic units from the universities in Ukraine and provide effective collaboration of specialists who perform much work together.

To do this, let us demonstrate the functionality of the instrumental system "Dictionary of the Ukrainian Language", created in the Ukrainian Lingua-Information Fund (ULIF) as a basis for creating DUL in 20 volumes.

The 11-volume explanatory "Dictionary of the Ukrainian language" issued over the years 1970 – 1980 became its prototype. It should be noted that the process of creating a dictionary had been lasting for 40 years, as participants of that project say. We were tasked to form the main corpus of the dictionary in 20 volumes within five years.

This was done as follows:

[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 figure shows the process chain of creating the computer lexicographic database of the DUL. First, the text of the 11-volume dictionary was scanned. Then it was recognized using the optical character recognition program, transferred to the computer text format and printed on paper. After that the text was corrected by the linguists from ULIF, the Institute of Linguistics and the Institute of the Ukrainian Language.

After that the so-called Dictionary parsing was carried out. It means, that the formal attributes of each of the DUL structural elements were identified according to the DUL formal structure (it was described by V.A.Shyrokov in the monograph "Information Theory of Lexicographic Systems"); and the computer database structure that meets the DUL formal structure was developed.

After this a program procedure was written, to analyze the structure formal features from the electronic text of the DUL, spread it automatically on the fields of the developed lexicographic database. Thus, the initial database of the DUL was created just in two weeks.

It's not an easy task. You can understand it of the fact that we are aware of some groups, which tried to make the DUL database, worked on it more than 10 years, but there was no result of their work.

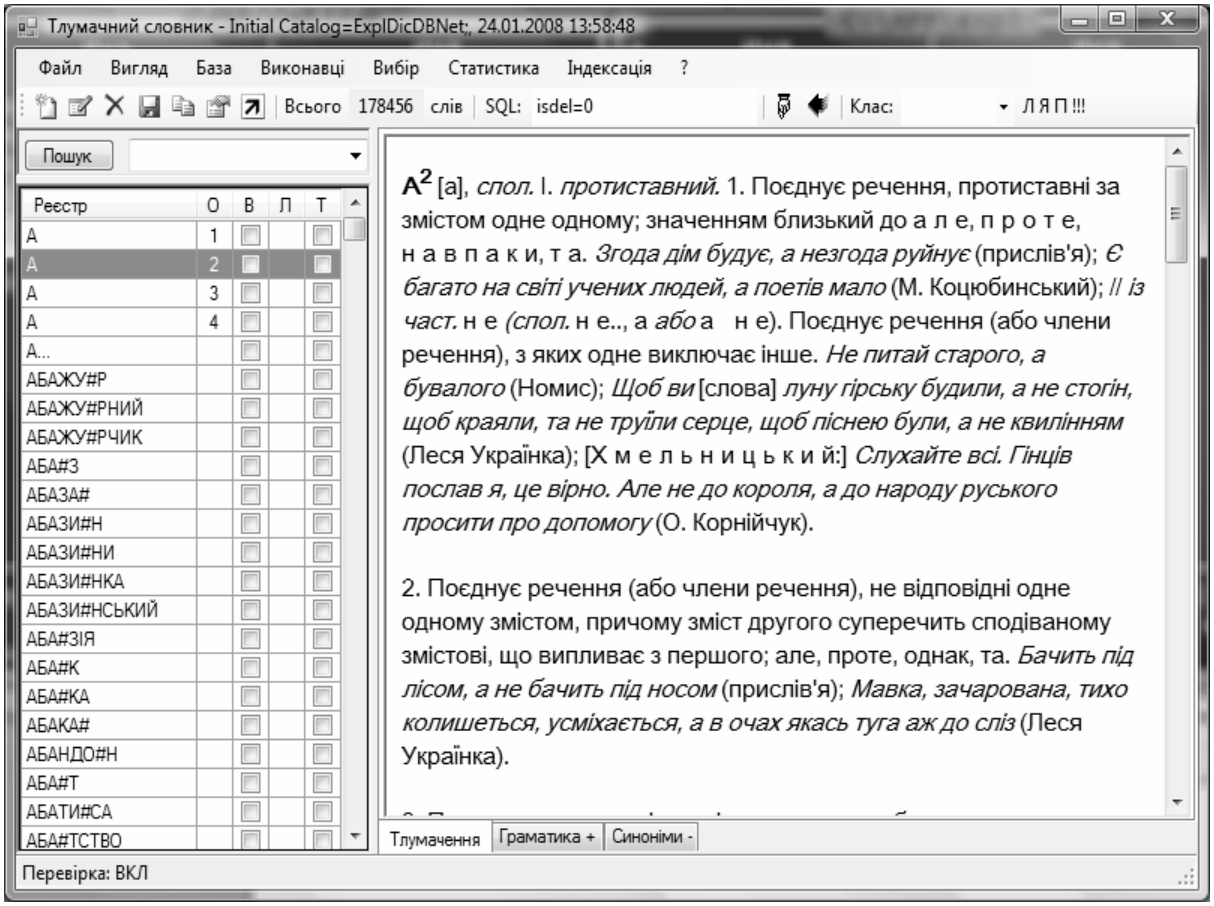
An instrumental complex was developed on the basis of this DUL database. It is important, that till 1 January 2007 technologists worked directly with the system and lexicographers gave them material on paper.

Since 2007 our lexicographers worked directly with this system and they have prepared and given the first volume of the DUL to the publishing house.

The main functions of the DUL are described below:

- Users authentication and authorization;
- Filling new dictionary entries into the lexicographic database of the information system;
- Deleting dictionary entries from the database;
- Editing dictionary entries in the lexicographic database;
- Providing modification of dictionary entry structure of certain lexicographic system and creating new lexicographic systems;
- Dynamic creation of dictionary entries in the lexicographic databases in print format;
- Fulfilling different types of sorting and creating subsystems;
- Data analysis;
- Adding new users;
- Deleting users;
- Managing users' access rights.

Let’s return to the main screen.



A dictionary register with such columns is displayed on the left side of the main window:

- Register – register word of the dictionary entry;
- O – homonym number;
- B – attribute, that indicates whether the dictionary entry was processed by the publishing editor;
- Л – attribute, that indicates whether the dictionary entry was added or changed;
- T – attribute, that indicates whether the dictionary entry was processed by the technologist.
- The dictionary entry text is displayed on the right side of the main window. The buttons on the toolbar have such destinations:
- “Add” – adding a new dictionary entry (it is duplicated in menu “Base > Add");
- “Edit” – editing the current dictionary entry (it is duplicated in menu “Base > Edit”);
- “Delete” – deleting the current dictionary entry (it is duplicated in menu “Base > Delete”);
- “Write to file” – saving selected dictionary entries or a range of them in HTML format (it is duplicated in menu “Base > Write HTML”);
- “Copy” – creating a copy of the selected dictionary entry (it is duplicated in menu “Base > Copy”);
- “… words at all” – the number of words in the dictionary registry or in the selected subset;
- “SQL:” – filtering the dictionary registry on arbitrary condition in SQL format, which is entered in the next line;
- “Go to” – transition to the article on the word that is selected in the text of the current dictionary entry;
- “Back” – back to the previous article;
- "GOTCHA!!!" – marking the selected dictionary entry as a problem (or unmarking it).

Also such functions are available through the program menu and other controls:

- Saving and editing of phraseological index;
- Setting parameters of viewing dictionary entries: font, stress mark, using style files, need for illustrations etc.;

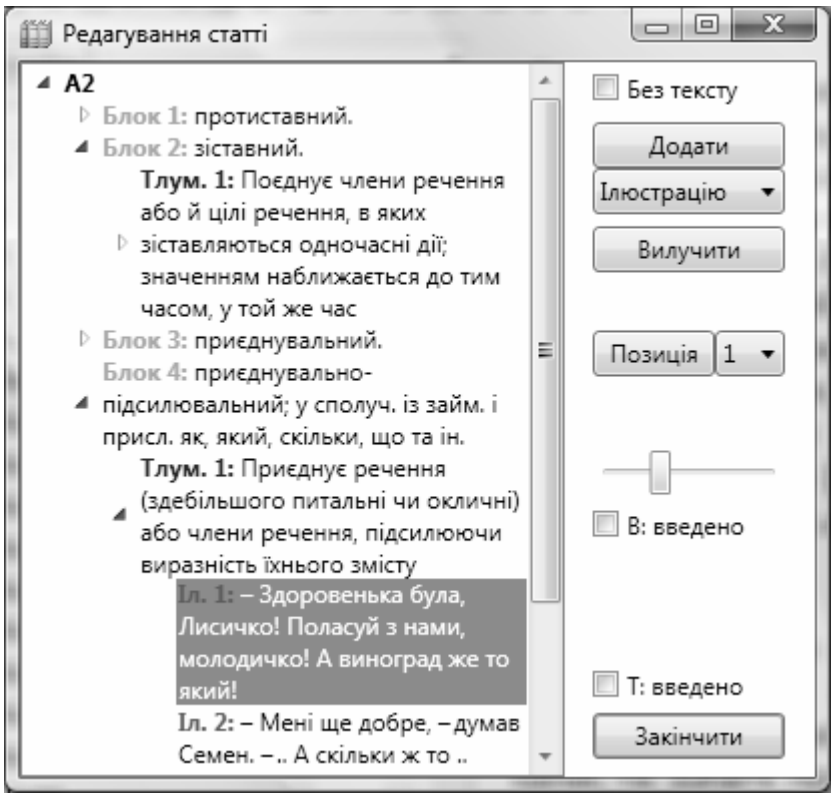
- Selecting register ranges by the performers – lexicographers, research editors etc.;

- Register filtering by various criteria: available or missing dictionary entries in the 11-volume dictionary; added or changed entries; by part of speech; missing interpretations or illustrations; stylistic remarks; typical interpretation formulas – quasisemantics formulas etc. ;

- Obtaining statistics: a number of all, added or changed entries, interpretations or illustrations in any register range, a number of characters in the entries, diagrams on the number of stylistic remarks and typical interpretation formulas;

- Indexing dictionary entries in the selected range;

The window of dictionary entry editing looks as follows:



Here on the left side there is a structure of dictionary entry as a hierarchical tree, where:

- "Block" – interpretation block;
- "Int." – separate lexical meaning;
- "Sh/mean." – shade of meaning;
- "Phrase" – phrase (idiom, terminological phrase, word equivalent or stable expression);
- "Mean." – phrase meaning;
- "*)int." – parts of interpretations or shades of meaning;
- "Deriv." – idiom derivatives;
- "Ill." – text illustration.

The instrumental system gives to users wide opportunities for replenishment, editing, modifying, and correcting lexicographic database. It provides many functions that help prevent errors.

Also there are prompts in the system. For example, when entering a new word into the register automatically, depending on its part of speech, the system offers lexicographer to add derivative variants to the register according to certain lexicographic clichés like:

Абстр. ім. до

Вищ. ст. до

Властивість за знач

Властивість і стан за знач.

Властивість і якість за знач.

Дієпр. акт. до
Дієпр. пас. до
Дія за знач.
Дія і стан за знач.
Док. до
Друга частина складних слів, що відповідає слову
Жін. до
Збільш. До
Збірн. До
Зменш. До
Зменш.-пестл. До
Найвищ. ст. до
Однокр. До
Пас. До
Перша частина складних слів, що відповідає словам
Перша частина складних слів, що відповідає слову
Перша частина складних слів, що відповідає:
Пестл. До
Підсил. До
Прикм. До
Присл. До
Стан за знач.
Стан і властивість за знач.
Те саме, що
Уживається як пред. за знач.
Уживається як присудок за знач.
Числівник порядковий, відповідний до кількісного числівника
Якість за знач.
Якість і властивість за знач.

A number of other instrumental lexicographic systems have been built in the ULIF on the DUL technology. The list is below:

1. Russian Language Explanatory Dictionary
2. Turkish Language Explanatory Dictionary
3. Ukrainian Language Grammar Dictionary
4. Russian Language Grammar Dictionary
5. Turkish Language Grammar Dictionary
6. German Language Grammar Dictionary
7. English Language Grammar Dictionary
8. French Language Grammar Dictionary
9. Spanish Language Grammar Dictionary
10. Ukrainian Language Etymological Dictionary
11. Ukrainian Language Dictionary of Synonyms
12. Russian Language Dictionary of Synonyms.

Together with colleagues from the Warsaw Institute of Slavic Studies we have begun the work on creating the Polish Language Grammar Dictionary.

Some software technology was developed in ULIF, which provides access to the instrumental lexicographic systems – in principle, to all mentioned above. This access is provided via Internet with a full range of functions that were mentioned in the local variant.

This is a Virtual Lexicographic Laboratory, because it gives to user an impression that he is working with a local system.

The first, pilot VLL was tested in the ULIF on instrumental system of the 20-volume explanatory Dictionary. Now a lot of our staff lexicographers work with it in the virtual mode. A pilot version of the virtual laboratory was put into operation between the Ukrainian Lingua-Information Fund and the Center for Applied and Cognitive Linguistics at Taurida National V.I.Vernadsky University in 2007. That is, it connects Kiev and Simferopol. The VLL includes: the explanatory "Dictionary of the Ukrainian Language", "Russian Language Explanatory Dictionary", "Turkish Language Explanatory Dictionary", "Turkish Language Grammar Dictionary", and the "Ukrainian National Linguistic Corpus".

The Virtual Lexicographic Laboratory "Dictionary of the Ukrainian Language" was also installed at Kharkov National University of Radio Electronics at the Interdepartmental Research Center of Mathematical and Applied Linguistics.

Users of VLL in Kiev:

- Ukrainian Lingua-Information Fund, NAS of Ukraine;
- publishing house "Naukova dumka";
- V.M.Glushkov Institute of Cybernetics, NAS of Ukraine;
- Specialists of ULIF (from home computers)

Thus, we see that actually the overall organization of work over the "Dictionary of the Ukrainian Language" by large linguistic groups is possible. This idea was continued in the project «Ukrainian Linguistic Dialogue» with a goal to unite all lexicographic groups in Ukraine via the system of virtual lexicographic laboratories for the implementation of large dictionary and other linguistic projects.

We hope that this task will form a part of the national project program on the creation of the Ukrainian segment of GRID.

On the other hand, we believe that this system engineering and technology could be used by the consortium MONDILEX for solving its problems.

II. Digital Lexicographic Resources (Corpora and Dictionaries) and their Applications

Standardised Encoding of Morphological Lexica for Slavic Languages[1]

Simon Krek, Tomaž Erjavec
Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
simon.krek@ijs.si, tomaz.erjavec@ijs.si

Abstract
The paper presents a proposal for lexical encoding concentrating on morphological properties of words, with special emphasis given on the rich inflectional properties of Slavic languages. The encoding format is an application of the recently adopted ISO standard LMF, while the core lexical structure and morphosyntactic annotation come from the MULTEXT-East proposal. The paper explains the structure of the MULTEXT-East type of lexica and morphosyntactic annotations, with emphasis on recent extensions introduced for Slovene. Next, the ISO standard LMF is introduced and discussed. On the example of Slovene, we detail the representation of inflectional paradigms, regular derivational relations, variant spellings, etc. The paper concludes with a discussion and directions for further research.

Keywords: *Morphological Lexica, Standardisation, Lexical Markup Framework, MULTEXT-East*

1. Introduction

The paper presents a proposal for lexical encoding, meant to serve as a foundation for the lexicon being developed in the recently started national project "Communication in Slovene". While the lexicon format is being developed for encoding of Slovene language lexica, it is general enough to be applicable to other, esp. Slavic languages, as they all share complex morphology, which the proposal aims to address.

The lexicon currently concentrates on word-forms, and is meant to cover two application areas:

- the use of the lexicon in applications of human language technologies
- the use of the lexicon as a resource to interconnect with other language resources to be developed in the scope of SSJ, in particular a style guide for Slovene.

The two goals impose different requirements on the lexicon: in HLT applications the lexicon must cover as many as possible of words which appear in real texts, including spoken language, and be machine-processable. For the Style guide, however, the lexicon must define normative aspects of the language, and contrast them with the contemporary reality of the Slovene language; and, as far as possible, the lexicon should be human readable.

The lexicon is encoded in XML, with the schema being based on the ISO standard "Lexical Markup Framework",[2] which is the last in long tradition of HLT standardisation projects, starting with EAGLES.[3]

The lexicon, as an HLT resource, must be integrated with morphosyntactically annotated corpora and taggers. The SSJ proposal here relies on the MULTEXT-East[4] defined tagsets, more specifically the tagset defined in the JOS[5] project, itself an outgrowth of the MULTEXT-East proposal for Slovene. The annotations of the corpora are thus a direct mapping from the word-form features defined in the lexicon.

[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] In November 2008 LMF became the international standard ISO-24613:2008 . The Web page or LMF is http://www.lexicalmarkupframework.org/
[3] EAGLES, Expert Advisory Group on Language Engineering Standards: http://www.ilc.cnr.it/EAGLES/home.html
[4] MULTEXT-East, Multilingual Text Tools and Corpora for Central and Eastern European Languages: http://nl.ijs.si/ME/
[5] JOS: Linguistic Annotation of Slovene: http://nl.ijs.si/jos/

2. MULTEXT-East morphosyntactic specifications and lexica

MULTEXT-East morphosyntactic specifications set out the grammar and vocabulary of valid morphosyntactic descriptions, MSDs, which can then serve as a compact representation of word-level syntactic tags, used in tagging of corpora. The specifications determine what, for each language, is a valid MSD and what it means, e.g., that Ncms is a valid MSD for English and is equivalent to the feature-structure PoS:Noun, Type:common, Gender:masculine, Number:singular.

The MULTEXT-East morphosyntactic specifications, currently at Version 3 (Erjavec, 2004) have been developed in the formalism and on the basis of specifications for six Western European languages of the MULTEXT project and in cooperation with EAGLES, the Expert Advisory Group on Language Engineering Standards. Originally, these specifications were released as a report of the MULTEXT-East project but have been revised for both subsequent releases, and have become, if not a standard, then at least a reference for comparison (Erjavec et al., 2003). The complete specifications are structured as a report, and contain introductory chapters, followed by the list of defined categories (parts-of-speech), and then, for each category, a table of attribute-values, and the languages the features are appropriate for. These so called common tables are followed by language particular sections. Each language section is further subdivided, and can contain feature co-occurrence restrictions, examples, notes, and full lists of valid MSDs, as well as localisation information. The formal core of the specifications resides in the common tables, as they define the features, their codes for MSD representation, and their appropriateness for each language - an example is given in Figure 1.

In MULTEXT-East the complete specifications were encoded as a LaTeX document, however, given preliminary work described in Erjavec (2006) the encoding was, in the JOS project, moved to XML, with the schema is based on the TEI,[1] version P5. While the JOS morphosyntactic guidelines are instantiated only for Slovene, the principles could be adopted to MULTEXT-East multilingual specifications, which is, in fact, work in progress. Below we give an example from the specifications, giving the first part of the definition of Noun. As can be seen, the specification is bi-lingual, in Slovene and English, and defines the names of the attributes and their values, as well as the position of the attribute and its code, for mapping to MSDs.

```
− <div type="section" xml:id="msd.N">
    <head xml:lang="sl">SAMOSTALNIK</head>
    <head xml:lang="en">Noun</head>
  − <table n="msd.cat" xml:id="msd.cat.N">
      <head xml:lang="sl">Tabela atributov in vrednosti za samostalnik</head>
      <head xml:lang="en">Attribute-Value Table for Noun</head>
    − <row role="type">
        <cell role="position">0</cell>
        <cell role="name" xml:lang="sl">samostalnik</cell>
        <cell role="code" xml:lang="sl">S</cell>
        <cell role="name" xml:lang="en">Noun</cell>
        <cell role="code" xml:lang="en">N</cell>
      </row>
    − <row role="attribute">
        <cell role="position">1</cell>
        <cell role="name" xml:lang="sl">vrsta</cell>
        <cell role="name" xml:lang="en">Type</cell>
      − <cell role="values">
        − <table>
          − <row role="value">
              <cell role="name" xml:lang="sl">občno_ime</cell>
              <cell role="code" xml:lang="sl">o</cell>
              <cell role="name" xml:lang="en">common</cell>
              <cell role="code" xml:lang="en">c</cell>
            </row>
          − <row role="value">
              <cell role="name" xml:lang="sl">lastno_ime</cell>
              <cell role="code" xml:lang="sl">l</cell>
              <cell role="name" xml:lang="en">proper</cell>
              <cell role="code" xml:lang="en">p</cell>
            </row>
          </table>
        </cell>
      </row>
```

[1] TEI, Text Encoding Initiative: http://www.tei-c.org/

The XML encoding of JOS morphosyntactic specifications brings with it a number of benefits, in particular that the source XML specifications can be, with XSLT stylesheets, directly transformed either into HTML suitable for browsing, or into tabular files, which give the conversion of the MSD set into various feature-structure representations, a key requirement if the corpus MSDs are to be converted to lexical feature-structures.

The MULTEXT-East / JOS type specifications are quite expressive, and meet both criteria for expressivity and compact representation necessary for corpus annotation.

MULTEXT-East also defined the format for lexicon encoding, which is a simple tabular format, with one entry per line comprising three fields: the word-form, its lemma, and its MSD, e.g. walks walk Ncnp. In MULTEXT-East the convention was to include in the lexicon complete inflectional paradigms of words, i.e. they give an extensional account of the inflectional morphology of the languages.

But while MULETEXT-East lexica offer a simple and relatively compact representation, they are not very expressive: they do not distinguish between homonymous lemmas, and do not allow for including further information about entries, e.g. frequency of usage, normative reference, or derivational relations. These are the main reasons why it was decided to move to a more expressive encoding, in particular the LMF standard.

3. Lexical Markup Framework

Lexical Markup Framework[1] (LMF) is the ISO International Organization for Standardization ISO/TC37 standard for natural language processing (NLP) and machine-readable dictionary (MRD) lexicons. The scope is standardization of principles and methods relating to language resources in the contexts of multilingual communication and cultural diversity. The goals of LMF are to provide a common model for the creation and use of lexical resources, to manage the exchange of data between and among these resources, and to enable the merging of large number of individual electronic resources to form extensive global electronic resources.

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

LMF is composed of the following components:

- The core package which is the structural skeleton which describes the basic hierarchy of information in a lexical entry.

- Extensions of the core package which are expressed in a framework that describes the reuse of the core components in conjunction with the additional components required for a specific lexical resource.

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

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

4. The SSJ lexicon proposal

The SSJ lexicon proposal takes LMF as its format while using the JOS morphosyntactic specifications to express parts-of-speech, their attributes and values. Currently, the attribute values of various features and the morphosyntactic properties are expressed in Slovene, however, it should be noted that it is not difficult to localise these into English, as the JOS specification is bi-lingual.

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

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

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

As can be seen, a lexical entry is assigned an ID, which uniquely identifies the entry; in case several entries have the same lemma, the ID is decorated with a number, to distinguish homonymous entries. The lexical entry then specifies which part of speech it belongs to. More generally, the top level features contain all the invariant features of the lemma, such as gender for nouns. Next comes the lemma form, with a feature specifying how the lemma form is written. The lemma is still an abstract form, not meant as a particular word-form to be found in text. Finally, the lexical entry specifies the word-form or word-forms that constitute its paradigm.

4.2 Inflectional paradigms

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

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

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

4.2 Derivational relations

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

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

In SSJ we plan to mark with related form only regular derivational relations.

4.3 Variant spellings

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

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

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

Conclusions

The paper has shown a standardised way of encoding word-level syntactic information for morphological lexica, using the Lexical Markup Framework and the MULTEXT-East proposal. We introduced the encoding, giving examples of inflection, derivation, and variant forms. There are other principles that were taken into account when laying down the guidelines for lexicon construction, however, they are mostly language specific and could well be different for other Slavic languages.

In the next stage we plan to operationalise the suggested format by constructing a reference lexicon for Slovene, which will also validate the proposal in practice. Not mentioned in the paper is also the question of meta-data, where LMF provides the structure for only very basic information. We are currently looking at ways to extend the meta-data set, say, by taking the TEI header as the starting point.

References

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

Erjavec, T., C. Krstev, V. Petkevič, K. Simov, M. Tadić, and D. Vitas (2003). The MULTEXT-East Morphosyntactic Specifications for Slavic Languages. In Proceedings of the EACL 2003 Workshop on Morphological Processing of Slavic Languages. Budapest.

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004 (1535–1538). Paris: ELRA.

Erjavec, T. MULTEX-East morphosyntactic specifications and XML (2006). In: SLAVCHEVA, M., SIMOV, K., ANGELOVA, G. (eds). Readings in multilinguality : selected papers for young researchers. Sofia: Institute for Parallel Processing, Bulgarian Academy of Science, 2006, pp 41-48.

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

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

A New Version for Bulgarian MTE Morphosyntactic Specifications for Some Verbal Forms[1]

Ludmila Dimitrova[1], *Peter Rashkov*[2]
[1]Institute of Mathematics and Informatics
Bulgarian Academy of Sciences, Sofia, Bulgaria
ludmila@cc.bas.bg
[2]Jacobs University, Bremen, Germany

Abstract
In this paper we propose a new version for MULTEXT-East morphosyntactic specifications for Bulgarian participles.

Keywords: *Bulgarian, impersonal verbal forms, grammatical attribute, morphosyntactic descriptors*

Introduction

The morphosyntactic descriptors for Bulgarian in the framework of the MULTEXT-East (MTE for short) project were developed 12 years ago. Some of them are not strictly adequate to the particular morhosyntactic properties of the respective parts-of-speech – Tense, Number, Gender, Voice, Definiteness – especially in the system of impersonal verbal forms (participles) (MTE 2004).

The system of participles in modern Bulgarian contains a multitude of forms with different origins. Its main elements are the two past participles (perfect and imperfect) and the past passive participial which are characteristic of the spoken language for a long time, as well as the present active participial, which has been re-introduced in the literary language in the nineteenth century. The indeclinable gerund has its own special place, because its distribution does not cover all Bulgarian dialects, yet it is widely used in the written language.

According to the Bulgarian grammarians, Bulgarian participles do not possess the grammatical attribute Voice because the relation between the subject and the verb action conveyed by a participle is attributive (*плачещо дете, подранила зима*), not predicative (*детето плаче, зимата подрани*). In fact, the Bulgarian language (as well as all other Slavic languages) have no special forms for passive voice at all and express it periphrastically.

Yet the Bulgarian classification distinguishes between active participles (деятелни причастия)
- present active participial (Bulg. *сегашно деятелно причастие*):
четящ, ходещ, разказващ,
- perfect active participle (in Bulg. *минало свършено деятелно причастие, formed from the aorist stem of the verb + suffix* ***-l-***):
*живя-**х** → живя-**л**, ходи-**х** → ходи-**л**, писа-**х** → писа-**л***;
imperfect active participle (Bulg. *минало несвършено деятелно причастие, formed from the imperfect stem of the verb + suffix* ***-l-***):
*живее-**х** →живее-**л**, ходе-**х** → ходе-**л**, пише-**х** → пише-**л***;

and passive participles (*страдателни причастия*)
- past passive participial (Bulg. *минало страдателно причастие, formed from the imperfect stem of the verb + suffix* ***-н-*** *or* ***-т-***):
*писа-**х** → писа-**н**, игра-**х** → игра-**н**, разказва-**х** → разказва-**н***,
*ши-**х** → ши-**т**, би-**х** → би-**т**, чу-**х** →чу-**т**, обу-**х** → обу-**т***;

[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

- present passive participial (Bulg. *сегашно страдателно причастие*): *любим, необходим, неделим*, which existed in its own right in Old Bulgarian, but survives only in a few cases and is considered an adjective.

The participles are differentiated according to tense and aspect:

The grammatical category tense is not fully developed either, and the distinction between the so-called present (*сегашни причастия*) and past participles (*минали причастия*), according to the Bulgarian classification is rather due to their historical origins, namely the present or past verb stem, from which the participle is formed, than to any kind of temporal relations.

The grammatical category aspect has also a limited place in the participle system – only imperfective verbs can form the present active participial, while both perfective and imperfective verbs form past participles.

The grammatical categories gender, number, and definiteness apply to participles in a similar way as to adjectives. The categories gender and number are natural to all participles, and in this way the participles resemble adjectives. Definiteness applies to the active participial, past passive and perfect active participles. Hence we may speak of the Bulgarian participle as a hybrid part-of-speech, where each grammatical attribute is only partially represented. This hybridism has attracted the attention of the first modern Bulgarian grammarians (such as Neofit Rilski[1]) and made them classify the participles as a separate POS.

Present active participial

The MSD of this type of participle (Bulg. сегашно деятелно причастие) contains VForm=participle(p), Tense=present(p), Voice=active(a), for example,

въртя**що**то	въртя	Vmpp-sna-y
дължа**ща**	дължа	Vmpp-sfa-n
изгрява**ща**та	изгрявам	Vmpp-sfa-y
изгрява**щи**	изгрявам	Vmpp-p-a-n

The present active participial is a new phenomenon in modern Bulgarian. It disappeared from the spoken language in the historical development of the language during 15-16 century, but was frequently used in Old Bulgarian, for example:

- **Ходѧ же при мори галилеѥсцѣ видѣ Симона и Аньдрѣѩ, братра того Симона, въмѣтаѩшта мрѣжѣ вь морѥ.** (Mar.) /As he went along the Sea of Galilee, he saw Simon and Andrew, Simon's brother, casting a net into the sea./ [Mark 1:16]

The reintroduction of this verbal form started in the 19 century under Russian and Church Slavic influence. Today this participial has its own specific structure (stem and suffix) in contrast to the same category in other Slavic languages. It represents a property/attribute of a person or an object resulting from an action of the person or the object regardless of any temporal orientation. This participial expresses the property/attribute independently and regardless of the tense of the main verb in the sentence. It corresponds semantically to the English *ing*-form (English present participle), when it modifies a noun or serves as a noun with active sense. For example:

- *Преподавателят посреща* (*посреща* is verb in present tense) ***идващите*** *ученици*. /The lecturer meets the *coming* students/

- *Преподавателят посрещна* (*посрещна* is verb in aorist tense) ***идващите*** *ученици*. /The lecturer met the coming students/

- *Преподавателят ще посрещне* (*ще посрещне* is verb in future tense) ***идващите*** *ученици*. /The lecturer will meet the incoming students/.

It also functions as an attribute and behaves as a noun or an adjective in the sentence:

- *В стаята имаше много* ***правостоящи***. /There were many *standing* [people] in the room/.

- *Пази се от* ***падащи*** *предмети*! /Beware of *falling* objects!/.

Furthermore, it can serve as a marker for a subordinate clause, for instance:

- *По улиците се мяркаха тук-там минувачи, неспокойно* ***поглеждащи*** *към сивото небе*. /On the streets one caught here and there a glimpse of passers-by, restlessly *glancing* toward the grey sky./

[1] Neofit Rilski published *Bolgarska Grammatika*, the first systematic Bulgarian grammar, in 1835.

This participial cannot possess the Tense attribute because it expresses the property/attribute independently and regardless of the tense of the main verb in the sentence (Bulgarian Grammar 1990). The Voice attribute is also implicit from the context. Hence we propose a new attribute VForm=active_participial(a) in MSD, which replaces the old descriptor in the following manner:

		MSD old version		MSD new proposal
въртящото	въртя	Vmpp-sna-y	->	Vma--sn--y
дължаща	дължа	Vmpp-sfa-n	->	Vma--sf--n
изгряващата	изгрявам	Vmpp-sfa-y	->	Vma--sf--y
изгряващи	изгрявам	Vmpp-p-a-n	->	Vma--p---n

Perfect active participle

The MSD of the perfect active participle (Bulg. *Минало свършено деятелно причастие*) contains VForm=participle(p), Tense=aorist(a), Voice=active(a), for example,

отишл**а** отида Vmpa-sfa-n
отишл**и** отида Vmpa-p-a-n
отишъл отида Vmpa-sma-n
отрекл**и** отрека Vmpa-p-a-n

From a historical perspective, this participle is quite old, and it is frequently used in Old Bulgarian in the formation of compound tenses, for example,

- **И по чъто нє въдастъ моєго сръебра пѣнѧзьникомъ и азъ пришєдъ съ лихвоѭ истѧзалъ є бимъ?** (Mar.) /Why then didn't you put my money in the bank, so that when I returned I *could have collected* it with interest?/ [Luke 19:23]

It is consequently stable across all Bulgarian dialects in form, semantics and functionality. Its forms are derived from the aorist stem of the verbs + suffix *-л, -ла, -ло,-ли: писа-х > писа-л, живя-х > живя-л, пи-х > пи-л*, etc. This participle has significantly expanded its usage and in modern Bulgarian it expresses a property/attribute resulting from an action that has been completed before the action of the main verb in the sentence, but regardless of the moment of speaking. For example:

- *Падналите ни другари* /our fallen comrades/
- *Гледам* ***дошлия*** *войник*. /I am looking at the soldier who came./
- *Гледах* ***дошлия*** *войник*. /I was looking at the soldier who had come./
- *Ще гледам* ***дошлия*** *войник*. /I will look at the soldier who will come./

The perfect active participle shares all morphosyntactic characteristics of the noun and adjective – number, gender and definiteness. For example:

- ***Закъснелите*** *ученици стояха настрана*. /The students who were late stood aside./
- *Пътниците слизаха бързо от* ***пристигналия*** *влак*. /The passengers quickly got off the arrived train (the train that had arrived)./
- ***Дошлите*** *се бяха насъбрали пред училището*. /The ones who had come had gathered in front of the school./

Furthermore, it can serve as a marker for a subordinate clause (Bulgarian Grammar 1990), for instance:

- *Година беше се минало откак Василчо,* ***яхнал*** *белия си кон, се изгуби от очите на Божура*. (Y. Yovkov, 1976) /A year passed since Vasilčo, *who had straddled* his white horse, had disappeared from Božura's sight./

This participle plays a major role in the formation of compound verb tenses and moods (Bulgarian Grammar 1990), for example:

- Indicative: *ходил съм* (perfect), *бях ходил* (plusquamperfect), *ще съм ходил* (future perfect), *щях да съм ходил* (future perfect in the past)
- Re-narrative (inferential): *ходил съм* (aorist) /I [allegedly] went/, *ходил съм бил* (perfect, plusquamperfect) /I [alledgedly] had gone/, *щял съм да ходя* (future and future in the past) /I [allegedly] will go, I [allegedly] would go/, щял съм да съм ходил (future perfect and future perfect in the past) /I [allegedly] will have gone, I [allegedly] would have gone/
- Conditional: *бих ходил /I would go/*.

Taking all this into account, we see that its MSD is not quite adequate since the attribute Tense=aorist(a) merely reflects the historic origin of the perfect active participle , derived from the aorist stem of the verbs (*живя-, лежа-, ходи-, гледа-*). We see that this participle is used to form many more verb tenses and moods, hence its MSD should not contain the morphosyntactic attribute Tense. This participle has no morphosyntactic attribute Voice, because it is an impersonal verbal form; the grammatical voice is implicit from the context. Hence, we only propose the following simplification of the MSD:

		MSD old version		MSD new proposal
буталото	бутам	Vmpa-sna-y	->	Vmp--sn--y
бухналата	бухна	Vmpa-sfa-y	->	Vmp--sf--y
возилото	возя	Vmpa-sna-y	->	Vmp--sn--y
отишла	отида	Vmpa-sfa-n	->	Vmp--sf--n
отишли	отида	Vmpa-p-a-n	->	Vmp--p---n
отишъл	отида	Vmpa-sma-n	->	Vmp--sm--n
отрекли	отрека	Vmpa-p-a-n	->	Vmp--p---n

Imperfect active participle

The MSD of the imperfect active participle (Bulgarian *минало несвършено деятелно причастие*) contains VForm=participle(p), Tense=imperfect(i), Voice=active(a), for instance,

благодаря**ла**	благодаря	Vmpi-sfa-n
вадe**л**	вадя	Vmpi-sma-n
грешa**ла**	греша	Vmpi-sfa-n
деря**ла**	дера	Vmpi-sfa-n
правe**л**	правя	Vmpi-sma-n
правe**ла**	правя	Vmpi-sfa-n

This participle is a relatively new POS in Bulgarian morphology, as it did not exist in Old Bulgarian and has developed in parallel to the re-narrative mood of the Bulgarian verb. Hence, its MSD is not quite adequate – the Tense=imperfect(i) attribute reflects merely the historic origin of the imperfect participle, as its forms are derived from the imperfect stem of the verb (*живее-, лежа-/леже-, ходе-, гледа-*). ***The main role of this participle is to express renarrative, or non-witnessed modality, hence it cannot be used as an attribute. It possesses only the grammatical attributes number and gender, and is exclusively used to form some compound verb tenses in renarrative (or inferential) mood.***

It is formed from the imperfect stem + suffix *-л: пише-х > пише-л, живее-х > живее-л, пие-х > пие-л*, etc. Sometimes the form of this participle coincides with the form of the first past participle because the imperfect and aorist stems of the verbs coincide. This is the case for some verbs from second conjugation (*вървял, търпял*, etc.) and all verbs from third conjugation (*гледал, стрелял*, etc.).

Some example of the usage are for forming present and imperfect tense in renarrative mood, e.g. ***ходел*** *съм* /I [allegedly] have been going/, as well as in the modal form *нямало* used to form negative forms of future, future in the past and future perfect in renarrative mood: ***нямало*** *да ходя,* ***нямало*** *да съм донесъл*.

Due to this highly specialized semantic function we propose a new language-specific VForm=renarrative(r) in MSD, which replaces the old descriptor in the following manner:

		MSD old version		MSD new proposal
благодаряла	благодаря	Vmpi-sfa-n	->	Vmr--sf
вадел	вадя	Vmpi-sma-n	->	Vmr--sm
грешала	греша	Vmpi-sfa-n	->	Vmr--sf
деряла	дера	Vmpi-sfa-n	->	Vmr--sf
правел	правя	Vmpi-sma-n	->	Vmr--sm
правела	правя	Vmpi-sfa-n	->	Vmr--sf
отидели	отида	Vmpi-p-a-n	->	Vmr--p

Past passive participial

The MSD of the past passive participial (Bulgarian *минало страдателно причастие*) contains VForm=participle(p), Tense=past(s), Voice=passive(p), for instance,

отрязан**ни**	отрежа	Vmps-p-p-n
отрязан**ни**те	отрежа	Vmps-p-p-y
предаде**н**	предам	Vmps-smp-n
прочете**н**	прочета	Vmps-smp-n

This participial is also an old form, which has not changed much semantically or formally over the centuries, for example it is used in Old Bulgarian as an attribute:

– **Придѣтє кь мънѣ вьси трѫждаѩштєи сѧ и обрѣмєнєнии.** (Mar.) /Come to me, all you who are weary and *burdened*./ [Matthew 11:28]

It is formed from the aorist verb stem + suffix *-н/-т: казах > казан, пренесох > пренесен, ших > шит, носих > носен, измих > измит*, etc. The main function of this participial is to convey the result of an action as an attribute of a person or object (Bulgarian Grammar 1990). In this way it is very close to the adjective and the noun, with which it shares the morphosyntactic attributes Number, Gender and Definiteness. It corresponds to the English past participle when it modifies or serves as a noun with passive sense. For example:

- *Вечер* […] *грееха огнените сияния на* ***запалените*** *села*. (Y. Yovkov, 1976) /In the evenings the fiery radiance of the *burnt* villages was shining./

- ***Поканените*** *чакат в приемната зала*. /The *invited* ones wait in the parlour./

Furthermore, it can serve as a marker for a subordinate clause, for instance:

И те носели една книга, ***намерена*** *в Света гора*, […] (Y. Yovkov, 1976) /And they supposedly carried a book *found* in Mount Athos, […]/

Additionally it is used to form one type of periphrastic passive voice in Bulgarian, similar to the English formation.

Стоките ***ще бъдат прегледани*** *от комисията*. /The merchandise will be inspected by the commission./, but compare the reflexive-passive form: *Стоките* ***ще се преглеждат*** *от комисията*.

This participial cannot possess the Tense or Voice attribute because it is an impersonal verb form and both the temporal characteristic and the relationship between the action (or state) expressed by the verb stem and the participant (object) are presented attributively, not predicatively. Hence we propose a new attribute VForm=passive_participial(v) in MSD, which replaces the old descriptor in the following manner:

		MSD old version		MSD new proposal
отрязани	отрежа	Vmps-p-p-n	->	Vmv--p---n
отрязаните	отрежа	Vmps-p-p-y	->	Vmv--p---y
предаден	предам	Vmps-smp-n	->	Vmv--sm--n
прочетена	прочета	Vmps-smp-n	->	Vmv--sf--n

Conclusion

We hope that these changes, which bring the morphosyntactic description in line with the grammatical characteristics of the Bulgarian participles, will be more useful for automated processing of Bulgarian texts.

References

MTE, 2004: MULTEXT-East Morphosyntactic Specifications – version 3, edition 10th May 2004.

Bulgarian Grammar, 1993: Главна редакция Д. Тилков, Ст. Стоянов, К. Попов. Граматика на съвременния български книжовен език. Том 2 / МОРФОЛОГИЯ. Издателство на БАН. София. (In Bulgarian).

Y. Yovkov, 1976: Йордан Йовков. Старопланински легенди. В Събрани съчинения в шест тома. Том втори, Издателство "Български писател", София, под общата редакция на Симеон Султанов. (In Bulgarian).

Mar.: Codex Marianus (10 century manuscript), 7-bit ASCII code at http://www.slav.helsinki.fi/ccmh/marianus.html

Appendix

1. Common table for Verb:

Verb (V)
14 Positions

PoS	Type	VFrm	Tens	Pers	Numb	Gend	Voic	Neg	Def	Cltc	Case	Anim	Clt2
=	==========	==========	==========	==========	==========	==========	==========	===	EN	RO	SL	CS	**BG**	ET	HU
P	ATT		VAL					C	x	x	x	x	x	x	x
=	==========	==========	==========	==========	==========	==========	==========	===							
1	Type		main					m	x	x	x	x	x	x	x
			auxiliary					a	x	x	x	x	x	x	x
			modal					o	x	x	x	x		x	
			copula					c		x	x	x			
		l.s.	base					b	x						
-	----------	----------	----------	----------	----------	----------	----------	-							
2	VForm		indicative					i	x	x	x	x	x	x	x
			subjunctive					s		x					
			imperative					m		x	x	x	x	x	x
			conditional					c	x		x	x		x	x
			infinitive					n	x	x	x	x		x	x
			participle					p	x	x	x	x	x	x	
			gerund					g		x			x	x	
			supine					u			x			x	
			transgressive					t			x				
		l.s.	quotative					q						x	
		l.s.	**renarrative**					**r**					**x**		
		l.s.	**active_participial**					**a**					**x**		
		l.s.	**passive_participial**					**v**					**x**		
-	----------	----------	----------	----------	----------	----------	----------	-							
3	Tense		present					p	x	x	x	x	x	x	x
			imperfect					i		x			x	x	
			future					f			x	x			
			past					s	x	x	x	x		x	x
		l.s.	pluperfect					l		x					
		l.s.	aorist					a					x		
-	----------	----------	----------	----------	----------	----------	----------	-							
4	Person		first					1	x	x	x	x	x	x	x
			second					2	x	x	x	x	x	x	x
			third					3	x	x	x	x	x	x	x
-	----------	----------	----------	----------	----------	----------	----------	-							
5	Number		singular					s	x	x	x	x	x	x	x
			plural					p	x	x	x	x	x	x	x
		l.s.	dual					d			x				
			collective					l							
-	----------	----------	----------	----------	----------	----------	----------	-							
6	Gender		masculine					m			x	x	x		
			feminine					f			x	x	x		
			neuter					n			x	x	x		
**********															
7	Voice		active					a				x	x	x	x
			passive					p				x	x		x
-	----------	----------	----------	----------	----------	----------	----------	-							

2. Application to Bulgarian Verb

Verb (V)

P	ATT		VAL	Example	C
1	Type		main	govorya	m
			auxiliary	sym	a
2	VForm		indicative	govorya	i
			imperative	govorete	m
			participle	govoril	p
		l.s.	renarrative	govorel	r
		l.s.	active_participial	govorest	a
		l.s.	passive_participial	govoreno	v
			gerund	govorejki	g
3	Tense		present	govorya	p
			imperfect	govoreh	i
		l.s.	aorist	govorih	a
4	Person		first	govorya	1
			second	govorish	2
			third	govori	3
5	Number		singular	govorya	s
			plural	govoryat	p
6	Gender		masculine	govoril	m
			feminine	govorila	f
			neuter	govorilo	n
7	Voice		active	govorya	a
8	Negative				-
9	Definiteness		no	govoril	n
			yes	govorilite	y
			short_art	govoriliya	s
			full_art	govoriliyat	f
10	Clitic				-
11	Case				-
12	Animate				-
13	Clitic_s				-

Combinations

PoS	Type	VFrm	Tens	Pers	Numb	Gend	Voic	Neg	Def	Cl1	Case	An	Cl2	Example
V	[ma]	i	[pai]	[123]	[sp]	-	a	-	-	-	-	-	-	1.
V	[ma]	m	-	2	[sp]	-	a	-	-	-	-	-	-	2.
V	m	p	-	-	s	[mfn]	-	-	n	-	-	-	-	3.
V	m	a	-	-	s	m	-	-	[sf]	-	-	-	-	4.
V	m	p	-	-	s	[fn]	-	-	[ny]	-	-	-	-	5.
V	m	r	-	-	s	[mfn]	-	-	-	-	-	-	-	6.
V	m	a	-	-	p	-	-	-	[ny]	-	-	-	-	7.
V	m	v	-	-	s	[mfn]	-	-	n	-	-	-	-	8.
V	m	v	-	-	s	m	-	-	[sf]	-	-	-	-	9.
V	m	v	-	-	s	[fn]	-	-	y	-	-	-	-	10.
V	m	v	-	-	p	-	-	-	[ny]	-	-	-	-	11.
V	a	[pr]	-	-	s	[mf]	-	-	n	-	-	-	-	12.
V	a	p	-	-	s	m	-	-	[sf]	-	-	-	-	13.
V	a	p	-	-	s	[fn]	-	-	y	-	-	-	-	14.
V	a	[pr]	-	-	p	-	-	-	[n-]	-	-	-	-	15.
V	a	p	[a]	-	p	-	-	-	y	-	-	-	-	16.
V	[ma]	g	-	-	-	-	-	-	-	-	-	-	-	17.
V	m	r	-	-	p	-	-	-	-	-	-	-	-	18.
V	m	a	-	-	s	[mfn]	-	-	n	-	-	-	-	19.

Examples:

1. govorya, sym
2. govori, bydete
3. govoril, govorila, govorilo - like noun/adjective
4. govorestiya, govorestiyat - like noun/adjective
5. govorila, govoriloto - like noun/adjective
6. govorel, govorela, govorelo -renarrative, only in verbal use
7. govoresti, govorestite - like noun/adjective
8. govoren, govorena, govoreno - like noun/adjective
9. govoreniya, govoreniyat - like noun/adjective
10. govorenata, govorenoto - like noun/adjective
11. govoreni, govorenite - like noun/adjective
12. bil, bila - like noun/adjective
13. biliya, biliyat - like noun/adjective
14. bilata, biloto - like noun/adjective
15. bili -renarrative and participle form coincide - see table
16. bilite - like noun/adjective
17. govorejki, bidejki
18. govoreli -renarrative, only in verbal use
19. govorest, govoresta, govoresto - like noun/adjective

Comparing Bulgarian and Slovak Multext-East morphology tagset[1]

Ludmila Dimitrova[a)], *Radovan Garabík*[b)], *Daniela Majchráková*[b)]
a) Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
1113 Sofia, Bulgaria
ludmila@cc.bas.bg
b) Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
813 64 Bratislava, Slovakia
korpus@korpus.juls.savba.sk, http://korpus.juls.savba.sk

Abstract
We analyse the differences between the Bulgarian and Slovak languages Multext-East morphology specification (MTE, 2004). The differences can be caused either by inherent language dissimilarities, different ways of analysing morphology categories or just by different use of MTE design guideline. We describe all the parts of speech in detail with emphasis on analysing the tagset differences.

Keywords: *Bulgarian, Slovak, grammatical category, morphology specification, morphology tagset*

Introduction

The EC project MULTEXT *Multilingual Tools and Corpora* produced linguistics resources and a freely available set of tools that are extensible, coherent and language-independent, for seven Western European languages: English, French, Spanish, Italian, German, Dutch, and Swedish (Ide, Veronis, 1994). The EC INCO-Copernicus project MULTEXT-East *Multilingual Text Tools and Corpora for Central and Eastern European Languages* is a continuation of the MULTEXT project. MULTEXT-East (MTE for short; Dimitrova et al., 1998) used methodologies and results of MULTEXT. MTE developed significant language resources for six Central and Eastern European (CEE) languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene, as well as English. Three of these languages (Bulgarian, Czech, and Slovene) belong to the Slavic language group. The results of the two projects MULTEXT and MTE are:

- tools, corpora, and linguistic resources for thirteen western and eastern European languages, with extensions to regional languages (Catalan, Occitan) and non-European languages (Bambara, Kikongo, Swahili);
- experience of developing standards and specifications for encoding of linguistic corpora;
- experience of using the same program tools for the processing of linguistic corpora.

These results show how important the development of common, harmonised and unified resources for different European languages and the language independence of the tools employed are.

The MTE electronic linguistics resources include a multilingual corpus and datasets of language-specific resources. The language-specific resources that the MTE project developed are: morphosyntactic specifications, language-specific data, and lexica.

Bulgarian morphosyntactic specifications have been made in the frame of the MTE project, but they are based on a semantic part-of-speech classification of the traditional Bulgarian grammar.

Slovak language morphology specification compatible with the MTE tagset has been developed as a projection of the Slovak morphology tagset used at the Ľ. Štúr Institute of Linguistics (Garabík, 2006), which (pragmatically) influences some parts of the specification design.

The aim of this article is to compare the differences between Slovak and Bulgarian MTE specification. Specifically, our goal is *not* to compile a list of grammar differences between the languages – we only gloss over them as far as they influence the morphosyntactic tagsets used.

[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 tagset differences describer can be separated into three different categories:

1. Differences due to inherent differences between the languages. For example, Bulgarian has lost (up to few exceptions) the Proto-Slavic case system, while Slovak keeps it almost fully – subsequently, the Case attribute is present only sporadically in the Bulgarian tagset, while the Slovak Case category is ubiquitous. We include also the differences resulting from orthography tradition here, since we are primarily dealing with the written language, where the orthography forms an inseparable part of language analysis.

2. Differences due to different way of analysing the morphology, either as described by traditional grammars, or by different design decisions in our tagsets. Most notably, Slovak tagset strives to cover the morphology at the lowers possible level and assumes thorough tokenization into the smallest possible units – there are no multi word tokens in the Slovak tagset (each part of such an expression will be assigned its own tag), while in Bulgarian multiword expressions are common (e.g. Bulgarian expression ***дявол да го вземе*** will be classified as interjection, while the Slovak ***čert ho ber*** will be analysed as three independent words, noun+pronoun+verb – the two expressions are otherwise identical in both languages).

3. Different way of putting grammar information into the Multext East tagset. Since the Bulgarian and Slovak tagsets were created independently, using only the MTE guidelines as a common references, there are some features that have no base neither in the primary grammar differences, nor in traditional descriptions, but rather reflect the ambiguity of categorization of grammar features in the scope of MTE. The Slovak MTE tagset is secondary to a morphosyntactic tagset developed to analyse Slovak language in the Slovak National Corpus (Garabík et al., 2004) – in fact, there is also an automatic algorithm mapping the corpus tagset into the MTE one, therefore its design is in some points influenced by the primary tagset as well.

Several words on terminology used: Category is a part-of-speech, consisting of Noun(N), Verb(V), Adjective(A), Pronoun(P), Determiner(D), Article(T), Adverb(R), Adposition(S), Conjunction(C), Numeral(M), Interjection(I), Residual(X), Abbreviation(Y), Particle(Q). Each category has one or several attributes, and each attribute can have exactly one value (including special value '-', meaning 'not applicable'). Throughout the article, we write the one letter abbreviation of a specific category or value in parentheses after the full name. To differentiate the established meaning of *grammar* category from the MTE Category term, we always use the expression *grammar category* for the former.

Values (but not the whole categories or attributes) used only in one of the MTE languages are denoted as 'language specific' in the MTE specification and we mark them with the [l.s.] abbreviation following the value name.

Common differences

There are some features specific for both – Bulgarian and Slovak - languages, which occur repeatedly in several categories, and which we describe here at the beginning, to avoid unnecessary repetition.

Case attribute

Old-Bulgarian had an elaborate case system – there were three numbers for nouns, for example, and seven cases for each of these three numbers. In the process of development of Bulgarian from a synthetic/inflectional language to an analytic/flectional language, case forms were replaced with combinations of different prepositions with a common case form. Case forms then dropped out, and only some have remained in the language until current day. Bulgarian has lost most of the traditional old Slavic case system. For nouns, best preserved is the vocative form, which has survived in the proper names (mostly in given names and some other typically addressee nouns (***Иване, жено, народе*** /*Ivan, woman, folks*/). In some local dialects, the genitive-accusative form is well preserved with proper male name noun forms: ***Тичай до Ивана, до Стояна*** (instead of ***до Иван, до Стоян***) /*to Ivan, to Stoyan*/, ***Кажи на Димитра*** (instead of ***на Димитър***) /*to Dimităr*/.

Most case forms have been preserved, in a systematic form, as related to pronouns (Bulgarian Grammar, 1993). Some of the Bulgarian pronouns keep the difference in nominative(n), dative(d) and accusative(a) cases.

There are no cases anywhere else, and the Case attribute is marked as 'not applicable'.

Slovak keeps the complete case paradigm for nouns, adjectives, (nominal and adjectival) pronouns, participles, and numerals, with the old Slavic vocative surviving only in some fossilized forms (***pane, bože, otče*** /*sir, god, father*/) and a new vocative emerging for some given names or close family relations (***Zuzi, Pali, oci, babi*** /*Zuza, Paľo, dad, grandma*/).

Definiteness attribute

One of the most important grammatical characteristics of the new Bulgarian language which sets it apart from the rest of the Slavic languages is the existence of a definite article. The definite article is a morphological indicator of the grammatical category determination (definiteness). The definite article is not a particle (particles are a separate category of words – parts-of-speech, while the article is not a separate word), nor is it a simple suffix, but a meaningful compound part of the word. It is a word-forming morpheme, which is placed at the end of words in order to express definiteness, familiarity, acquaintance (Bulgarian Grammar, 1993). In Bulgarian, nouns, adjectives, numerals, and full-forms of the possessive pronouns and participles can acquire an article.

For singular masculine, there are two forms: a full article(f)[l.s.] and a short article(s)[l.s.]. The full article is used when a singular masculine form is the syntactic subject of the clause, otherwise a short one is used – a purely orthographic rule. The distinction of full vs. short is not made for feminine, neuter and plural forms, and we use just the yes(y) or no(n) to mark definiteness or respectively lack thereof. Therefore, the definiteness attribute can take overall 4 different values: indefinite(n), definitive(y), short article(s), full article(f).

Examples:

Feminine:

жена, жената */a woman, the woman/*	***жени, жените*** /women, the women/
жена = Ncfs-n	**жени = Ncfp-n**
жената = Ncfs-y	**жените = Ncfp-y**

Neutrum:

дете, детето /a child, the child/	***деца, децата*** */children, the children/*
дете = Ncns-n	**деца = Ncnp-n**
детето = Ncns-y	**децата = Ncnp-y**

Masculine:

мъж, мъжа, мъжът */a man, the man – short art., the man – full art./*	***мъже, мъжете*** */men, the men/*
мъж = Ncms-n	**мъже = Ncmp-n**
мъжа = Ncms-s	**мъжете = Ncmp-y**
мъжът = Ncms-f	

Slovak lacks the definiteness attribute altogether.

Animate attribute

For Slovak, the Animate attribute can be thought of as a subattribute of the masculine gender, where the words in masculine split into two categories, the animate and inanimate one. The feminine and neuter do not have this grammar category[1]. The animate is mostly used for nouns related to persons and animals. Animals are animate in the singular, but in the plural they can be both animate and inanimate, depending on the level of human characteristics assigned to them (often metaphorically). There are some borderline cases, which can be thought of as animate or inanimate in the singular as well (robot, as a thinking being is mostly animate, but as a mechanical tool is inanimate), or the animate feature distinguishes homonyms (***kohútik*** /*rooster*/ *is* animate*, but* ***kohútik*** /*water tap*/ inanimate).

For Bulgarian there is no animate attribute at all, and it is marked as 'not applicable'.

[1] Sometimes a different description is used, where all the non-masculine words are inanimate by default. This is however not according to the mainstream linguistic terminology and leads to some singularities, like the word ***žena*** */woman/* being inanimate.

Part of speech specific differences

Noun

The noun in

Bulgarian possesses the grammatical categories gender, number, definiteness, and (traces of) case. The noun in Slovak possesses the categories gender, number, case, and (sometimes) animateness. In both Slovak and Bulgarian, the gender is invariable and independent of word-formation. Every noun possesses one of three grammatical genders – a masculine, feminine or neuter[1].

Nouns have a singular and plural form, i.e. grammatical meaning of singular number and grammatical meaning of plural number, determined by given suffix morphemes. While in Slovak Number=singular(s) and Number=plural(p) are the only allowed values for the Number attribute, in Bulgarian there is the third value, the so-called count form, marked by Number=count(t)[l.s.]. This special count form in -a/-я originates from the proto-Slavic dual form. The count form appears after a cardinal numeral form (for example, ***два*** /*two*/, ***три*** /*three*/, ***четири*** /*four*/ etc.) or after the adverbs ***колко*** /*how many*/, ***толкова*** /*that many*/, ***няколко*** /*several, a few*/ with masculine nouns that end with a consonant and that do not denote persons, for example: ***два града*** /*two towns*/, ***три стола*** /*three chairs*/, ***четири цвята*** /*four colours*/, ***колко лева*** /*how many levs*/, ***няколко броя*** /*a few copies, issues*/. The count form does not appear after other adverbs such as ***много*** /*many*/, ***малко*** /*few*/, for example ***много столове*** /*many chairs*/ vs. ***три стола*** /*three chairs*/ (Bulgarian Grammar, 1993).

Slovak keeps full featured case morphology, while Bulgarian distinguishes only nominative(n) and vocative(v) – see the discussion on cases above.

In Slovak, there is the Animate attribute, which is completely absent from Bulgarian.

Animate is differentiated only for Gender=masculine(m) and only in these cases:

1. Type=proper(p)
2. Type=common(c) & Case=accusative(a)
3. Type=common(c) & Number=plural(p) & Case ∈ { nominative(n), accusative(a), vocative(v) }

This corresponds to situations where the animateness has influence on the morphology and/or syntax. Although the animateness could be easily (with only little homonymy) assigned to all the masculine nouns, we opted for the described, rather complicated schema in order to be consistent with other MTE languages.

Pavol = Npmsn--y
Žiar = Npmsn--n
pes = Ncmsn
psa = Ncmsa--y
psov = Ncmpg (genitive)
psov = Ncmpa--y (accusative animate, homonymous with the genitive)
psi = Ncmpa--n (accusative inanimate, different from the animate)
žena = Ncfsn
ženu = Ncfsa

Verb

Almost all verb forms and the related grammatical meanings that existed in Old-Bulgarian have been preserved in the contemporary Bulgarian language. Unlike Bulgarian, the other Slavic languages have considerably simplified their old verb systems. The most characteristic peculiarity of Bulgarian is its very well developed system for expressing the grammar category of tense – there are forms for nine distinct verb tenses. Another important feature of the Bulgarian verb system is the presence of mood (so-called *inferential or re-narrative* mood) for the expression of non-witnessed modality or second-hand information. Bulgarian verbs have the grammatical categories person, number, voice, type, tense and mood. According to their lexical meaning, verbs can be transitive and intransitive. All these featured add to the complexity of the MTE tagset for Bulgarian verbs.

---

[1] It can be argued that some Slovak pluralia tantum do not follow this classification. However, in traditional grammars, a given word is always assigned (often arbitrarily and forcibly) its gender, to *make* the description fit.

Some examples:

чета	=	Vmia1s
чета	=	Vmip1s
пиша	=	Vmip1s
заминавам	=	Vmia1s
заминавам	=	Vmip1s

Both languages keep the so-called reflexive verbs. Reflexive verbs are formed from transitive verbs with the help of the personal reflexive pronoun ***sa, ce***, or from transitive and intransitive verbs with the personal reflexive pronoun ***si, cu***, for example: ***obliekať – obliekať sa, обличам – обличам се*** */dress – dress oneself/*; ***myslieť – myslieť si, мисля – мисля си*** */think – think by oneself/*. Reflexive verbs are not marked in the MTE tagset, reflexivity is shown only implicitly by the reflexive pronoun presence.

Bulgarian has only main(m) and auxiliary(a) vaues for the Type attribute, but again, Bulgarian verbs could be easily categorised in different ways (e.g. the Bulgarian ***(аз) мога*** (described as Type=main(m)) corresponds almost exactly with Slovak ***(ja) môžem*** (described as Type=modal(o)).

Slovak differentiates main(m), auxilliary(a), modal(o) and copula(c). However, this description is highly arbitrary and does not follow the traditional Slovak grammar description in detail, rather it was made for compatibility with the MTE tagset.

Vform=participle(p) corresponds to Slovak L-participle, in Bulgarian called just the participle and is used to form the past tense or the conditional. In Bulgarian, it also includes past participle (***говорено***) */spoken/*).

Vform=transgressive(t)[l.s]. in Slovak corresponds to VForm=gerund(g) in Bulgarian – this is just a difference in description.

In Slovak, imperative can be also present in the 1st person plural (***hovorme***), in Bulgarian the imperative would be formed analytically ((***хайде***) ***да говорим*** – (particle)+particle+verb).

In both Bulgarian and Slovak, the conditional is expressed roughly in the same way, by using a separate word ***би, by***, and the L-participle form (called just participle in Bulgarian). Slovak ***by*** is for the MTE purpose highly arbitrarily classified as a verb in conditional (Vform=conditional(c), the only such verb). No other grammar categories (person, gender, tense) are marked, purely for pragmatic reasons – to avoid the need of disambiguation. On the other hand, the Bulgarian ***би*** is classified as a full verb, Vform=active(a) (this is just a superficial difference in MTE tagset):

Slovak (lemma ***by***):

by =	Vcc

Bulgarian (lemma ***бъда***):

би =	Vaia2s
би =	Vaia3s
бих =	Vaia1s
биха =	Vaia3p
бихме =	Vaia1p
бихте =	Vaia2p

Verbs in participle form in Bulgarian can be classified for definiteness, Slovak verbs have no definiteness attribute.

In Bulgarian, there is a language specific Tense=aorist(a) value for the Tense attribute.

Past perfect tense "aorist" expresses a past action (event) carried out or completed in a given moment or during a given period and finished before the state of speaking.

Aorist is completely absent from Slovak.

In Slovak, voice attribute is always Voice=active(a), because passive voice occurs only in participles, which are categorised as adjectives. In Bulgarian, participles are classified as verbs, with Voice=passive(p) (past tense) or Voice=active(a) (present tense) types.

In Bulgarian, verbs can be negated with a special particle ***не*** written separately in front of the verb. In Slovak, verbs are negated by a prefix ***ne-***, which forms an unseparable part of the verb, and the lemma of a negative verb remains negative – this is more a feature of an orthography than an inherent difference in the languages. The only exception is the negation of the verb ***byť*** */to be/*, which is formed by a special particle ***nie*** written separately in front of the verb – this will be analysed as a particle, followed by a (positive) verb

lemmatised as ***byť***. In Slovak MTE, there is a Negative attribute, with (rather confusing) possible values Negative=no(n) for positive verbs and Negative=yes(y) for negative ones. Bulgarian does not have this attribute.

In Slovak, there is an Aspect attribute, which appeared in MTE in version 3. The Bulgarian tagset has been designed earlier and lacks the Aspect attribute, even if the aspect in Bulgarian is roughly the same as in Slovak (and other Slavic languages). The ambivalent aspect[l.s.] is present in a special class of verbs that have the same form in perfective and imperfective/progressive aspect (the difference is only semantic/syntactic, not morphological).

Adjectives

Slovak adjectives can have either qualificative or possessive Type.

Slovak adjectives have the degree attribute, while in Bulgarian degree is formed with a separate, auxiliary particles comparative ***по*** and superlative ***най***, written with a hyphen (***хубав***, ***по-хубав***, ***най-хубав***). This can be arguably considered just a matter of different orthography tradition, however, the Bulgarian description is justified by the adjective being always in the same form, regardless of the degree.

Gender, number and person are the same in Bulgarian and Slovak.

Slovak has a full case paradigm, while Bulgarian lacks cases (there is not even a separate vocative for the adjectives, and the attribute has empty value in MTE).

Bulgarian has definiteness.

Slovak has animateness, which is governed by the agreement between adjectives and nouns.

Pronouns

Classification of Bulgarian pronouns is according to their meaning – personal, possessive, reflexive, demonstrative, interrogative, relative, indefinite, negative and general. Bulgaran has Type=relative(r) (e.g. ***който***), which in Slovak would be formed by two consequent pronouns (***ten, ktorý***).

All the other values are compatible, there are only differences between specific classification of pronouns.

There are some traces of cases for Bulgarian pronouns, nominative(n), dative(d) and accusative(a) for personal pronouns, and their use depends on their syntactic function in the sentence – for example 1 p. sing.: ***аз*** (nom.), ***мене, ме*** (acc.), ***мене, ми*** (dat.), etc.

Slovak has full featured case paradigm for personal, adjectival and some other pronouns.

Owner_Number has the same function in Bulgarian and Slovak, however it is not described in the Bulgarian MTE (the type is left empty).

Although the Owner_Gender could be described for 3rd person possessive pronoun, both for Slovak and Bulgarian, both the Slovak and Bulgarian MTE description leave this type empty.

Clitic is the same for Bulgarian and Slovak.

Referent_Type is personal, possessive, attributive and quantitative in Bulgarian, but only personal and possessive in Slovak – the rest of pronouns do not have this type set (Referent_Type=-), which is just a deficiency in the Slovak MTE description. Otherwise the types are quite compatible between Bulgarian and Slovak.

Syntactic_Type in Slovak can be nominal(n) or adjectival(a) (e.g. ***ktorý, môj***), which is absent in the Bulgarian language (there are no adjectival pronouns of this type). Slovak also has several quasi-adjectival pronouns classified as Syntactic_Type=a (e.g. ***tvoj***), equivalents of which do exist in Bulgarian as well, but due to lack of the clear distinction of adjectival paradigm it was not felt unnecessary to introduce this value in Bulgarian MTE.

Bulgarian has definiteness, but it is present only for the possessive and reflexive types of pronouns, and for some general pronouns. Examples include:

Possessive:

Мой – моя - моят /my/

Твой - твоя – твоят /your, 2 p. sing/

Негов – неговия – неговият /his/

Reflexive:

Свой – своя – своят, своя – своята, свое-своето, свои - своите /his, her, its, their own/

Adverb

Bulgarian has language specific Type=adjectival(a), for words like ***умно*** /cleverly, wisely, sensibly/, which are derived from adjectives.

Slovak does not differentiate these two kinds of adverbs, but this is just a difference in description.

Slovak adverbs have the degree attribute, while in Bulgarian degree is formed with a separate, auxiliary particles ***по*** and ***най*** (see the discussion of degree for adjectives).

Adposition

Both languages have only prepositions, no postpositions.

Type is always preposition(p).

Slovak can contract some preposition with the following pronoun (***preň*** instead of ***pre neho***). These are described as Formation=compound(c).

Bulgarian has no compound prepositions.

Slovak tags for prepositions have the case attribute, which marks the case the preposition binds with.

Some Slovak prepositions can be vocalized, i.e. a vowel is appended to the preposition, if a following word starts with certain consonants (***v→vo, k→ku, s→so, z→zo, nad→nado***). This vocalization is not marked in the MTE tagset at all.

Conjunction

Type is the same in Slovak and Bulgarian – coordinating(c) or subordinating(s).

In Slovak, the class of two-part conjunctions has not been introduced, thus we ignore the Formation attribute.

In Bulgarian, Formation can be either simple(s) or compound(c).

Numeral

Slovak has Type cardinal(c), ordinal(o), multiple(m) and special(s), Bulgarian only cardinal(c) and ordinal(o).

In both Bulgarian and Slovak, the numerals are divided into two main categories: cardinal (quantitative) and ordinal (qualitative). Cardinal numerals signify a numerical (quantitative) property of objects: ***jeden dom, dve ženy, tri knihy; един дом, две жени, три книги*** /*one home, two women, three books*/. Ordinal (qualitative) numerals have an enumerating property, through which one can determine the consecutive position of an object in an ensemble of homogenous objects: ***prvý deň, druhý mesiac, tretia sekunda; първи ден, втори месец, трета секунда*** /*first day, second month, third second*/. Ordinal numerals cannot express degrees of comparison[1], but in Bulgarian they can accept an article (definiteness is the same in Bulgarian as for nouns). The two categories of numerals are distinguished not only by meaning, but also grammatical characteristics. Cardinal numerals do not have a grammatical gender (with the exception of ***jeden, jedna, jedno, dva, dve; един, една, едно, два, две,*** which were adjectives in Old Slavic) and do not change in number (with the exception of ***jeden, jedni, jedny; един, едни***), as they determine a given quantity. Ordinal numerals change gender and number just like adjectives. In Slovak, both cardinal and ordinal numerals keep morphological cases, and ordinal numerals are marked for animateness.

According to composition, numerals can be simple, complex or compound. Simple are single word numerals: ***jeden, dva, desať, sto; един, две, десет, сто*** /*one, two, ten, hundred*/, complex consist of several words fused together: ***jedenásť, dvanásť, päťsto; единадесет, дванадесет, петстотин*** /*eleven, twelve, five hundred*/, while compound ones are formed from two or more separate words – in Bulgarian, numerals connected with the conjunction ***и***, like ***двадесет и пет, хиляда и двеста*** /*twenty five, one thousand two hundred*/, in Slovak whenever the constituents are declinable (mostly ordinals bigger than 20) – ***dvadsiaty prvý, stoosemdesiaty druhý*** /21st, 182nd/ . In MTE tagset, this distinction is not described, and compound numerals are analysed as a sequence of several separate numerals (sometimes with the conjunction ***и***).

Example:

един	=	Mcms-ln
сто	=	Mc-p-ln
единадесет	=	Mc-p-ln
единадесети	=	Moms—ln
jeden	=	Mcmsnl--1
sto	=	Mcnpnl--f
jedenásť	=	Mcnpnl--f
jedenásty	=	Momsnl--fy

For cardinals, a number is singular only for the number 1 (***jeden, един***) and ratios.

---

[1] Neverthless, in Slovak there exist comparative and superlative degrees formed from the numeral ***prvý*** */the first/* – ***prvší, najprvší***. In Bulgarian only the form ***най-първи*** is used in colloquial speech.

Ratios in both Slovak and Bulgarian are compound – they are composed of two numerals: ***jedna štvrtina, една четвърт*** /*a quarter*/, ***tri desatiny, три десети*** /*three tenths*/. In Slovak, when the numerator equals „one“, it can be optionally left out. In Bulgarian, when the numerator is one “***единица***”, the numeral is formed using suffixes: ***-ин-а*** (***половина*** /*one half*/), ***-тин-а*** (***третина*** /*one third*/). In Slovak, both numerator and denominator are analysed as two separate numerals, while in Bulgarian they are analysed as one token:

една-четвърт = Mcfs-ln

една-пета = Mcfs-ln

In Slovak MTE, the Form attribute can be one of digit(d), roman(r), letter(l),

Bulgarian has an additional Form=m_form(m), used only for people, formed with suffix ***-(и)ма: двама, трима, петима*** /*two(people), three(people), five(people)*/ and Form=approx(a), used for approximate numerals (***десетина*** /*about a ten*/, ***стотина*** /*about a hundred*/):

десетина = Mc-p-an

стотина = Mc-p-an

Nouns derived from cardinal numerals with the suffixes ***-ina, -ica, -(or)ka, -ojka, -ица, -(ор)ка, -ойка*** will be classified as regular nouns – ***единица*** /*a one*/, ***stovka, стотица*** /*a hundred*/, ***sedmica, седморка*** /*a seven*/, ***osmica, осмица*** /*an eight*/.

единица = Ncfs-n

стотица = Ncfs-n

десетка = Ncfs-n

jednotka = Ncfsn

stovka = Ncfsn

desiatka = Ncfsn

Bulgarian has no Class attribute. Slovak has possible values according to the cardinality of the number, definite1(1) for “one”, definite2(2) for “two”, definite34(3) for “three” or “four”, definite(f) for “five or more”, demonstrative(d) (***toľko***/*that many*/), indefinite(i) (***niekoľko***/*several*/), interrogative(q)(***koľko***/*how many*/). Definite1, definite2, definite34 and definite are separated according to syntactical structures the numerals impose on the governed nouns – definite1 requires the corresponding noun to be in nominative singular, definite2 in nominative plural, definite34 nominative plural, definite genitive plural.

Bulgarian equivalents of demonstrative, indefinite, interrogative are classified as pronouns of a respective Type (including relative), e.g. ***няколко ученика*** /*a few students*/ – indefinite pronoun + noun. or sometimes as adverbs.

Interjection

Bulgarian has Formation=simple(s) or Formation=compound(c). Compound are those consisting of two (or more) words: ***боже мой!, има-няма, къде-къде, хайде де, кой знае, дявол да го вземе***. Note that some of them are written with a hyphen, but some with a space, and it is the task of the tokenizer to prepare the correct tokens.

In Slovak, corresponding interjections are mostly written together (***ktovie, dočerta, čerthovie***), but sometimes separately or with a hyphen (***dovidenia***, but also ***do videnia, bum bác*** but also ***bum-bác***), and these are tokenized as several separate words and analysed as either several interjections or as a residual + interjection.

Residual

In Slovak, special 'adverb prepositions' (***po, na, do***), encountered in expressions like ***po anglicky, na zeleno, do modra*** are classified as residuals. Traditional Slovak grammars do not like to consider them separate words, but rather see them to be different part-of-speech, mostly an adverb (see interjections above), with a space inside. In corresponding Bulgarian expressions (e.g. ***на български***), the residual will be classified as ***Sp*** (preposition). This is however just a difference in grammar description, not an inherent difference in the languages.

Abbreviation

In Slovak, trailing full stop is considered to be a separate token (punctuation character). In Bulgarian, the full stop is part of the abbreviation. Otherwise the descriptions in both languages are identical.

Particle

In the Bulgarian MTE tagset, particles are characterised by the Type attribute. Type attribute is one of negative, general, comparative, verbal, interrogative, modal.

Type=negative(z) is used for particles expressing negation (***не, ни, нито***)

Type=verbal(v) is used to form different type of verbal syntactical relationships, e.g. to create future tense (***ще*** *говориш*), or particles like ***се, да*** – Slovak uses very different verbal syntactical structures.

Type=interrogative(i) are particles used to form yes/no-questions or exclamations (***ли, дали, нали, нима, мигар***) – this type of particles is not present in Slovak at all.

Type=comparative(c) is for particles used to create comparatives or superlatives (***по, най***) – Slovak comparatives are formed through a morphology suffix, ***naj-*** is written together with superlatives. (this could be considered just a difference in orthography).

Type=modal(o) – used to express urge or order, mostly homonymous with other types of particles, for instance ***да, дано, нека, хайде***.

Type=general(g) is for all the other, non-specialised particles.

The Formation attribute can be either simple(s) (single word particles) or compound(c) (multiple word particles, e.g. ***хайде де***).

In the Slovak MTE tagset, we simplified our task enormously by resigning the classification attempts (which can be analysed ad nauseam to an arbitrary precision (Šimková, 2004)), and all the articles have the same simple tag **P**. The classification has no morphology effect anyway.

Concluding Remarks and Recommendations

Multext East morphological tagset attempts to describe the morphology of several languages using the same principles and the same set of tags. Ideally, the differences in the respective tagsets reflex inherent underlying differences in the languages. Our analysis show that at least between Bulgarian and Slovak, there are many differences due to different way of analysing morphology in traditional grammars, as well as different Multext East tags assigned to the same categories across languages. However, we have successfully analysed the differences and pointed out categories and attributes where the discrepancies occur. In any comparative analysis of the languages based on the Multext East morphology annotation, it is necessary to take these results into account, to reveal superficial differences not based on real dissimilarities of the languages' grammars in question.

The Multext East tagset is suitable for Slavic languages. We recommend MTE morphology tagset for annotation of corpora (parallel or comparable), either as a sole morphology tagset or in addition to an established one. However, special care needs to be taken when analysing morphology across languages, because the Multext East tagset differences are sometimes artificial, based on different grammar description, not on real differences between the languages. There are also morphology and syntax categories that the Multext East tagset does not map the same way between the languages, and therefore cannot be used uncritically in cross-linguistic analysis.

References

1. Dimitrova, 1998: Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.-J., Petkevic, V., and Tufis, D. 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.

2. Garabík et al, 2004: Garabík, R. Gianitsová, L., Horák, A., Šimková, M., Šmotlák, M.: Slovak National Corpus. In: Proceedings of the conference TSD 2004. Brno, Czech Republic: Springer-Verlag 2004.

3. Garabík, 2006: Garabík, Radovan: Slovak morphology analyzer based on Levenshtein edit operations. Proceedings of the WIKT'06 conference, p. 2–5. Bratislava, Slovakia, 2006.

4. Ide, Véronis, 1994: Ide, N., and Véronis, J.: Multext (multilingual tools and corpora). In COLING'94, p. 90–96, Kyoto, Japan, 1994.

5. Ružička, 1966: Morfológia slovenského jazyka. Ed. J. Ružička. Bratislava: Vydavateľstvo Slovenskej akadémie vied 1966.

6. MTE, 2004: MULTEXT-East Morphosyntactic Specifications – version 3, edition 10th May 2004.

7. Šimková, 2004: Šimková, Mária: Funkcie častíc v komunikácii. In: Jazyk v komunikácii. Medzinárodný zborník venovaný Jánovi Bosákovi. Ed. S. Mislovičová, p. 168 – 176. Bratislava: Veda 2004.

8. Bulgarian Grammar, 1993: Главна редакция Д. Тилков, Ст. Стоянов, К. Попов. Граматика на съвременния български книжовен език. Том 2 / МОРФОЛОГИЯ. Издателство на БАН. София. (In Bulgarian).

Annotation of Parallel Corpora (on the Example of the Bulgarian–Polish Parallel Corpus)[1]

Ludmila Dimitrova[1], *Violetta Koseska-Toszewa*[2], *Ivan Derzhanski*[1], *Roman Roszko*[2],
[1] IMI-BAS, *ludmila@cc.bas.bg*, [2]ISS-PAS

Abstract
In this paper we briefly describe a comparison of the morphosyntactic characteristics of the words of the first Bulgarian-Polish parallel corpus from the point of view of a prospective unification.

Keywords: *Bulgarian, Polish, parallel corpus, corpus annotation, morphosyntactic description, POS tagging*

1. Introduction

Corpus linguistics is a dynamic field which boasts many accomplishments in recent years. Among them are the MULTEXT corpus (Ide, Veronis, 1994), initially in seven West European languages (Dutch, English, French, German, Italian, Spanish and Swedish, with more in later editions, including Bambara, Catalan, Kikongo, Occitan and Swahili), and the MULTEXT-East annotated parallel corpus (Dimitrova et al., 1998), initially in six East European languages (Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovenian, plus English as a "hub" language, in later editions including Croatian, Lithuanian, Resian[2], Russian and Serbian). MULTEXT-East is an extension of the language engineering project MULTEXT, one of the largest EU projects in the domain of language tools and resources.

The first Bulgarian–Polish corpus (currently under development in the framework of the joint research project "Semantics and Contrastive linguistics with a focus on a bilingual electronic dictionary" between IMI—BAS and ISS—PAS, coordinated by L. Dimitrova and V. Koseska) contains a total of approx. 3 million words and comprises two corpora: parallel and comparable (Dimitrova, Koseska, 2007, 2008). The first Bulgarian–Polish parallel corpus contains more than 1 million words, mostly fiction (a small part comprises official documents of the European Commission available through the Internet). The corpus is composed of two parts: original Bulgarian texts with Polish translations or vice versa and texts in other languages translated into both Bulgarian and Polish. The comparable corpus includes texts in Bulgarian and Polish, excerpts from newspapers, literary works, Internet textual documents, with the text sizes being comparable across the two languages. Some of the texts have been annotated at paragraph level. The bilingual Bulgarian–Polish corpus will be annotated according to the digital language resource annotation standards and will provide a sample of the vocabulary, which is to be included in an initial experimental version of the Bulgarian–Polish digital dictionary.

We endeavoured to perform a comparison of the morphosyntactic characteristics of the words of parallel texts in the two languages from the point of view of a prospective unification.

[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] Resian is a distinct dialect of Slovenian spoken in the valley Resia in Italy, close to the border with Slovenia. Resian and standard Slovenian are mutually unintelligible due to archaisms not preserved in modern Slovenian and significant Italian influence on Resian pronunciation and vocabulary, as well as Italian-induced innovations in Resian grammar (including prepositive definite and indefinite articles).

2. Corpus annotation

Corpus annotation is the process of adding linguistic information in an electronic form to a text corpus (Ide et al. 2000, Leech 2004, Monachini, Calzolari, 1996). Among the most common and important types of corpus annotation are **morphosyntactic annotation** (also called **grammatical tagging** or *part of speech* (*POS*) *tagging*), whereby a label or tag is associated with each word token in the text in order to indicate its grammatical classification, and *lemma annotation*, where the lemma of each word-token is indicated in the text. These two types may be regarded as mutually complementary.

POS tagging is the task of labelling each word in a sequence of words with its appropriate part-of-speech. Words are often ambiguous with respect to their POS; for example, in Bulgarian the neuter singular forms of most adjectives serve double duty as adverbs.

вероятно 'probable (neuter), probably'

вероятно → POS: adjective, Gender: neuter, Number: singular,

Definiteness: no

вероятно → POS: adverb, Type: adjectival

A tagset is a set of part-of-speech tags. The size and choice of the tagsets vary across languages. The classical system is based on a set of parts of speech including noun, verb, adjective, pronoun, adverb, numeral, preposition, conjunction, particle, interjection, and often (depending on the language) article, participle, etc. Morphologically rich languages need more detailed tagsets reflecting various inflexional categories.

The applications of POS tagging include lexicography, parsing, language models in speech recognition, disambiguation clues for ambiguous words (machine translation), information retrieval, spelling correction, etc.

3. Morphosyntactic descriptions for Bulgarian

For the purposes of morpho-lexical processing of corpora, the MULTEXT-East consortium developed language-specific word-form lexical lists covering at least the words appearing in this corpus. For each of the six MTE languages, a lexical list containing at least 15,000 lemmata were developed for use with the morphological analyser. Each lexicon entry includes information about the inflected-form, lemma, POS, and morphological specifications. A mapping from the morphosyntactic information contained in the lexicon to a set of corpus tags (used by the part-of-speech disambiguator) was also provided, according to the MULTEXT tagging model.

A lexicon entry has the following structure:

word-form ‹TAB› **lemma** ‹TAB› MSD ‹TAB› **comments**

where word-form represents an inflected form of the lemma, characterised by a combination of feature values encoded by *MSD*-code (**MSD**: **M**orpho**S**yntactic **D**escription); the fourth (optional) column, comments, is currently ignored and may contain either comments or information processable by other tools.

Here is an excerpt from the Bulgarian Lexicon:

Word-Form	*Lemma*	*MSD*
вещества	вещество	Ncnp-n
веществата	вещество	Ncnp-y
вещество	=	Ncns-n
веществото	вещество	Ncns-y

(*вещество* 'substance')

The **MSDs** are provided as strings, using a linear encoding; a relatively efficient and compact way to represent the flat attribute-value matrices. In this notation, the position in a string of characters corresponds to an attribute, and specific characters in each position indicate the value for the corresponding attribute. That is, the positions in a string of characters are numbered 0, 1, 2, etc., and are used in the following way: the character at position 0 encodes part-of-speech; each character at position 1, 2, …, n, encodes the value of one attribute (person, gender, number, etc.), using the one-character code; if an attribute does not apply, the corresponding position in the string contains the special marker "-" (hyphen). By convention, trailing hyphens are not included in the lexical MSDs. Such specifications provide a simple and relatively compact encoding, and are similar to feature-structure encoding used in unification-based grammar formalisms. When the word form is the very lemma, then the equal sign is written in the lemma field of the entry ("=").

For Bulgarian morphosyntactic annotation was implemented in 1996–1997 for the purposes of the MULTEXT-East project. The morphosyntactic descriptions were designed on the basis of the traditional part-of-speech classification (Bulgarian Grammar 1993). Each word form is assigned a label encoding the major category (part of speech), type where applicable (e.g., proper versus common noun) and inflexional features. Punctuation is also included, as are abbreviations, numbers written in digits, and unidentified objects (residuals). A further non-standard category contains markers of degrees of comparison. Those are formed in Bulgarian with the particles *по* (comparative) and *най* (superlative), preposed to the adjective or adverb but separated from it by a hyphen (*лек* 'light', *по-лек* 'lighter', *най-лек* 'lightest'; *леко* 'lightly', *по-леко* 'more lightly', *най-леко* 'most lightly'). These particles are annotated as separate words:

по → POS: Particle, Type: comparative, Formation: simple,

най → POS: Particle, Type: superlative, Formation: simple.

4. The morphosyntactic descriptions for Polish

For Polish morphosyntactic analysis is performed by Marcin Woliński's Morfeusz (Woliński 2003, 2006). This analyser is based on an extended set of parts of speech (15 in total), so that groups of words traditionally counted under the same part of speech (and even parts of the same paradigm) are separated if they differ significantly in their inflexional categories or syntactic meaning. It constructs all possible analyses of each word and singles out one analysis as the most likely one (a suggestion that the user is free to endorse or decline).

The analyser has the shortcoming that it does not cover the bound clitic forms of the copula *-(e)m* 'I am', *-(e)ś* 'you are', etc., except when the host of the clitic is a past tense verb form (giving regular past tense forms inflected for person and number, though represented as two-word sequences here, which shows that the treatment of groups of words unseparated by blank space or punctuation is not a problem in principle). For example, the sentence *Coś zrobił?* is only analysed as 'Did he do something?' (*coś* 'something, anything' + *zrobił* '[he] did'), missing the alternative meaning 'What did you do?' (*co* 'what' + *-(e)ś* 'you' + *zrobił* 'did'), also expressible as *Co zrobiłeś?* (*co* 'what' + *zrobił* 'did' + *-(e)ś* 'you', analysed correctly by Morfeusz).

One further possibly questionable point is the treatment of gender. The category of animacy is unusually ramified in Polish, so that three varieties of the masculine gender are counted (human, animal and inanimate[1]). The analyser treats these as three separate genders (of a total of five in the language, according to Saloni's simplified version[2]), which gives rise to a proliferation of possible analyses due to the massive syncretism of gender in all parts of speech inflecting for this category (adjectives, numerals, pronouns, quasiparticiples and participles).

For example, the verb form (here called a quasiparticiple) *był* '(he) was' is assigned three genders (the three masculines); *były* '(they) were' is assigned four (masculine animal, masculine inanimate, feminine and neuter). Personal pronouns for the first and second persons are also assigned all possible genders, the most likely one being chosen on semantic grounds if possible. Thus the pronoun *ja* 'I' in *Ja przyszedłem* 'I came (m.)' is proclaimed most likely masculine human, although the other four analyses are also generated, because the verb form is in the masculine; in *Ja idę* 'I go' the same pronoun, now with a gender-neutral verb, is labelled as most likely feminine, perhaps because it happens to end in *-a*. This adds to the complexity of the analysis.

5. The experiment

We took two short stories for children ('The Gluttonous Little Bear' by Emilian Stanev and 'Soap Bubbles' by Svetoslav Minkov) in the original Bulgarian and in V. Koseska-Toszewa's Polish translation (just under 1000 words).

The translation is literary rather than literal. Some frequently recurring differences are due to the stylistic preferences characteristic of the two languages: Polish makes active use of constructions with participles and gerunds, which Bulgarian also possesses but employs significantly less, especially in informal speech and writing, preferring constructions with finite verb forms. Of course this is just a general tendency, and in individual sentences the correspondences may be of any complexity:

'"Great that I broke off from that vulgar straw!" said the first soap bubble, flushed with joy and floated above the bed of daisies.'

[1] Words denoting animals behave as human in the singular number and as inanimate in the plural.

[2] Different accounts distinguish between three and nine (or more) genders in Polish.

Bulgarian:

– Добре, че се откъснах от тая проста сламка — рече първият, като почервеня от радост и се понесе над лехата с маргаритките.

(*flushed and floated are coordinated and both are subordinated to said*)

Polish:

– Doskonale! oderwałam się od tej brzydkiej słomki! — odezwała się pierwsza bańka i zarumieniona z dumy i radości poleciała nad grządkę ze stokrotkami.

(*flushed is subordinated to floated, which is coordinated with said*)

We ran the tagger on the Polish and Bulgarian texts. Then we compared the tags.

The result of the automatic disambiguation of the first sentence of 'Soap Bubbles' by Svetoslav Minkov

Имаше едно малко момиченце с червена панделка на косата.

Była sobie raz dziewczynka z piękną czerwoną wstążką we włosach.

'There was once a little girl with a red ribbon in her hair.'

can be found in the Appendix.

The table below shows the tags assigned to the words; where there are two or more possible analyses, the one which is actually chosen is shown first.

Bulgarian	Bulgarian MSDs	Bulgarian ctags	Polish	Polish ctags
имаше	Vmii3s Vmii2s	VMII3S VMII2S	*była*	adj:sg:nom:f:pos praet:sg:f:imperf
			sobie	siebie:dat siebie:loc
			raz	subst:sg:nom:m3 subst:sg:acc:m3
едно	Mcns-ln	MC		
малко	A--ns-n Ra Ncns-n	ANS RA NCNS-N		
момиченце	Ncns-n	NCNS-N	*dziewczynka*	subst:sg:nom:f
с	Sp	SP	*z*	prep:gen:nwok prep:inst:nwok qub
			piękną	adj:sg:acc:f:pos adj:sg:inst:f:pos
червена	A--fs-n Vmps-sfp-n	AFS VMPS-SF	*czerwoną*	adj:sg:acc:f:pos adj:sg:inst:f:pos
панделка	Ncfs-n	NCFS-N	*wstążką*	subst:sg:inst:f
на	Sp Qgs	SP QG	*we*	prep:loc:wok prep:acc:wok
косата	Ncfs-y	NCFS-Y	*włosach*	subst:pl:loc:m3
.		PERIOD	.	interp

The Bulgarian tags stand for, as follows:

AFS	Adjective feminine singular
ANS	Adjective neutral singular
MC	numeral cardinal
NCFS-N	noun common feminine singular indefinite
NCFS-Y	noun common feminine singular definite
NCNS-N	noun common neuter singular indefinite
PERIOD	Period
QG	particle general
RA	adverb adjectival
SP	Adposition prepositive
VMII2S	verb main indicative imperfect 2nd singular
VMII3S	verb main indicative imperfect 3rd singular
VMPS-SF	verb main participle past singular feminine

The Polish tags stand for, as follows:

adj:sg:acc:f:pos	adjective : singular : accusative : feminine : positive
adj:sg:inst:f:pos	adjective : singular : instrumental : feminine : positive
adj:sg:nom:f:pos	adjective : singular : nominative : feminine : positive
interp	punctuation
praet:sg:f:imperf	quasiparticiple : singular : feminine : imperfective
prep:acc:wok	preposition : accusative : vocalised
prep:gen:nwok	preposition : genitive : unvocalised
prep:inst:nwok	preposition : instrumental : unvocalised
prep:loc:wok	preposition : locative : vocalised
qub	qublik (particle-adverb)
siebie:dat	siebie : dative
siebie:loc	siebie : locative
subst:pl:loc:m3	noun : plural : locative : masculine (inanimate)
subst:sg:acc:m3	noun : singular : accusative : masculine (inanimate)
subst:sg:inst:f	noun : singular : instrumental : feminine
subst:sg:nom:f	noun : singular : nominative : feminine
subst:sg:nom:m3	noun : singular : nominative : masculine (inanimate)

Regarding the tagsets, the main differences between them (ignoring the mismatches in the names of matching tags, which can be amended easily) are due to the different morphological makeup of the two languages: Polish has morphological case pattern for all nominal parts of speech (seven cases) which Bulgarian has almost entirely lost (with the exception of a vestigial vocative in the noun and rudimentary declension of the personal pronoun); by contrast, Bulgarian has a definite article which was originally an enclitic but has merged with the noun, adjective or numeral into a single word form, giving an inflexional category of definiteness. For example:

Bulgarian

косата коса Ncfs-y

[wordform *косата* 'the hair', lemma *коса* 'hair'] POS: Noun, Type: common, Gender: feminine, Number: singular, Definiteness: yes;

Polish

włosach włos subst:pl:loc:m3

[wordform *włosach*, lemma *włos* '(strand of) hair'] POS: substantive (noun), Number: plural, Case: locative, Gender: masculine 3 (inanimate).

Bulgarian has also preserved more of the verb conjugation of Old Slavic, whereas in Polish verb conjugation is relatively simple, especially if the floating cliticised copula (a Polish innovation) is considered a separate word, as in this analyser (e.g., *przyszedłem* 'I came' is analysed as *przyszedł* 'came' + *-em* 'I').

The number of MSDs was reduced with respect to the c-tags in Bulgarian (from 324 MSDs, used in Bulgarian MTE corpus, to 117 c-tags to run the ISSCO tagger) due to software limitations 15 years ago. For example, instead of the three MSDs for masculine singular forms of adjectives (A--ms-f with full article, A--ms-s with short article, A--ms-n with none) the single c-tag AMS was used. The five MSDs for the demonstrative pronoun (Pd-------q, Pd--p----p, Pd-fs----p, Pd-ms----p, Pd-ns----p) were collapsed to the c-tag PD; for the 15 MSDs for relative pronoun (Pr-------q, Pr--p----a, Pr--p----p, Pr--p----s, Pr-fs----a, Pr-fs----p, Pr-fs----s, Pr-ms----a, Pr-ms----s, Pr-msa---p, Pr-msd---p, Pr-msn---p, Pr-ns----a, Pr-ns----p, Pr-ns----s) the c-tag PR is used. Due to increased computing power today we think that such a reduction is no longer necessary, so that the morphosyntactic descriptions are fully preserved at POS annotation.

We do not consider the different nomenclature of POS tags in Polish and Bulgarian to be a significant problem because a one-to-one correspondence could easily resolve it.

It is interesting to compare the set of tags for Polish used by Morfeusz to MTE's set of tags for synthetic Slavic languages (Czech, Slovak, Slovene), whose grammatical categories are closer to those of Polish. Some of the differences are obvious (e.g., Polish has no dual number, whereas Czech has a vestigial dual and Slovene a full-fledged one). The following table summarises the less trivial differences between the approaches:

	MTE (Czech, Slovak, Slovene)	Morfeusz (Polish)
noun class	split into orthogonal categories of gender (m, f, n) and animacy (no, yes)	treated as a single multi-value category of gender
cliticised copula	treated as a feature of the host word	treated as a separate word
past tense	treated as a value of the category of tense (past)	treated as a compound of a quasi-participle and a cliticised copula
imperfective present vs perfective future	distinguished as different tenses	distinguished only by aspect
prepositional form of 3rd person pronoun	disregarded	labelled by means of an express category as postprepositional

Conclusion

Our scrutiny of the outcome of the tagging of the short texts leads us to believe that a unification of the morphosyntactic annotation for Bulgarian and Polish should be done within the perspective of the elaboration of a general tagset for Slavic languages.

References

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

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

3. Dimitrova, Koseska, 2008: Dimitrova, L., V. Koseska–Toszewa. Some Problems in Multilingual Digital Dictionaries. In: International Journal *ÉTUDES COGNITIVES*. Vol. 8. SOW, Warsaw. 2008. Pages 237-254. ISSN 1641-9758.

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

5. Ide et al. 2000: Ide, N., Bonhomme, P., and Romary, L. XCES: An XMLbased Encoding Standard for Linguistic Corpora. Proceedings of the Second International Language Resources and Evaluation Conference. Paris: European Language Resources Association, 2000. 825-830.

6. Leech 2004: Geoffrey Leech. Developing Linguistic Corpora: a Guide to Good Practice Adding Linguistic Annotation. 2004.

7. http://ahds.ac.uk/guides/linguistic-corpora/chapter2.htm

8. Monachini, Calzolari, 1996: Monachini, M. and Calzolari, N. Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora: A Common Proposal and Applications to European Languages. EAGLES Report EAG-CLWG-MORPHSYN/R. 1996. http://www.ilc.pi.cnr.it/EAGLES96/morphsyn/

9. Piasecki, 2007: Piasecki M., Polish Tagger TaKIPI: Rule Based Construction and Optimisation. Task Quarterly. 2007, 11, p. 151-167.

10. Piasecki, Godlewski, 2006: Piasecki, M., Godlewski, G.. Reductionistic, Tree and Rule Based Tagger for Polish. In: Kłopotek, M. A., Wierzchoń, S. T., i Trojanowski, K., red. (2006). Intelligent Information Processing and Web Mining — Proceedings of the International IIS: IIPWM'06 Conference held in Zakopane, Poland, June, 2006. Advances in Soft Computing. Springer, Berlin.

11. Piasecki, Wardyński, 2006: Piasecki, M., Wardyński, A., Multiclassifier Approach to Tagging of Polish. In: Proceedings of 1st International Symposium Advances in Artificial Intelligence and Applications. 2006.

12. Woliński, 2006: Woliński, M., Analizator morfologiczny *Morfeusz SIAT*. (In Polish) http://nlp.ipipan.waw.pl/~wolinski/morfeusz/

13. Woliński, 2003: Woliński, M., System znaczników morfosyntaktycznych w korpusie IPI PAN. Polonica XII, 2003, p. 39-55. (In Polish)

14. Bulgarian Grammar, 1993: Главна редакция Д. Тилков, Ст. Стоянов, К. Попов. Граматика на съвременния български книжовен език. Том 2 / МОРФОЛОГИЯ. Издателство на БАН. София. (In Bulgarian).

ISSCO tagger: http://www.issco.unige.ch/staff/robert/tatoo/tagger.html#design

Appendix

The first sentence of 'Soap Bubbles' by Svetoslav Minkov is:

Имаше едно малко момиченце с червена панделка на косата.

'There was a little girl with a red ribbon in her hair.'

(1) in Bulgarian (ISSCO tagger)

```
<tok type=WORD>
      <orth>Имаше</orth>
      <disamb><base>имам</base><ctag>VMII3S</ctag></disamb>
      <lex><base>имам</base><msd>Vmii2s</msd><ctag>VMII2S</ctag></lex>
      <lex><base>имам</base><msd>Vmii3s</msd><ctag>VMII3S</ctag></lex>
    </tok>
<tok type=WORD>
      <orth>едно</orth>
      <disamb><base>едно</base><ctag>MC</ctag></disamb>
      <lex><base>едно</base><msd>Mcns-ln</msd><ctag>MC</ctag></lex>
    </tok>
<tok type=WORD>
<orth>малко</orth>
   <disamb><base>малък</base><ctag>ANS</ctag></disamb>
   <lex><base>малък</base><msd>A--ns-n</msd><ctag>ANS</ctag></lex>
   <lex><base>малко</base><msd>Ra</msd><ctag>RA</ctag></lex>
   <lex><base>малко</base><msd>Ncns-n</msd><ctag>NCNS-N</ctag></lex>
   </tok>
<tok type=WORD>
      <orth>момиченце</orth>
      <disamb><base>момиченце</base><ctag>NCNS-N</ctag></disamb>
      <lex><base>момиченце</base><msd>Ncns-n</msd><ctag>NCNS-N</ctag></lex>
    </tok>
<tok type=WORD>
      <orth>с</orth>
      <disamb><base>с</base><ctag>SP</ctag></disamb>
      <lex><base>с</base><msd>Sp</msd><ctag>SP</ctag></lex>
    </tok>
<tok type=WORD>
    <orth>червена</orth>
    <disamb><base>червен</base><ctag>AFS</ctag></disamb>
    <lex><base>червен</base><msd>A--fs-n</msd><ctag>AFS</ctag></lex>
    <lex><base>червя</base><msd>Vmps-sfp-n</msd><ctag>VMPS-SF</ctag></lex>
     </tok>
<tok type=WORD>
      <orth>панделка</orth>
      <disamb><base>панделка</base><ctag>NCFS-N</ctag></disamb>
      <lex><base>панделка</base><msd>Ncfs-n</msd><ctag>NCFS-N</ctag></lex>
    </tok>
<tok type=WORD>
      <orth>на</orth>
      <disamb><base>на</base><ctag>SP</ctag></disamb>
      <lex><base>на</base><msd>Qgs</msd><ctag>QG</ctag></lex>
      <lex><base>на</base><msd>Sp</msd><ctag>SP</ctag></lex>
    </tok>
<tok type=WORD>
      <orth>косата</orth>
      <disamb><base>коса</base><ctag>NCFS-Y</ctag></disamb>
      <lex><base>коса</base><msd>Ncfs-y</msd><ctag>NCFS-Y</ctag></lex>
    </tok>
<tok type=PUNCT>
     <orth>.</orth>
     <ctag>PERIOD</ctag>
    </tok>
```

(2) in Polish (Piasecki, 2007)

Była sobie raz dziewczynka z piękną czerwoną wstążką we włosach.

```
<tok>
    <orth>Była</orth>
    <lex><base>były</base><ctag>adj:sg:nom:f:pos</ctag></lex>
    <lex disamb="1"><base>być</base><ctag>praet:sg:f:imperf</ctag></lex>
</tok>
<tok>
    <orth>sobie</orth>
    <lex disamb="1"><base>się</base><ctag>siebie:dat</ctag></lex>
    <lex><base>się</base><ctag>siebie:loc</ctag></lex>
</tok>
<tok>
    <orth>raz</orth>
    <lex><base>raz</base><ctag>subst:sg:nom:m3</ctag></lex>
    <lex disamb="1"><base>raz</base><ctag>subst:sg:acc:m3</ctag></lex>
</tok>
<tok>
    <orth>dziewczynka</orth>
    <lex disamb="1"><base>dziewczynka</base><ctag>subst:sg:nom:f</ctag></lex>
</tok>
<tok>
    <orth>z</orth>
    <lex><base>z</base><ctag>prep:gen:nwok</ctag></lex>
    <lex disamb="1"><base>z</base><ctag>prep:inst:nwok</ctag></lex>
    <lex><base>z</base><ctag>qub</ctag></lex>
</tok>
<tok>
    <orth>piękną</orth>
    <lex><base>piękny</base><ctag>adj:sg:acc:f:pos</ctag></lex>
    <lex disamb="1"><base>piękny</base><ctag>adj:sg:inst:f:pos</ctag></lex>
</tok>
<tok>
    <orth>czerwoną</orth>
    <lex><base>czerwony</base><ctag>adj:sg:acc:f:pos</ctag></lex>
    <lex disamb="1"><base>czerwony</base><ctag>adj:sg:inst:f:pos</ctag></lex>
</tok>
<tok>
    <orth>wstążką</orth>
    <lex disamb="1"><base>wstążka</base><ctag>subst:sg:inst:f</ctag></lex>
</tok>
<tok>
    <orth>we</orth>
    <lex disamb="1"><base>w</base><ctag>prep:loc:wok</ctag></lex>
    <lex><base>w</base><ctag>prep:acc:wok</ctag></lex>
</tok>
<tok>
    <orth>włosach</orth>
    <lex disamb="1"><base>włos</base><ctag>subst:pl:loc:m3</ctag></lex>
</tok>
<ns/>
<tok>
    <orth>.</orth>
    <lex disamb="1"><base>.</base><ctag>interp</ctag></lex>
</tok>
```

The current state of work on the Polish-Ukrainian Parallel Corpus (PolUKR)[1]

Natalia Kotsyba
Institute of Slavic Studies PAS (Warsaw)
gnatko@gmail.com

Abstract
The article describes the present state of work on PolUKR, the Polish-Ukrainian parallel corpus, developed in the Institute of Slavic Studies of the Polish Academy of Sciences since 2004. Presented are the ways of bitexts' acquisition, their structure and pre-processing stages; the solutions concerning the common morphosyntactic annotation pattern for Polish and Ukrainian, as well as annotation methods; the alignment format and the software used or developed for the corpus needs.

Keywords: *corpus linguistics, parallel corpus, bitexts, translation memory, morphosyntactic annotation, Polish, Ukrainian, Slavic languages, sentence aligner, electronic dictionary*

Objectives of creating the corpus

PolUKR, a Polish-Ukrainian parallel corpus was launched as a pilot corpus project in the Institute of Slavic Studies of the Polish Academy of Sciences in 2004. The corpus is intended for use as a tool for both human and machine users, as well as language material for compiling bilingual Polish<>Ukrainian dictionaries and a contrastive grammar for Polish and Ukrainian. It can also be used as a translation database and language learning materials.

Acquisition and preprocessing of parallel texts

Currently the corpus contains ab. 2 mln tokens (500K tokens in 70 parallel texts are publicly available for search through the web interface at http://corpus.domeczek.pl) that represent mostly modern Ukrainian and Polish literature (the 2nd part of the XXth century). Part of them was received from the translators[2], then the corresponding original was sought for and prepared accordingly. Another group was downloaded from existing digital libraries[3]. The quality of the texts was often unsatisfactory, as in most cases electronic texts were acquired through scanning the paper editions that were later submitted to the automatic Optical Character Recognition (OCR) procedure and needed further corrections. A large group of the texts was originally in the hard copy format, they were scanned, cleaned from images, page numbers and other unnecessary information, then OCRed with the help of the FineReader 9.0 program, checked for mistakes that appeared as a consequence of a poor OCR, recorded as MS Word documents and converted into simple UTF-8 encoded XML files that contain information about division into paragraphs extracted from DOC files with the help of the AutoReplace function.

The text metadata are recorded into a MySQL database placed on the server. They include (if available): author, title, language, year of creation, publication place, year and publishing house, genre, translator,

[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] We would like to thank Katarzyna Kotyńska, Anna Łazar, Ola Hnatiuk and Helena Krasowska for sharing their texts.
[3] Some of the libraries used can be found at: http://lib.ru, http://www.ae-lib.org.ua/, http://www.4shared.com/dir/3997557/7fe59813/ebooki.html, http://exlibris.org.ua/, http://ukrcenter.com/library/default.asp, http://www.share.net.ua/, http://lib.proza.com.ua.

year of translation, source and original format of the text, etc. This information may be used to restrict the scope of search, e.g., one can choose only the texts created after a specific date or by a specific author.

Structural annotation

The texts are segmented into chunks that can be of two types: paragraphs and sentences. Sentences are always parts of paragraphs. Such structure of the document is encoded in a corresponding Document Type Definition file.

Morphological annotation

We use the freely available TaKIPI toolset developed by Marcin Woliński, Adam Przepiórkowski, Adam Radziszewski and Maciej Piasecki, that includes a text chunker, a lemmatizer, a morphological tagger and a disambiguator, for adding the morphosyntactic information for the Polish texts.

Morphological tags are stored as value lists containing morphological class and grammatical categories adequate for a given class, e.g., the grammatical characteristics of *jedziecie* (*you*$_{\text{pl}}$ *go*) will be *fin:pl:sec:imperf* (finite verb form, plural, second person, imperfective aspect). If an ambiguity occurs for a given segment, several tags are listed. After the disambiguation procedure the most verisimilar "candidate" is given the disambiguation value "1".

An example of a tagged chunk "Dokąd jedziecie?"

```
<chunk type="p" xlink:href="#p5">
<chunk type="s">
  <tok>
    <orth>dokąd</orth>
    <lex disamb="1">
     <base>dokąd</base>
     <ctag>qub</ctag>
    </lex>
  </tok>
  <tok>
    <orth>jedziecie</orth>
    <lex disamb="1">
      <base>jechać</base>
      <ctag>fin:pl:sec:imperf</ctag>
    </lex>
  </tok>
  <tok>
    <orth>?</orth>
    <lex disamb="1">
      <base>?</base>
      <ctag>interp</ctag>
    </lex>
  </tok>
</chunk>
</chunk>
```

For the Ukrainian language we use the UGS (Ukrainian Grammatical Dictionary) developed at the ULIF NASU by Igor Shevchenko and Oleksandr Rabulets, that enables lemmatization and morphological annotation of texts, although its does not support disambiguation at the moment.

A common morphosyntactic tagset for Polish and Ukrainian was developed by us for the corpus needs based on the mentioned resources, see [Kotsyba et al. 2008, Коциба 2009] for details. Language specific categories and values are preserved, as our intention was not to lose any information. All the details will not be seen at the GUI search-level, but will be accessible for advanced users through self-defined regex-based corpus queries. The basic changes we had to introduce include a higher POS granulation for Ukrainian and

regrouping some word classes for Polish to fit a more traditional understanding of the parts of speech. These quasi-changes are realized with the help of the mechanism of aliases and effect only the GUI search level. Reorganizing of information about the degree for Ukrainian adjectives and adverbs from the lexical to grammatical level has also been done to keep to the standards both in traditional grammars and current commonly accepted NLP treatment of the degree as a grammatical category. The special treatment of predicatives that was followed by us as well is described in detail in [Derzhanski, Kotsyba 2008].

The above format was also used for the Ukrainian language while converting the original annotated files.

Alignment

Presently the parallel texts are aligned at the paragraph level dynamically, i.e. paragraphs are enumerated during the searching procedure and those with the same order number that the ones where the searched fragment is found are shown along with the KWICs. The difference in the paragraph division had to be removed manually, so that their order numbers and content where equal. This situation is provisionary – the paragraph level of the alignment is unsatisfactory as most paragraphs are too lengthy to easily spot the searched equivalent. The intended alignment level are sentences and, eventually, words.

One of freely available programs that aligns parallel texts at the sentence level is the language independent HunAlign. The result of the alignment is recorded either as an intertwined text or as sets of corresponding sentences, so called link groups, represented by sentence numbers or other identifiers. Additional numeric information about the accuracy of alignment can be included as well. The program foresees the use of a corresponding bilingual dictionary to ensure a higher accuracy of alignment. Such a dictionary can also be generated by the program itself from the currently fed in bitexts, if not available otherwise. The results of aligning Polish and Ukrainian texts without a dictionary were far from satisfactory. For the purpose of a more accurate alignment we have developed a bilingual dictionary structured according to the HunAlign demands. It is recorded as a plain text where each entry takes a separate line: the original word or expression, @-sign, the equivalent word or expression. Since many words and expressions have several equivalents due to polysemy, the same entries on the left side can be repeated with different equivalents. The alignment dictionary was generated automatically from the database version of the Polish-Ukrainian dictionary that is currently developed as a joint project of ULIF NASU and ISS PAS, and contains 31088 entries.

Fragment of the dictionary:

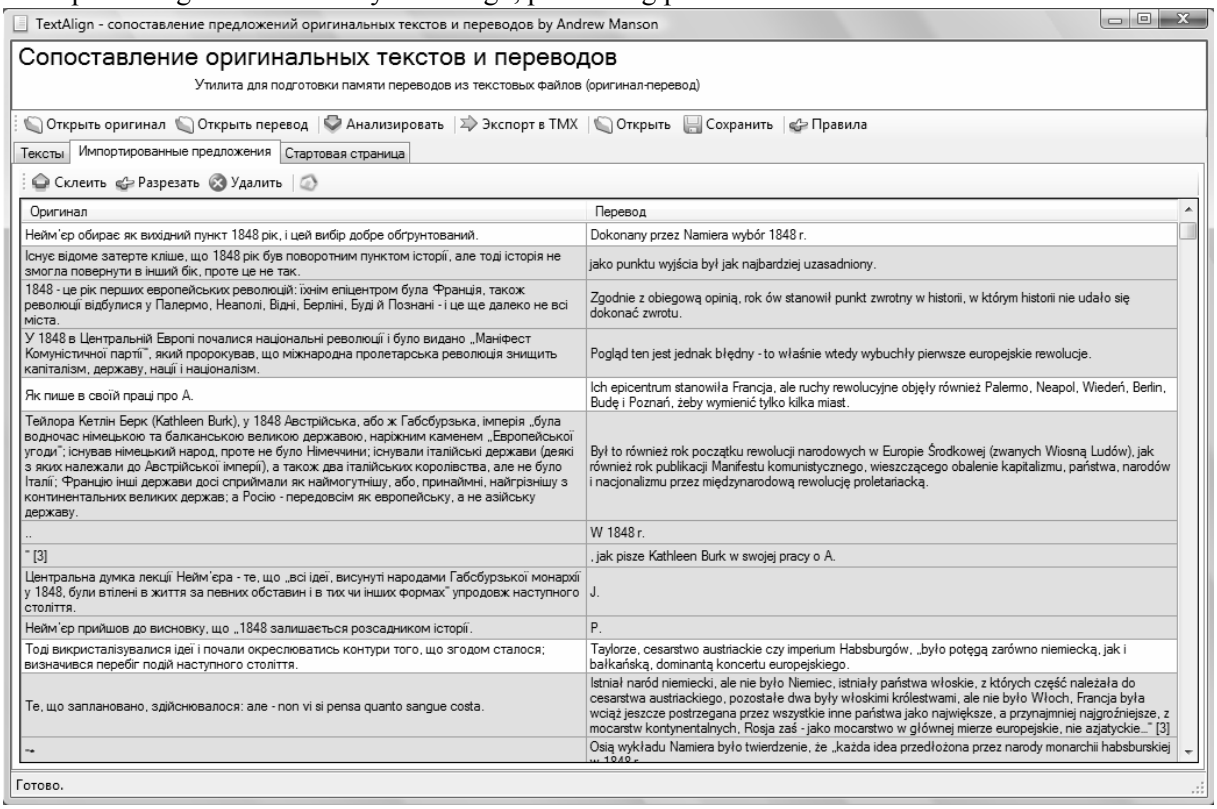
białokrusz @ окис свинцю
blałolicy @ білолиций
blałoramienny @ білоплечий
białoruszczyzna @ білорущина
białoruszczyzna @ все, що білоруське
bibułka @ папіросний папір
bibułka @ цигарковий папір
bibułomania @ манія збирати старі рукописи
biczować @ батожити
biczowanie @ батоження
biczyk @ батіжок
biczykowaty @ подібний до батіжка
bić @ бити
biję @ б'ю
biec @ бігти

Since both Polish and Ukrainian are highly inflected languages, basic dictionary forms are not enough. Either we need lemmatized texts, or a dictionary with all possible forms generated. The first option seems to be easier to realize, but for this we need to adjust the alignment algorithm and to work with already annotated texts.

Another option for aligning is the TextAlign, a user friendly software with GUI and editing possibilities. The only possible input format there is RTF (rich text format), the output is a TMX file with an

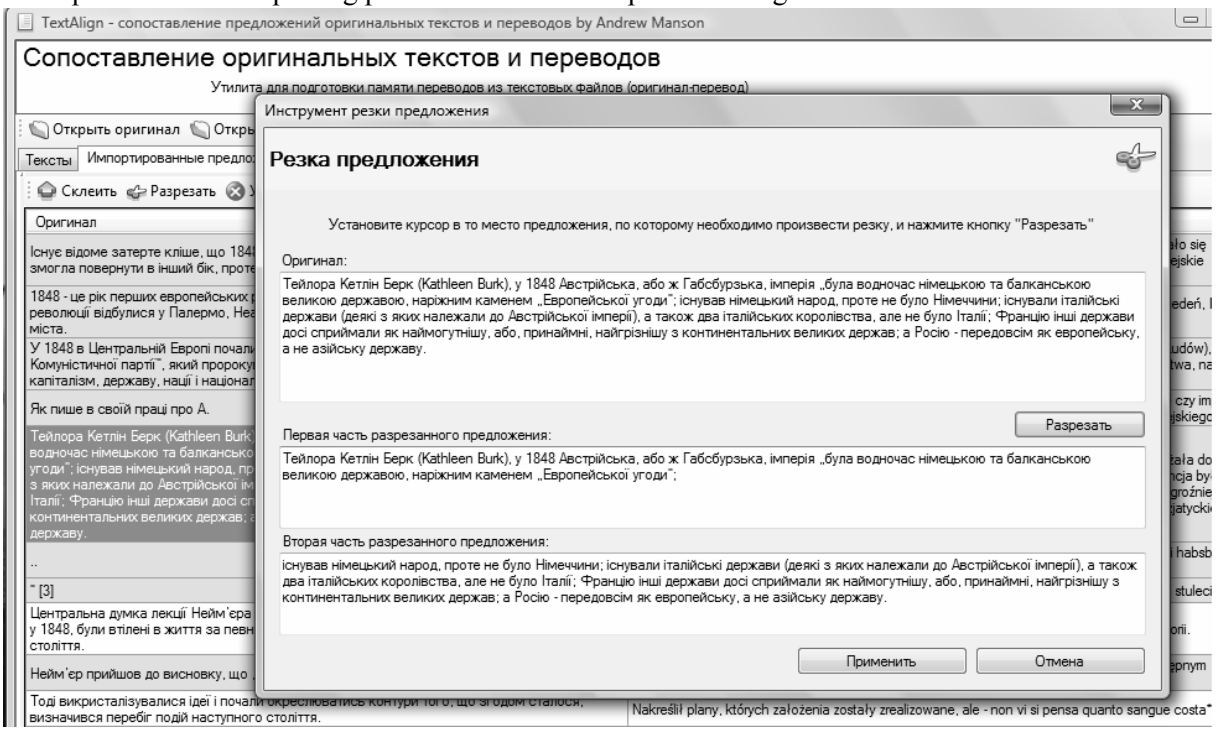
intertwined parallel text. The main problem with the unequal number of sentences in parallel texts that effected the quality of the results produced by the fully automatic and hardly controllable Hunalign is compensated by the possibility of an easy and quick alignment edition in the TextAlign. However, the sentence segmentation algorithm in the TextAlign is too simple for satisfactory results.

Example of alignment results by TextAlign, pre-editing phase



It can be seen from the example above that sentence borders are defined basing on punctuation marks without considering common abbreviations ended with full stops, which can generate wrong sentence segmentation.

Example of a manual splitting procedure with the help of TextAlign.



At the moment we are developing a PLUczeK program that will combine the features of the HunAlign and the TextAlign. It will include an editable plugging-in module of text-segmentation at the paragraph and sentence levels, which has to ensure language independence of the program. The sentence segmentation module is rule based, it presupposes the use of such heuristics as frequently used abbreviations to function as a stop list, combinations and sequences of abbreviations and punctuation marks, forms of the reported speech presentation (that can also be different across languages), cf. also [Rudolf, 2004]. The program will work with both plain texts and morphologically annotated XML files, addressing either information about the actual form of a token, or its lemma, as well as using grammatical information for sentence segmentation (a verb or a preposition cannot be a proper name, hence, written with a capital letter they signal about the beginning of a sentence, etc.). The program will also have a GUI interface and a possibility of manual edition of segmentation.

We have chosen the XCES format for alignment records. Information about corresponding sentences is stored in a separate file. An example fragment of an alignment file is below (sentences 1 i 2 of the second link group are translated as one sentence).

```
<cesAlign>
 <cesHeader>
...
...
  <translations xml:base="http://corpus.domeczek.pl/corpus">
    <translation trans.loc="exampleAna.ua.xml" lang="ua" xml:lang="ua" n="1" />
    <translation trans.loc="exampleAna.pl.xml" lang="pl" xml:lang="pl" n="2" />
  </translations>
</profileDesc>
 </cesHeader>

  <linkList>
    <linkGrp id="p1" targType="s">
      <link>
        <align xlink:href="#p1s1" />
        <align xlink:href="#p1s1" />
    </link>
      <link>
        <align xlink:href="#p1s2" />
        <align xlink:href="#p1s2" />
    </link>
  </linkGrp>
    <linkGrp id="p2" targType="s">
      <link>
        <align xlink:href="#xpointer(id('p2s1')/range-to(id('p2s2')))" />
        <align xlink:href="#p2s1" />
    </link>
      <link>
        <align xlink:href="#p2s3" />
        <align xlink:href="#p2s2" />
    </link>
  </linkGrp>
</linkList>
</cesAlign>
```

Even sentence alignment cannot reach a 100% accuracy due to objective reasons. In the table below, fragments that are parts of one sentence are highlighted with the same shade.

Dokonany przez Namiera wybór 1848 r. jako punktu wyjścia był jak najbardziej uzasadniony.	Нейм'єр обирає як вихідний пункт 1848 рік, і цей вибір добре обґрунтований.
Zgodnie z obiegową opinią, rok ów stanowił punkt zwrotny w historii, w którym historii nie udało się dokonać zwrotu.	Існує відоме затерте кліше, що 1848 рік був поворотним пунктом історії, але тоді історія не змогла повернути в інший бік,
Pogląd ten jest jednak błędny –	проте це не так.
to właśnie wtedy wybuchły pierwsze europejskie rewolucje.	1848 – це рік перших європейських революцій:
Ich epicentrum stanowiła Francja, ale ruchy rewolucyjne objęły również Palermo, Neapol, Wiedeń, Berlin, Budę i Poznań, żeby wymienić tylko kilka miast.	їхнім епіцентром була Франція, також революції відбулися у Палермо, Неаполі, Відні, Берліні, Буді й Познані – і це ще далеко не всі міста.

This means that mistakes are practically unavoidable, especially with large amounts of texts, but it is still possible to keep the general quality of the corpus sufficient for working with it and receiving objective results.

Conclusions and further work

The current state of PolUKR enables already searching for translation equivalents and can be used as a translation memory database by both human translators or researchers and machines. But the corpus can be enhanced in a number of ways, like finer alignment level, enriching with further annotation of different types, including also semantic and referential information. Automatic word-level alignment can be of significant help while compiling bilingual dictionaries. The search engine has to be adjusted to enable searching for the new information as well.

Literature

Broda Bartosz, Piasecki Maciej & Radziszewski Adam. Towards a Set of General Purpose Morphosyntactic Tools for Polish. Proceedings of Intelligent Information Systems, Zakopane Poland, 2008. Institute of Computer Science PAS, 2008.

Ivan Derzhanski and Natalia Kotsyba. The Category of Predicatives in the Light of Consistent Morphosyntactic Tagging of Slavic Languages. Proceedings of the International Workshop within MONDILEX project, Moscow, 2-4 October 2008.

Hunalign - sentence level aligner: http://mokk.bme.hu/resources/hunalign.

Natalia Kotsyba, Olha Shypnivska and Magdalena Turska. *Linguistic principles of organizing a common morphological tagset for PolUKR (Polish-Ukrainian Parallel Corpus)*. Proceedings of Intelligent Information Systems, Zakopane, Poland, 2008. Institute of Computer Science PAS, 2008.

Adam Przepiórkowski and Marcin Woliński. *A Flexemic Tagset for Polish*. In: The Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL 2003. http://nlp.ipipan.waw.pl/~adamp/Papers/2003-eacl-ws12/ws12.pdf

Michał Rudolf. Metody automatycznej analizy korpusu tekstów polskich. Pozyskiwanie, wzbogacanie i przetwarzanie informacji lingwistycznych. Warszawa, 2004.

TextAlign in MT2007 (Memory Translation Computer Aided Tool): http://mt2007-cat.ru/index.html.

Magdalena Turska and Natalia Kotsyba. Polsko-Ukraiński korpus równoległy (PolUKR). „Materiały LXIII Zjazdu Polskiego Towarzystwa Językoznawczego", Warszawa.

Magdalena Turska and Natalia Kotsyba. Polish-Ukrainian Parallel Corpus and its Possible Applications, Proceedings of the International Conference "Practical Applications in Language and Computers, 7-9 April, Łódź", Peter Lang GmbH, 2007.

v. Waldenfels, R. Compiling a parallel corpus of slavic languages. Text strategies, tools and the question of lemmatization in alignment. In: Brehmer, B., Zdanova, V., Zimny, R. (Hrsg.); Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9. München, 123-138, 2006.

Коциба Наталія. Принципи морфосинтактичного тагування польсько-українського паралельного корпусу (PolUKR). Proceedings of the International Conference "MegaLing'2008. Horizons of Applied Linguistics and Linguistic Technologies, Parthenit – Crimea, Ukraine, September 2008", 2009 (in preparation).

Широков В.А, О.В.Бугаков, Т.О.Грязнухіна, О.М.Костишин, М.Ю.Кригін, Т.П.Любченко, О.Г.Рабулець, О.О.Сидоренко, Н.М.Сидорчук, І.В.Шевченко, О.О.Шипнівська, К.М.Якименко. Корпусна лінгвістика. Київ: Довіра, 2005.

Towards Creation of the Polish Grammatical Dictionary[1]

Igor Shevchenko
Ukrainian Linguo-Information Fund,
National Acacdemy of Sciences, Ukraine, Kyiv
ivislander@gmail.com

Abstract
The report has examined the formation of the Polish grammatical word-inflexion dictionary on the basis of linguistic similarities with the current Ukrainian grammar dictionary. The notions of the word-inflexion parameter and the word-inflexion class have been described. The cases of convergence and differentiation in word-inflexion classes of the related languages have been considered.

Key words: *grammatical dictionary (GD), word-inflexion class (WIC), part of speech (POS), highly inflected languages, Slavic languages, Polish grammatical dictionary (PGD), Ukrainian grammatical dictionary (UGD).*

For the inflective languages grammatical dictionaries that provide description of the word-declination and word-formation are of great importance. Creation of an exhaustive set of the variants of the lexeme as well as the rules of word formation is an inevitable step on the way of natural language processing to provide first of all, the lemmatization of word forms, i.e. their identification with the initial form available in dictionaries, morphological analysis and synthesis, grammatical tagging of text corpora.

A grammatical dictionary as we understand it deals with the word declination and has to contain all the forms of the inflected words of the certain language with their grammatical features. Variety of those forms especially in Slavic languages makes such a task far from being trivial. As regards grammatical dictionaries we should first of all mention the fundamental work by Andrey Zalizniak[1] that has made a breakthrough in making a uniform and exhaustive description of the word-declination in a Slavic language though not computer-aided initially but however quite applicable for computer processing.

The Grammatical dictionary of the Ukrainian language (UGD) developed in the ULIF NASU[2] provides by now a partition of the lexemes fixed in dictionaries into 2456 word-inflexion grammatical classes (WIC), each of them presenting a set of lexeme endings according to their grammatical meanings, unique and uniform inside the class and therefore embraces all the types of word-inflexion in Ukrainian[3]. The UGD have been instrumental in making the first ever Ukrainian integrated lexicographical system "Dictionaries of Ukraine"[4]. Eight editions of this digital dictionary have come off on CD-ROMs in 2001-2008.

The program shell has been prepared by Dr. Olexandr Rabulets. We also appreciate a lot of valuable corrections and amendments to our grammatical classification made by my colleagues from Ukrainian Lingua-Information Fund within these years. Some amendments were also proposed by users of the integrated lexicographical system "Slovnyky Ukrainy" (Dictionaries of Ukraine). All the research and development of the grammatical dictionaries has been carried out and is being carried out under the guidance of Prof. Volodymyr Shyrokov.

The uniformity of the word-inflexion inside a grammatical word-inflexion class means in terms of the computer processing of written texts that all the words belonging to a WIC have the same set of the grammatical meanings and in each of the grammatical meanings (and, besides, in each of the variants of grammatical meanings if there are several of them) the same number of characters from the right is replaced with the same line of characters. Thus, the words belonging to the same WIC differ in their invariable parts only.

[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

As a matter of fact, a WIC is a set of words with the same type of word-inflexion, which is characterized by a set of values of the word-inflexion parameters[5]. The conformity of the word-inflexion for a lot of lexemes different in meaning and form in an inflectional language allows us to specify the following grammatical word-inflexion parameters.

1. Part of speech (or as its word-inflexion generalization – word-inflexion type)
2. Type of word stem
3. Conjugation pattern
4. Type of consonant-vowel changes
5. Paradigm incompleteness
6. Atypical word-form features in certain grammatical meanings
7. Type of the accent distribution in the word-inflexion paradigm.
8. Aspect (for verbs)
9. Reflexivity (for verbs)
10. Imperative form (for verbs)
11. Passive participle suffix (for verbs)
12. Gender (for nouns)
13. Denotatum type (for nouns)
14. Form of the genitive case for masculine nouns
15. Form of the locative for masculine and neutral nouns
16. Form of the dative for masculine nouns
17. Form of the accusative case in plural (for nouns)

WIC is determined by a combination of the parameter values a given word-inflexion class, for example, all the masculine nouns of the second declension that indicate persons, end in a soft consonant, drop the vowel -e- in indirect cases and do not have atypical features in its inflexion belong to the same WIC (in our classification #1540), examples are: “вьборець’” (“voter”), “ірлfндець” (“Irishman”).

Lexeme	part of speech	declension	basis	change	anim	genitiv	WIC
вьборець видавtць іспfнець промислjвець	n	2dec	soft	-е	person	a	1540

lexeme	part of speech	conjugation	basis	final suffix	aspect	reflex	change	WIC
компенсувfти ліквідувfти нормалізувfти	v	1dec	iota	-ати	imperf+ perf	–	–	382

lexeme	part of speech	declension	basis	change	animal coord	pecul	WIC
Вітfлiїв змsїв ненціїв	a	possessive	hard	ї-є	+	–	2335
всzкий дtкотрий жjдний	pron	general	hard	–	+	–	1145

The object of identification is the initial form of the word as it is recorded in the dictionary entry. A word-inflexion parameter makes sense only for certain groups of words according to their grammatical function[6]. So, there is no use considering the option "form of the genitive" and the "gender" for a verb (if the matter is gender as characteristic of a lexeme), or the parameter "aspect" or "form of imperative" with respect to a noun or an adjective. Then, the parameter is characterized with its domain. The word-inflexion parameter can be regarded as a discrete function with a limited range of possible values (the value area). As an example, the well-known list of parts of speech can be made. The parameter "type of the word stem" can get one of 5 values: hard, soft, combined, iotacized and r-type. The parameter ”gender” has potentially 10 different values for a lexeme (each of three genders, six of their combinations by the order of two and,

besides, one combination of all three genders). A form of the genitive for masculine nouns has three values: -a (or -я, depending on the word ending), -у (-ю), or both -а (-я) and -у (-ю) are possible.

Each of the word-inflexion parameter values can be implemented only for words that have a certain form. For example, the Ukrainian verbs can end only in -ти, or -тися. Any word entry with a different form cannot be considered as a candidate for a verb in the process of grammatical identification. At the same time, the word that ends in -ти is not necessarily a verb. For example, it can be a pluralia tantum noun, for example: «грати» (bars). Thus, the words with the ending -ти and -тися set a domain for the parameter. Belonging to the domain does not imply that the parameter value for the word is really implemented. In such cases, it is logically to say about an optional parameter since a word that ends in -ти can be a verb but not necessarily. There are however some cases when the word determines a parameter value unconditionally. For example, any Ukrainian feminine noun which has the last consonant of -к, has a change to -ц in the dative and locative cases. Thus, for this very productive group of words, the above value of the change is mandatory, so one can talk about an area of mandatory implementation for the parameter value.

The same approach can be applied to other inflexional languages. We should stress that the Polish word-declination system is fully described and we did not intend to make any discoveries in this domain thoroughly examined and described by a number of well-known scholars as Jan Tokarski, Janusz Stanisław Bień, Zygmunt Saloni and others. First of all the comprehensive index of Polish word forms by Jan Tokarski[9] as a fundamental work in this domain should be mentioned and the Grammatical dictionary of the Polish language that came off in 2007[12]. Thus, our goal was to try to put the description of word-inflexion in two Slavic languages in conformity within one ideology and one structure. We must especially thank Dr Zygmunt Saloni and Dr Marcin Woliński for digital index of the Polish flexions we are using in our research.

Our leading idea to the conclusion of the Polish grammatical dictionary was a wide parallelism in word-inflexion of related languages, namely, Ukrainian and Polish, because of the proximity in lexical composition and an apparent parallelism in the word-inflexion systems of both languages, and in some cases non-flexion modifications (change proper, insert, omission)[7]. Thus, it is assumed that nouns of neutral gender of -nie: “czekanie” (“waiting”), “milczenie” (“silence”), “noszenie” (“carrying”), and others are similar in their inflexion to the Ukrainian WIC #2108 covering singularia tantum nouns of neutral gender ending with -ння: “стояння” (“standing”), “малювfння” (“drawing”). Those ending with -сць: “legalność” („legality”) „doskonałość” (“perfectness”) and many others meet the Ukrainian WIC #2143 covering singular nouns of the 3rd declension the change o-i some cases “актbвність” (“activity”), “раптjвість” (“suddenness”). Adjectives with the -y in the end “rodzimy” (“home, native”), “ogólny” (“general”) obviously have the Ukrainian corresponding WIC #2302 “бsлий” (“white”), “спsльний” (“common”), that brings together adjectives with the hard ending. In the verb ending with -ać “dbać” (“take care”), “spać” (“sleep”) we recognize the Ukrainian WIC #697 “дбfти” (“take care”), “спати” (“sleep”), i.e. verbs of the 1st conjugation with iotacized endings in present tense and without passive participle in the paradigm.

мова	Lexeme	part of speech	declension	basis	change	anim	genitive	WIC
Ukr.	стоzння	n	2dec	hard	–	inanimate	a	2108
Pol.	czekanie	n	2dec	hard	–	inanimate	a	2108
Ukr..	раптjвість	n	2dec	hard	i-o	inanimate	a	2143
Pol.	legalność	n	2dec	hard	–	inanimate	a	2143

мова	Lexeme	part of speech	declination	basis	change	animal coord	pecul	WIC
Ukr.	білий	adj	general	hard	–	+	–	2302
Pol.	ogólny	adj	general	hard	–	+	–	2302

мова	Lexeme	part of speech	conjugation	basis	final suffix	aspect	reflex	change	WIC
Ukr.	дбати	v	2dec	iota	-ати	imperf	–	–	697
Pol.	dbać	v	2dec	iota	-аць	imperf	–	–	697

Such analogies enable us to make the first rough partition of the Polish dictionary entries into specific word-inflexion protoclasses that still require more detail and further differentiation. By the ending of forms, as well as the elements of word-inflexion paradigm available in the dictionary through a series of global replacements we can ascribe the WIC numbers to a large part of lexemes presented in the dictionary. Further refinement of supplies, as well as the detection of rare classes is to be carried out manually.

Of course, there is a lot of difference in word-inflexion system of the two languages.

First of all, rather fortunately in Polish language there is no accent wandering and therefore no need in fixing types of the accent distribution. So, we should not keep in mind the accent position in Polish words and it is a great relief since to compare with, in Ukrainian the whole system of accent distribution embraces about 700 types of accentuation thanks to the phenomenon of wandering accent in word-forms of the same lexeme like in зірка (“star”) – зірки (“stars”) or брати (“to take”) – беремо (“we take”), ходжу (“I walk”) – ходиш (“you walk”) .

Secondly, we do not deal with reflexive verbs in Polish because the reflexive particles spell separately from the verbs they belong to unlike their Ukrainian counterparts, reflexive suffices spelled together and often considerably modifying the verb conjugation.

We should note that the existence of far-going analogies between the WICs of the Ukrainian and Polish languages does not mean a total coincidence of the endings as well as their word-inflexion parameters. For example, in Polish there is no change of a vowel in the feminine nouns ending with -ość: “twórczość” (“creativity”), “miejcowość” (“area”), that is inherent in similar Ukrainian nouns ending with -ість: “влечність – влечності” (“marksmanship”) in a number of indirect cases.

Consider the process of differentiation of Polish word-inflexion classes for example, WIC = 1607, which covers masculine nouns of the 2nd declension with hard consonant at the ending with -a flexion in the genitive without change, designating inanimate objects (Ukrainian representative “гриб” ("mushroom"). The clearest counterpart to this Ukrainian word-inflexion class can be found in similar vowel-invariable Polish entries like "chleb" (“bread”).

At the same time, there are some types of change for this group of Polish nouns that are not inherent in the Ukrainian word-inflexion. This includes the change t-c: “instytut – loc. instytucie” (“institute”), “pirat – loc. piracie” (“pirate”), d-dz: „sąd – loc. sądzie” (“court of justice”), Szwed – loc. Szwedzie (Swede), as well as the double change of the lexeme “obiad” – loc. “obiedzie” (“dinner”). These changes allow us to single out extra WICs absent from the Ukrainian system of word-declension.

мова	lexeme	part of speech	declension	basis	anim	genitiv	change	WIC
Ukr.	гриб	n	2dec	hard	inanimate	a	–	1607
Pol.	chleb	n	2dec	hard	inanimate	a	–	1607
Pol.	sąd	n	2dec	hard	inanimate	a	т-ц	1615
Pol.	instytut	n	2dec	hard	inanimate	a	д-дз	1627
Pol.	obiad	n	2dec	hard	inanimate	a	г-з, е-я	1635

One more example of divergence. In the Ukrainian language there are different types of inflexion for nouns designating people and those for animals, because the accusative plural case of words to designate animals is realized in two optional forms one of which matches with the nominative plural and another coincides with the genitive: "пасти коні" and "пасти коней" (to graze horses). When designating persons, only the form coinciding with the genitive is acceptable: “зустріти дівчат” (not “*зустріти дівчата”) (“to meet girls”). For the Polish language such a differentiation does not exist since the word-declination parameter “denotatum type” in Polish has less values than in Ukrainian and therefore is less differentiated than in Ukrainian because there is a difference in word-declension between lifeless and live objects but the difference between the “animal” and the “person” substantial for the Ukrainian word-declension does not matter in respect to the Polish as the variants as both “spotkać kobiety” and “spotkać kobiet” (“meet women”) and „kupić krowy” and „kupić krów” (“buy cows”) are acceptable though with a somewhat different shade of meaning. Another minor difference. The vocative case in Polish is traditionally ranking 5th in the declension table while in Ukrainian the case before recently considered to be rather an optional vocative form is placed in the last, 7th position.

Thus, this approach allows us, firstly, to obtain a detailed word-inflexion classification of the Polish language in conformity with our Ukrainian GD and on the other side, study and summarize the differences in the word-inflexion systems of closely related languages. At the moment we have in our classification about 400 word-inflexion classes. The work on the grammatical dictionary of the Polish language is going on.

In the end a few words about advantages of the approach we develop.

GDs can be used in a variety of ways, e.g. the statistics of usage given by a GD can help us trace more common patterns of word-inflexion in similar classes of words, which can be useful for recommendations on standardization, considering the current variability of existing forms in both Ukrainian and Polish. Statistics of WICs can be of use in grammatical homonymy disambiguation.

GDs can be a powerful tool for comparative studies too, a much neglected by computational linguistics area so far.

GDs are corpus-driven, so they help us reveal the information about a language that is not covered in grammars, or is not covered consistently or clear enough for the users.

We did not find in the Ukrainian prescriptive grammars an indication to the verbs that do not lose the -ва- suffix in the imperative forms as in "давай" (give) though displaying the above omitting in the indicative present tense forms; compare "даю, даєш,..." (I give, you give) unlike a much commoner type "шануй,.. шаную, шануєш (respect), працюй,.. працюю, працюєш (work). Such an explication may be substantial for instance in textbooks for foreign students.

As regards the updating of the GD a new lexeme can be automatically ascribed to a WIC of the lexeme already available in the dictionary that has a maximum affinity to the new word in its form. For instance, the new word "наркомафія" receives the WIC #1185 just as the well-known word "мафія". This approach allows us to identify grammatically large lists of words, for instance those incorporated into the GD from a certain terminological dictionary or another document. The identification of some words can however prove to be wrong or incomplete. So, the word "колайдер" may be put in accordance with the word "провайдер" as the closest in its form to the former in the dictionary. And therefore, the WIC #1766 for masculine nouns designating persons would be prescribed for the new lexeme instead of the correct WIC #1607 (masculine nouns designating lifeless objects). Thus, the automated grammatical identification requires a check-up by an expert. Another way is a dialogue expert-computer system (not realized as yet) or merely the entry of certain parameters by an expert beforehand, for example, specifying the part of speech, denotatum type (person, animal, lifeless), genitive case flexion and so on.

As regards the grouping of the lexemes in the GD bigger classes arouse naturally if grouping lexemes according to their word-declination parameters such as parts of speech, word-stem type, type of changes, aspect (for verbs), gender (for nouns) and so on and any of their combinations. Using those "bigger groups" is the only feasible way of operating the GD since dealing with isolated WIC numbers without any interdependence among them is far from being practicable.

The final goal of this activity can be considered as an intellectual computer-aided system capable of correcting errors and explain how and why a word or a word-combination should be spelled and pronounced.

Literature.

1. Зализняк А.А. Русское именное словоизменение. М.: Наука, 1967. 370 с.

2. Шевченко И.В., Широков В.А., Рабулець А.Г. Электронный грамматический словарь украинского языка. // Труды международной конференции «Megaling'2005. Прикладная лингвистика в поиске новых путей». 27 июня – 2 июля 2005 года. Меганом, Крым, Украина. – С. 124–129.

3. Шевченко І.В. Алгоритмічна словозмінна класифікація української лексики. // Мовознавство. – 1996. – №4–5. – С. 40–44.

4. Широков В.А., Рабулець О.Г., Шевченко І.В., Костишин О.М., , Якименко К.М. Інтегрована лексикографічна система „Словники України", версія 3.1. – К., 2007. – CD-видання.

5. Широков В.А. Інформаційна теорія лексикографічних систем. – К.: Довіра, 1998. – С. 16, 152-174, 274-288.

6. Ігор Шевченко. Параметризація як основа граматичної ідентифікації словникових одиниць української мови. // Прикладна лінгвістика та лінгвістичні технології. Megaling-2007. Збірник наукових праць. К.: Довіра, 2008. – С. 393–402.

7. Słownik polsko-ukraiński we dwóch tomach. Kolegium redakcyjne: A. I. Gęsiorski. T. Ł. Humecka (redaktor naczelny], M. Kiernycki, M.J. Onyszkiewicz, M.I. Rudnycki. 1958.

8. Gramatyka współczesnego języka polskiego : składnia, morfologia, fonologia. Warszawa : Państwowe Wydawnictwo Naukowe, 1984.

9. Jan Tokarski. Schematyczny indeks a tergo polskich form wyrazowych. Warszawa : Wydawnictwo Naukowe PWN, 1993. 384 s.

10. Fleksja polska. Edited by Jan Tokarski. Wyd. 3. z uzup. Warszawa : Wydawnictwo Naukowe PWN, 2001. 271 s.

11. Zygmunt Saloni. Czasownik polski. Warszawa. 2001. 260 s.

12. Zygmunt Saloni, Włodzimierz Gruszczyński, Marcin Woliński, Robert Wołosz, Słownik gramatyczny języka polskiego, Wiedza Powszechna, Warszawa, 2007, CD + 177 s.

Digital Etymology (Illustrated by the example of the Etymological Dictionary of Ukrainian language)[1]

Irina Ostapova
Ukrainian Lingua-Information Fund,
NAS of Ukraine, Kiev, Ukraine
iros@zeos.net

Abstract
On the basis of the formal model of the lexicographic system of the Etymological Dictionary of Ukrainian language the technology of the instrumental system for the functioning of this dictionary in the digital media was developed. The mechanism of the working out of the dictionary index is under consideration.

Keywords: *etymology, lexicographic model, dictionary indexing, lexicographic instrumental system.*

The Etymological Dictionary of Ukrainian language (EDUL) is the fundamental lexicographic work created as one of the projects of the Ukrainian National Dictionary Base program [1]. The first volume of EDUL was edited in 1982, the sixth will be edited in 2009. The indexes of the words in each of the languages genetically relative to the register words are the most effective search instrument for etymological dictionaries. The seventh volume should be the multilingual index for the whole dictionary corpus. Indeed it should be the sum of the indexes for each language presented in the EDUL.

At the moment 232 different languages are represented in the corpus of five EDUL's volumes. The individual index which allows to identify all of the true localization of each word of the given language should be formed for each language of the EDUL. The range of the full index of the EDUL will be approximately 120000 units. The work of building of such an index is so large that in the «manual» mode is not technologically justified. Therefore, it was necessary to create special digital lexicographic environment adapted to the EDUL structures and oriented to the creation of Multilanguage index in the automatic mode.

When working on the creation of the EDUL digital version we used the methods that have already been successfully tested to deal with such tasks, in particular, for creating the computer lexicographic database of the new Semantic dictionary of the Ukrainian language [3].

We consider the dictionary as an information system of a special type — lexicographic. According to the theory of the lexicographic systems this is the abstract lingual-information object, oriented to the realization of complex informative description of the lexical-grammatical structures of certain language or several languages [3].

System architecture meets the standard three-level architecture of the information systems ANSI/SPARK, according to which information system has conceptual, internal and external levels of data [2].

As a conceptual model we use the lexicographic data model [3]. You can see its simplified form below:

$$\{I_0(D),\ V(I_0(D)),\ \beta,\ \delta[\beta],\ \mathrm{Red}[V(I^Q(D)]\},$$

where D – modelling object – the Etymological Dictionary of Ukrainian language; $I_0(D) = \{x_i\}$ – multitude of the dictionary register words, so called multitude of the elementary information units in the theory of the lexicographic systems; $V(I_0(D))$ – multitude of the descriptions (interpretations) of the elementary information units, that is the dictionary entry texts: $V(I_0(D)) = \{V(x_i)\}$ – dictionary entry with the word heading (registry unit) x_i; β – multitude of the structural elements that were abstracted after the dictionary text analysis; $\delta[\beta]$ – structure that is generated on β by the δ operator; limitation $\delta[\beta]$ on $V(x)$ generate

[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

microstructure of δ(x) dictionary entry; Red[$V(I_0(D)$] – mechanism of the recursive reduction of the lexicographic system. It gives a possibility of sequential identifing more details of the structure for the lexicographic system, in particular, to carry out the distribution of structural elements on the entry for register and interpretation part.

A conceptual model of the dictionary is based on the analysis of EDUL printed version, that is we analyze the typography, organization and structure for the printed texts of dictionary entries, which are interpreted as identifiers of the relevant elements of β and δ(x) lexicographic structures.

As a basic structural element of the EDUL lexicographic system we define an etymological class, which is a block of linear text for dictionary entry. The etymological classes are identified on formal grounds: a structural unit is detected as the etymological class if the unique character sequence used as delimiters can be identified in the entry text. We identified the following types of etymological classes for EDUL: *register word class* (*HEAD*), *derivative class* (*DER*), *slavic matches' class* (*SLAV*), *language class* (*LANG*), *bibliographic class* (*BIBL*), *link class* (*LINK*). Each of these classes has a unique structure, which gave us the possibility to construct a formal procedure for identifying the type of each etymological class in the linear text of the dictionary entry according to the formal criteria.

Let us illustrate this by the example of two small, but rather representative dictionary entries. The texts are presented in the form, maximally close to the printed version.

*Example 1 (dictionary entry with the register word **абетка – "alphabet"**):*

абетка, [*абетло*] Пі, *абетний* (заст.) «елементарний»**;—** власне українська назва азбуки, утворена за вимовою перших двох букв алфавіту (*а, бе*), очевидно, під впливом назв *азбука, альфабет* і п. abecadło «тс.» (від вимови перших трьох букв a, be, ce)**.—** Sadn. – Aitz. VWb. I 42**.—** Пор. **азбука, алфавіт**.

*Example 2 (dictionary entry with the register word **абзац – "paragraph"**):*

абзац;— р. бр. *абзац*, болг. *абзац*, схв. *абзац*;**—** запозичення з німецької мови; нім. Absatz «перерва, пауза, уступ, абзац» є похідним від дієслова absetzen «відсувати, відставляти», утвореного з префікса ab- «від-, з-», спорідненого з гот. af «від», лат. ab «тс.», і дієслова setzen «садити», пов'язаного з двн. sęzzen, дангл. sęttan, англ. set і спорідненого з псл. saditi, укр. **садити.—** CIC 7; Фасмер I 56; Paul DWb 8, 10; Kluge — Mitzka 705.— Див. ще абажур, садити.— Пор. обцас.

*Example 3 (etymological classes for dictionary entry **абетка – "alphabet";** the texts are given in the angle brackets):*

HEAD ≡ <**абетка**>

DER ≡ <[*абетло*] Пі, *абетний* (заст.) «елементарний»>

LANG ≡ <власне українська назва азбуки, утворена за вимовою перших двох букв алфавіту (а, бе), очевидно, під впливом назв азбука, альфабет і п. abecadło «тс.» (від вимови перших трьох букв a, be, ce)>

BIBL ≡ <Sadn. — Aitz. VWb. I 42>

LINK ≡ <Пор. **азбука, алфавіт**>

*Example 4 (etymological classes for dictionary entry **абзац – "paragraph"**):*

HEAD ≡ <**абзац**>

SLAVIA ≡ <р. бр. *абзац*, болг. *абзац*, схв. *абзац*>

LANG ≡ <запозичення з німецької мови; нім. Absatz «перерва, пауза, уступ, абзац» є похідним від дієслова absetzen «відсувати, відставляти», утвореного з префікса ab- «від-, з-», спорідненого з гот. af «від», лат. ab «тс.», і дієслова setzen «садити», пов'язаного з двн. sęzzen, дангл. sęttan, англ. set і спорідненого з псл. saditi, укр. *садити*>

BIBL ≡ <CIC 7; Фасмер I 56; Paul DWb 8, 10; Kluge — Mitzka 705>

$LINK_1$ ≡ <Див. ще **абажур, садити**>

$LINK_2$ ≡ <Пор. **обцас**>

The connections of register word with certain words from other languages are established in the text of each etymological class. All of these words, including the register ones, are called *etymons*. When analyzing the texts of the etymological classes eight parameters for etymons description have been determined: *language marker* (P_L), *remark for language marker* (P_{RL}), *character representation of etymon* (P_A), *belonging to the dialectal vocabulary* (P_D), *homonymy marker* (P_o), *interpretation* (P_S), *remark* (P_R), and

bibliography (P_B). We have listed the parameters in the order in which they are usually followed in the text of the etymological class. Two parameters are required: P_L (*language marker*) and P_A (*character representation of etymon*). These two parameters provide the uniqueness of each etymon for the dictionary entry: etymons with the same character form may have different language marker, or the etymons of the same language may have different character forms. Other options are optional. A formal procedure that allows identifying the relevant parameter in the text for each etymological class is determined for each parameter.

We will call $\{P_L, P_{RL}, P_A, P_D, P_O, P_S, P_R, P_B\}$ parameter set as an etymon-structure and will mark it with $ETYM(e_i)$ symbol, where e_i is a relevant etymon; index i is a sequence number of this etymon in the text. The order of parameters in the etymon-structure is not significant.

Not all the parameters are relevant for each etymological class. The text we identify as an etymological class uses its subset of parameters. Not each etymon should be described by the complete set of parameters. However, one type of etymon-structure is built to achieve structural homogeneity for each class. If certain parameter is not applicable or can not be identified by formal characteristics, an empty text line corresponds to its value. Etymon-structure is built only if it was possible to identify P_A.

Let us illustrate etymon-structures by the example of the texts for etymological classes:

Example 5 (etymon-structures for register word class):

HEAD (**абзац**) ≡ <абзац>

ETYM (e_1) ≡ {P_L= <укр.>, P_A = <абзац>}

Example 6 (etymon-structures for derivative class):

DER (**абетка**) ≡ <[абетло] Пі, абетний (заст.) «елементарний»>

ETYM (e_1) ≡ {P_L= <укр.>, P_A = <абетло>, P_D = 1, P_B = <Пі>}

ETYM (e_2) ≡ {<P_L= <укр.>, P_A = <абетний>, P_R = <(заст.)>, P_S = <«елементарний»>}

Homonymy parameter P_O for etymon e_1 takes value 1, because the square brackets indicate word belonging to the dialectal vocabulary. By default, the value of this parameter for all etymons is 0.

Example 7 (etymon-structures for Slavic matches' class):

SLAV(**абзац**) ≡ <р. бр. *абзац*, болг. *абзац*, схв. *абзац*>

ETYM (e_1) ≡ {<P_L= <р.>, P_A = <абзац>}

ETYM e_2) ≡ {<P_L= <бр.>, P_A = <абзац>}

ETYM (e_3) ≡ {<P_L= <болг.>, P_A = <абзац>}

ETYM (e_4) ≡ {<P_L= <схв.>, P_A = <абзац>}

Example 8 (etymon-structures for language class):

LANG (абзац) ≡ <запозичення з німецької мови; нім. Absatz «перерва, пауза, уступ, абзац» є похідним від дієслова absetzen «відсувати, відставляти», утвореного з префікса ab- «від-, з-», спорідненого з гот. af «від», лат. ab «тс.», і дієслова setzen «садити», пов'язаного з двн. sęzzen, дангл. sęttan, англ. set і спорідненого з псл. saditi, укр. *садити*>

ETYM (e_1) ≡ {<P_L= <нім.>, P_A = <Absatz>, P_S = <«перерва, пауза, уступ, абзац»>}

ETYM (e_2)≡ {<P_L= <нім.>, P_A = <absetzen>, P_S = <«відсувати, відставляти»>}

ETYM (e_3) ≡ {<P_L= <нім.>, P_A = <ab->, P_S = <«від-, з-»>}

ETYM (e_4) ≡ {<P_L= <гот.>, P_A = <af>, P_S = <«від»>}

ETYM (e_5)≡ {<P_L= <лат.>, P_A = <ab>, , P_S = <«тс.»>

ETYM (e_6)≡ {<P_L= <нім.>, P_A = <setzen>, P_S = <«тс.»>}

ETYM (e_7)≡ {<P_L= <двн.>, P_A = <sęzzen>}

ETYM (e_8) ≡ {<P_L= <дангл.>, P_A = <sęttan>}

ETYM (e_9) ≡ {<P_L= <англ.>, P_A = <set>}

ETYM (e_{10}) ≡ {<P_L= <псл.>, P_A = <saditi>}

ETYM (e_{11}) ≡ {<P_L= <укр.>, P_A = <saditi>}

The main problem of computer dictionaries creation, based on their print versions, is forming the relevant databases directly from the dictionary text (parsing) in the automatic mode. Experience shows that forming the lexicographic databases «manually» from the large and complex dictionary texts is practically

impossible. The main task of parsing is an identifying the structural elements directly from the dictionary text, since they are the elements of the lexicographical database.

Before the conversion the texts of all the volumes have been converted to HTML format and unified according to the file structure, and to the character system. Different volumes of the dictionary have been prepared for publishing with various publishing technologies. The first three volumes have been prepared with the precomputer monotype technology. Therefore, first the printed texts have been scanned, recognized with the FINEREADER program and then read. The texts of the 4-th and the 5-th volumes have been prepared in the publishing system; for this a special set of computer fonts, similar to those used in the 1, 2, 3 volumes have been developed. The character system of the dictionary has been unified according to the UNICODE 3.0. This allows making an inventory of the alphabet characters to represent each language etymon.

For the connection between print and digital versions of the dictionary, each dictionary entry was marked as follows: volume number, page number for beginning of the text, page number for end of the text.

Example 9 (the format of digital text for the dictionary entry ***абзац*** *–* ***"paragraph"****; character # is used for identifying the stress):*

@1 38 38

а#бза#ц;-- р. бр. <I>абза#ц</I>, болг. <I>а#бзац</I>, схв. абзац;-- запозичення з німецької мови; нім. Absatz «перерва, пауза, уступ, абзац» є похідним від дієслова absetzen «відсувати, відставляти», утвореного з префікса ab- «від-, з-», спорідненого з гот. af «від», лат. ab «тс.», і дієслова setzen «садити», пов'язаного з двн. sęzzen, дангл. sęttan, англ. set і спорідненого з псл. saditi, укр. <I>сади#ти</I>.-- СІС 7; Фасмер І 56; Paul DWb 8, 10; Kluge -- Mitzka 705.-- Див. ще абажу#р, сади#ти.-- Пор. о#бца#с.

As a result of these operations prepared in special manner texts of the volumes of the Etymological Dictionary fully ready for the automatic conversion into the lexicographic database were obtained.

The structure of the database represents a number of related tables, as presented in Figure 1.

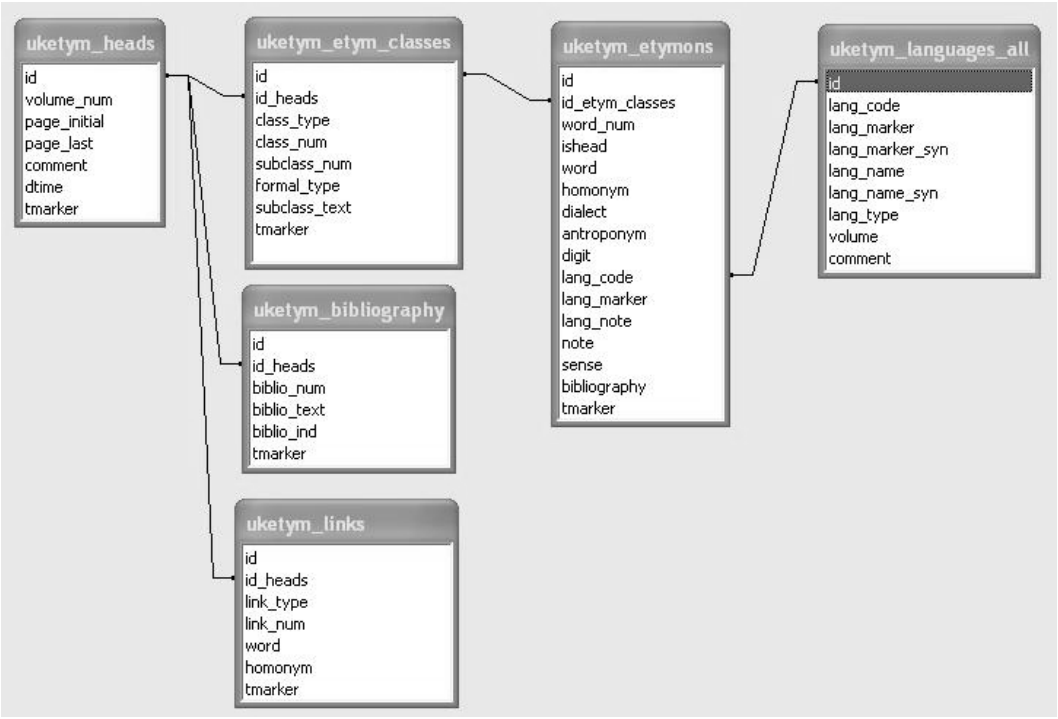


Fig. 1. Structure of the Lexicographical Database for the Etymological Dictionary of Ukrainian language.

Here is a brief description of database tables.

1. The **ukretym_heads** table organizes the dictionary entry text. It stores only the identification number of a dictionary entry, the numbers of the volume, the first and the last pages for dictionary entry text location in the printed version.

2. The **ukretym_etym_classes** table of language classes. It stores dictionary entry texts (excluding bibliography and references, which are organized in separate tables).

3. All inter dictionary entries links are organized in the **uketym_links** table.

4. The bibliography is organized in the **uketym_bibliography** table.

5. The **uketym_etymons** table contains parameters for etymon-structures.

6. All information on the languages used in the dictionary is organized in the **uketym_language_all** table. The user language registers for indexing are built on the basis of this table. Here is a fragment of the table:

id	Lang_code	lang_marker	lang_name	volume
127	127	п.	Польська	1
183	183	укр.	Українська	1
231	0	невизн.	невизначена	1

As we can see, there is one more, "technological" language in the table — indefinite (мова «невизначена»). This is done for the following reasons. Since indexing is automatic, we cannot always clearly determine the language according to the developed algorithm. In this case, the etymon is marked with an undefined language. On this characteristic we can select all the similar records from the database (for example, creating a language register, which consists only of technological language) and then make the identification of languages using the editing tools provided in the system.

Special tools that provide the following main functions were built to support a digital version of the dictionary [4]:

1) Traditional logon to the system by the register word and displaying the dictionary entry text;

2) Editing any structural element of the dictionary entry;

3) Creating the etymon-structure for the dictionary entry in the manual mode;

4) Automatic construction of etymon-structure for the dictionary entry;

5) Creating the dictionary entry with a certain structure.

Fig. 2 shows one of the dialogue boxes for editing the dictionary entries. On the left pane, the dictionary entry is presented as a tree of structural elements. An ordered list of etymons is displayed for each etymological class; thereby the depth of etymological research is visualized with graphics means. Using the buttons on the middle pane you can add or remove structural elements and change their sequence. The functions of the buttons vary depending on the selected structural element. For example, when choosing an etymon the button «Додати» (Add) allows adding only an etymon. An editing dialogue box, which reflects the specific character of the element, is developed for each structural element. The text of the relevant class is displayed for each etymon, but editing is prohibited. This gives the opportunity to verify the etymon parameterization, performed automatically.

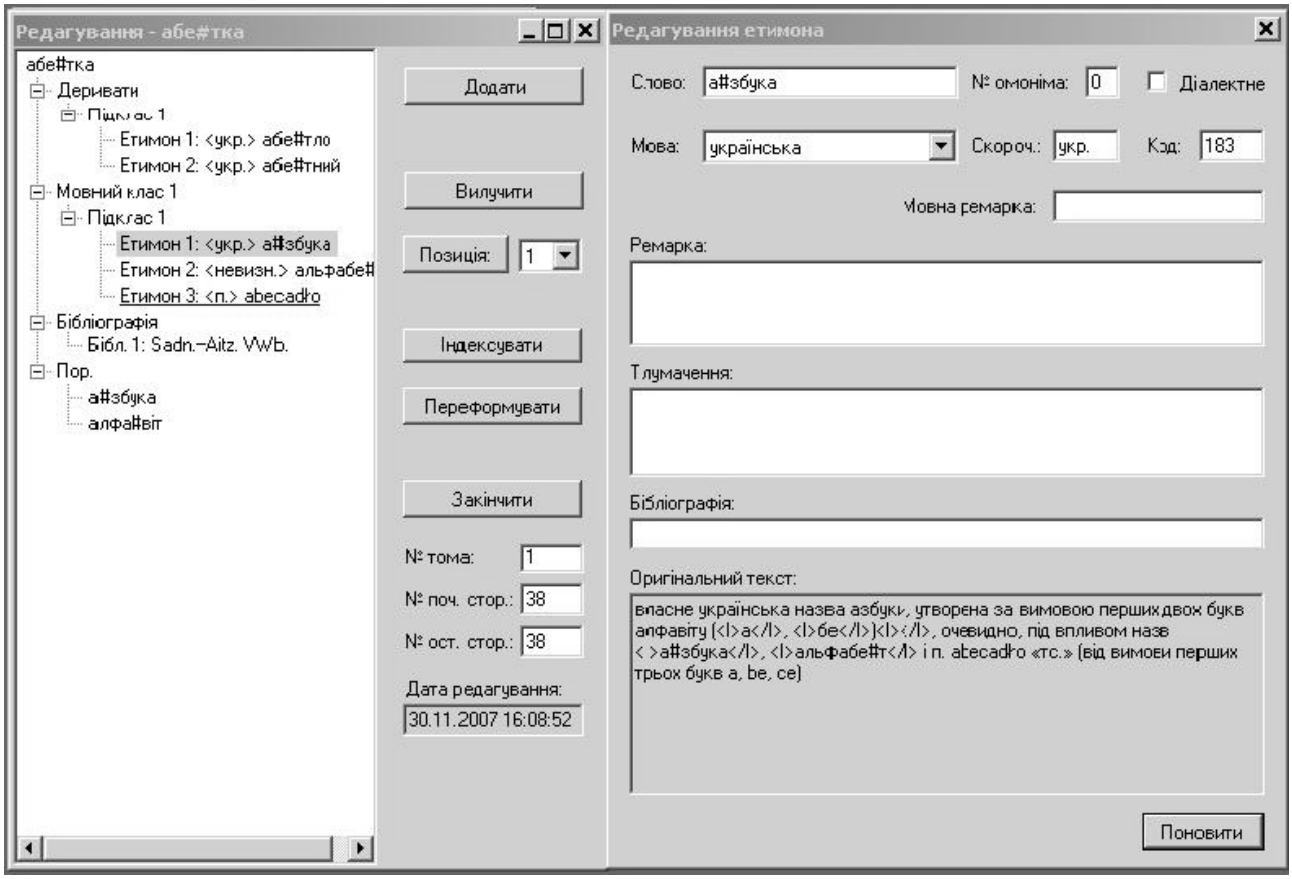


Fig. 2. Dialogue box for editing the dictionary entry with register word ***абетка – "alphabet".***

To automatically build language indexes special tools have been developed, which allows:

1) forming any amount of language registers on the multitude of all languages of the dictionary in the interactive mode;

2) setting the indexation spectra, taking into account the dictionary entry structure.

Fig. 3 shows the user dialogue box for language register forming.

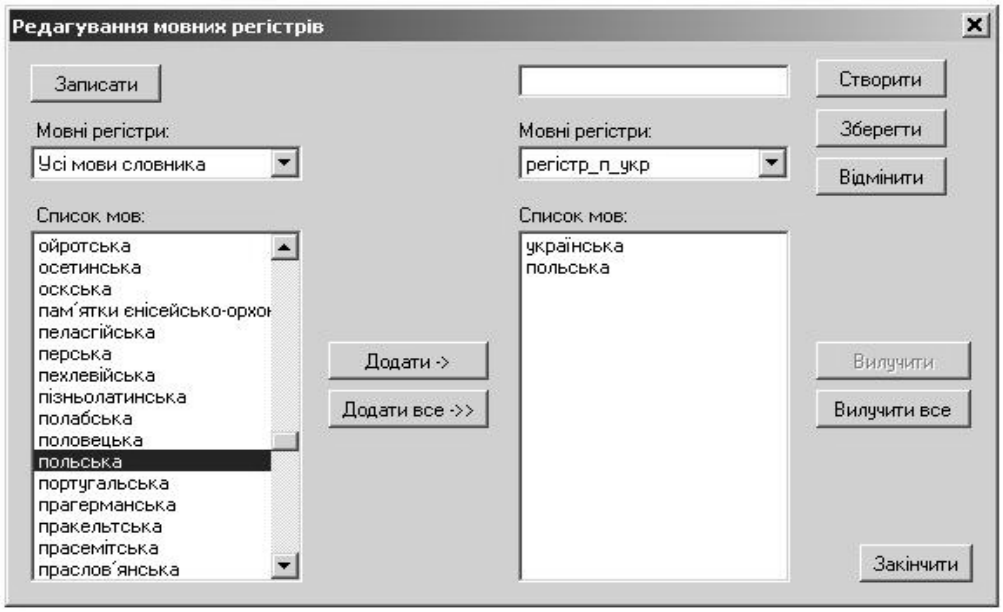


Fig. 3. Dialogue box for language register forming.

The left pane is used for selecting already formed registers as unalterable templates. The right pane is used for editing the existing registers and forming the new ones.

In our example, the new register was formed as follows: its name «регістр_п_укр» was assigned; after verifying the uniqueness the new register was added to the list of language registers; the register «Усі мови словника» (all the languages of the dictionary), which includes all the languages involved in the dictionary, was downloaded to the window on the left pane; the Ukrainian and Polish languages were selected consecutively from the list and moved to the list on the right pane. The register «Усі мови словаря» can be used only as a template. We are going to create a set of templates with meaningful names, for example «Slavic languages», «Romance languages», etc.

Fig. 4 shows the dialogue box of the main user interface for the dictionary with an index built on the formed register.

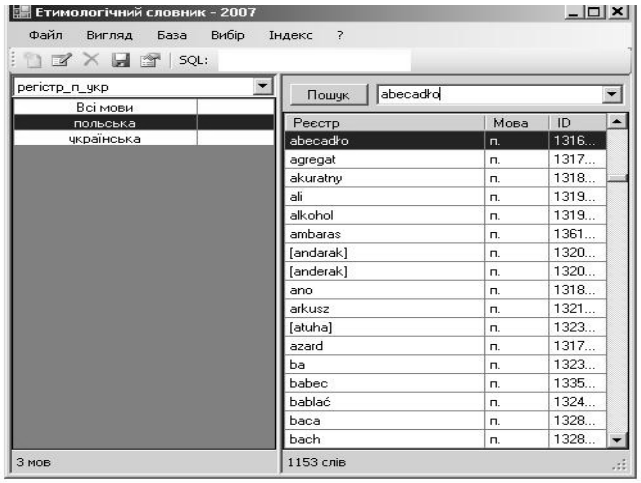


Fig. 4. Language index on the specified register.

Polish was selected from the proposed set of languages on the left pane (you can choose all languages in the register, or a certain subset of languages). A list of all the etymons, identified as words from the Polish language in the **uketym_etymons** table, is displayed in the registry window on the right pane.

The register words of the dictionary entries, in which etymological connections with the Polish language were fixed, may also be displayed in the register window. The formed index is output by the user to a text file showing the localization of each etymon.

The instrumental system allows setting localization of the indexed elements up to a structural element – etymological class type – of a dictionary entry (using menu «Вибір» (Select) on the top pane). In our case we have set only a language class.

When you activate any register element, the dictionary entry text is visualized.

The dictionary entry text for outputting is formed from the relevant database fields. The printing formatting of the dictionary entry is almost completely retained.

The described method of the dictionary modeling gave an opportunity to build a set of relevant etymon-structures as formal representatives of descriptions for the genetic connections of register words for a set of dictionary entries. All the indexing diagrams are built only on the basis of the etymon-structures. This approach gives an opportunity for constructing structures implied in the dictionary entries text, and for displaying the authentic text of the dictionary on the digital environment, which makes digital dictionary open to further interpretations.

The developed technologies and interfaces are offered as a basis for digital representations of etymological works.

REFERENCES

1. Етимологічний словник української мови: В 7 т. Київ: Наукова думка, 1982−2006. Т. 1−5.

1. ANSI/X3/SPARK DBMS study group interim report. FDT-Bull. ACM SIGMOD. 1975. V. 7. № 2.

3. Широков В.А. Елементи лексикографії. Київ: Довіра, 2005.

4. Остапова И.В., Якименко К.Н. Инструментальная лексикографическая система Этимологического словаря украинского языка // Прикладна лінгвістика та лінгвістичні технології. Київ: Довіра, 2008. С. 276−291.

Modelling of the Digital Grammar Dictionary of Russian[1]

Tetyana Lyubchenko,
Ukrainian Lingua-Information Fund,
National Academy of Sciences, Ukraine
ltp@i.com.ua

Abstract
The paper is devoted to the development of digital grammar dictionary (DGD) of Russian. The principles of word-inflexion paradigm formalizarion, the computer technology of DGD creation are considered. Functionalites of created DGD as well as the rezults of use of DGD in research of types of paradigm incompleteness and paradigm variability are described.

Keywords: *digital grammatical dictionary (DGD), word-inflexion class (WIC), Part of speech (POS), paradigmatical class (PC), type of paradigm incompleteness (TPI), type of paradigm variabililty (TPV).*

Introduction

The study is conducted within the framework of the program of the Ukrainian Lingua-information Fund of NASU on creation of digital grammatical dictionaries (DGD).

In particular, it is about the creation of computer databases and instrumental lexicographical systems, serving as the basis for development of word-inflectional models for languages of different types.

Created DGD are oriented to their use as a basic instrument of automatic morphological markup of texts.

We make efforts to develop the formalized apparatus, which would present as far as it is possible the full system of inflexion in languages with which we work, – Ukrainian, Russian, Polish, German, English, Turkish, Spanish.

Principles of the created formalized apparatus will be exposed on the example of forming the digital version of grammatical dictionary of Zaliznyak (Зализняк, 2003).

1. Formal model of Russian word-inflexion

Building a formal model of word-inflexion of inflectional language requires the ascertainment and formalization of linguistic criteria, under which the vocabulary of language units are divided into disjoint aggregate and within each of them word-inflexion occurs by the same rules. The group of the words with such properties is named word-inflexion or paradigmatic classes (PC).

When formalizing the word-inflexion processes we will use the formalism of a multiset. Multiset (MS) or a set with the repetitive elements, a new mathematical concept, as far as we know, mentioned for the first time in the works by D. Knuth (Кнут Д. 1977). In recent years a series of works by Petrovskiy (Петровский А.Б. 1995, 2000, 2003, 2004) dedicated to the development of the theory of multisets and to applications of this theory to decisions analysis in a fuzzy initial information, to the cluster analysis of multi-attribute objects and objects with contradictory properties in Petri nets, etc. has been published.

Partitioning of a multiset of words into paradigmatic classes is being implemented in several stages. First of all a paradigmatic type is defined in accordance with the grammatical categories, grammatical meanings and grammatical forms of a concrete language.

[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

Word-inflexion paradigm of words of different grammatical classes determined by their individual set of inflectional categories. In accordance with inflectional categories, the defining word-inflexion paradigm specific words are entered paradigmatic types (PT).

A detailed description of paradigmatic types of Russian language is presented in our publications (Shirokov, 2005), (Lyubchenko, 2006, 2008). For the inflective words of Russian we have defined the following paradigmatic types: substantival, adjectival, verbal and type of cardinal numerals, and for the invariable words – the so-called "zero" – paradigmatic type.

On the basis of membership in a concrete part of speech, and on additional grounds (non-inflexion characteristics) inside a certain part of speech, MSs of words are divided into submultisets, which will be referred to as grammatical classes to be denoted as follows.

Nouns in respect of the grammatical meaning of "gender" (which in this part of speech is a classification attribute) are divided into 3 grammar classes: masculine nouns, female nouns and neuter nouns; pluralia tantum nouns form a separate grammatical class. Thus, nouns form 4 grammatical classes.

Verbs in respect of the meaning of grammatical category "aspect" (which is regarded by us as classification) are allocated to three grammatical classes: perfective verbs, imperfective verbs and bi-aspectual verbs.

Pronouns are divided into two grammatical classes: pronouns-adjectives, and pronouns-nouns.

All the other words of language are attributed to their grammatical classes on the basis of membership in a particular part of speech.

Thus, in Russian we have identified the following grammatical classes: masculine nouns (P1), female nouns (P2), neuter nouns (P3), pluralia tantum nouns (P4), adjectives (adjectives + ordinal numerals) (P5), perfective verbs (P6), imperfective verbs (P7), bi-aspectual verbs (P8), participles (P9), pronouns (pronouns-nouns) (P10), pronouns-adjectives (P11), cardinal numerals (P12), adverbs (P13), interjections (P14), conjunctions (P15), particles (P16), prepositions (P17), predicatives (P18), abbreviations (P19).

Some paradigmatic type as a rule has several grammatical classes. In the grammatical classes we can single out paradigmatic classes (PC).

The diagram (Fig. 1.) presents the relationship among the paradigmatic types, grammatical classes and paradigmatic classes.

Paradigmatic classes are defined as follows.

1. Each word unit (word-form) x has a representation in the form of a combination of quasi-stem and quasi-inflexion:

$$x = c(x)*f(x), \qquad (1)$$

where c(x) – segment of lexeme x, which is invariable in all the word-forms (quasi-stem), f(x) - its variable component (quasi-inflexion), * – concatenation.

2. Word-inflexion paradigm is represented in the form:

$$\pi(x)=\{c(x)*\{f_j(x)\}\}, \qquad (2)$$

where fj(x), j=1,2,…, n(T_i) – is a quasi-inflexion of corresponding grammatical forms. Every grammatical form can be expressed in more than one word, i.e. in general:

$f_j(x) = \{f_j^l\}$, where l = 1,2, ..., ν – is a multiplicity of grammatical forms.

For example, the lexeme x = 'лёд' to be represented in the form x = 'л' * 'ёд', where 'л' – is a quasi-stem, and 'ёд' – is a quasi-inflexion of initial forms of lexeme. Its word-inflexion paradigm is represented in the form (2), where quasi-stem c(x) = 'л'; quasi-flexions: {f_1(x) = 'ёд'; $f_2^{1,2}$ (x) = {'ьда', 'ьду'}; f_3(x) = = 'ьду'; f_4(x) = 'ёд'; f_5(x) = 'ьдом'; $f_6^{1,2}$(x) = {'ьду', 'ьде'}; f_7(x) = 'ьды'; f_8(x) = 'ьдов'; f_9(x) ='ьдам'; f_{10}(x) ='ьду'; f_{11}(x) ='ьдами'; f_{12}(x) ='ьдах'}.

3. To partition a multiset of words into paradigmatical classes in every grammatical class P_i the paradigmatical relations π_i, are constructed. They are defined as follows:

$\forall\, x_1, x_2 \in P_i\; x_1\, \pi_i\, x_2 \colon x_1 = c(x_1)*f_k\;\; x_2 = c(x_2)*f_k,\;\; f_k \in [F]_k$,

where $[F]^k$ – is a set of quasi-flexions for a certain group of words.

Relation π_i is defined as a relation of paradigmatization.

The relation of paradigmatization is a relation of equivalence since it is reflexive, symmetric and transitive. Factor set forms the set of PCs $\{\Pi_j\}$ belonginig to a grammatical class P_i.

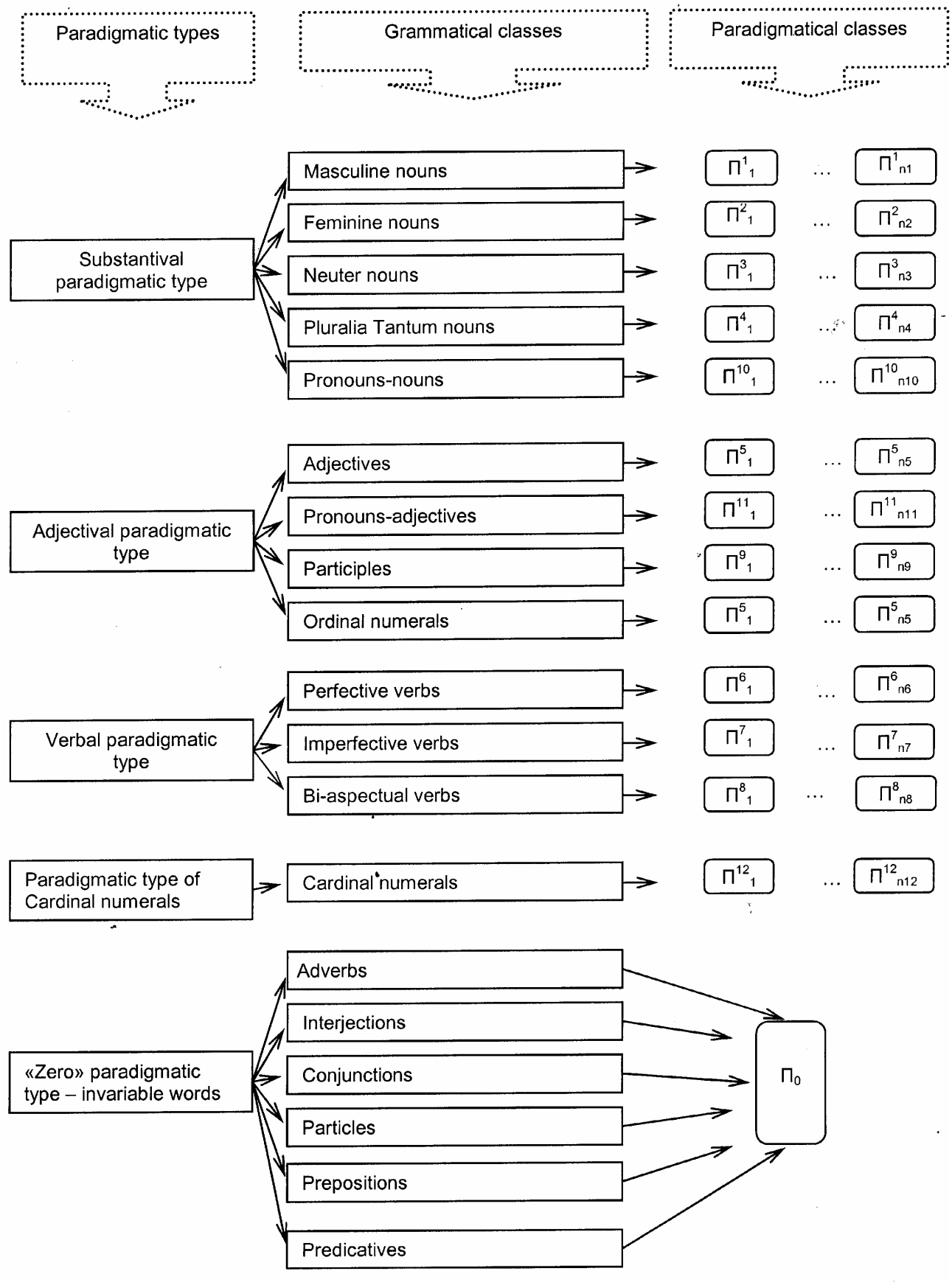


Fig. 1. Relationship among the paradigmatic types, grammatical classes and paradigmatic classes

One paradigmatical class includes all the words with the same pattern (multiset) of quasi-flexions for the corresponding grammatical forms. The words belonging to the same PC differ one from another only in their invariable component c(x) (quasi-stem).

4. To automatically build a complete paradigm by the initial form x0 the operator of paradigmatization (OP) H is determined:

H: $x_0 \rightarrow [x] = c(x)*\{f_0(x), f_1(x),\ldots,f_n(x)\} \equiv \{ c(x)*f_0(x), c(x)*f_1(x),\ldots, c(x)*f_n(x) \}$,

effect of which is determined by the relation of paradigmatization.

The operator of paradigmatization is determined independently for each paradigmatical class.

The operator of paradigmatization H maps lexeme x onto its full paradigm [x]. Algoritmic realization of OP ensures the construction (building) of all the word-forms by the initial form x_0.

Inverse operator H^{-1} returns the initial form of word for any of its word-forms. Algorithmic realization of the operator H^{-1} ensures the process of lemmatization.

The described morphological model is a conceptual base for computer modelling and the implementation of the functions of paradigmatic relations for a certain class of inflectional languages. According to the concept, building of the paradigmatic lexicographical system for inflectional languages can be represented in the form of (Fig. 2):

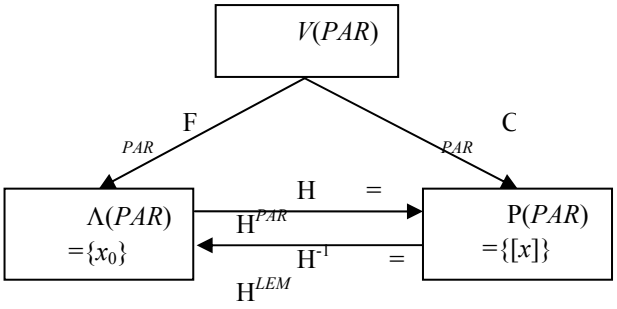


Fig. 2. Paradigmatic structure of L-Systems

Elements in the scheme have the interpretation:

V(PAR) - set of dictionary entries;

$\Lambda(PAR) = F^{PAR}V(PAR) = \{x_0\}$;

$P(PAR) = C^{PAR}V(PAR) = \{[x]\}$;

$H = H^{PAR}$ – operator of paradigmatization: $H^{PAR}x_0 = [x]$;

$H^{-1} = H^{LEM}$ – operator of lemmatization: $H^{LEM}\chi(x) = x_0$, where $\chi(x)$ – is any element of paradigm [x].

2. Formation of lexicographical databases of grammar dictionary

The formation of LDB of Russian grammar dictionary was based on the grammatical dictionary of Russian language by Zaliznyak (Зализняк А.А., 2003) (hereinafter – the GDZ), which gives a sufficiently complete model of the word-inflexion system in Russian.

The technological chain of building of Russian grammatical LDB includes such steps:

- the conversion of the GDZ from a hardware copy to the digital format by scanning;
- correcting the scanned text;
- development of the lexicographical GDZ system, the language of its markup and identifiers of its structural elements;
- automatic conversion of the digital text of the GDZ into the LDB according to the developed structure;
- building a paradigmatic classification (in accordance with the formal presentation of words in the digital grammatical dictionary (DGD));
- development of algorithms for the transition from the classification of GDZ to the classification of DGD, and their software implementation.

Thus a LDB of Russian language which includes about 100,000 word entries was formed. The word list was supplemented with proper names. In addition, the vocabulary needs to be permanently supplemented with new lexical units and the existing ones are to be updated.

For these purposes a software tool set was developed, i. e. the software system, to make the foundation of automated work places for linguists working with the LDB of the Russian language.

3. Functions of the software complex of grammar dictionary

Let us consider the functionality of the tool set designed to maintain the grammatical LDB of Russian.

DGD software interface is developed using the controls operating the Windows environment.

The main program window (Fig 3.) is divided into three zones: the functional area; the vocabulary zone, the zone of lexicographical information.

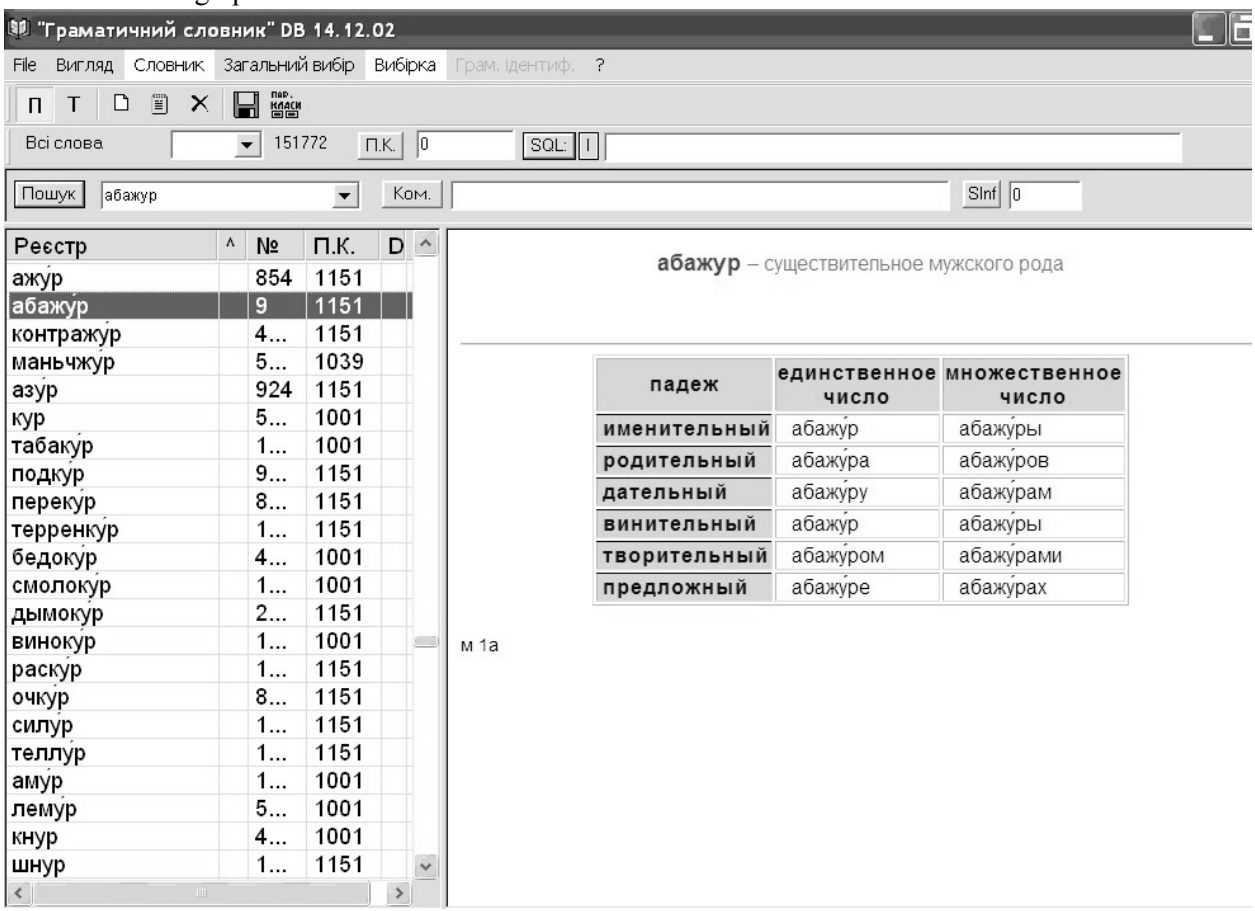


Fig.3. The main program window

The functional area consists of subareas: the general menu, tools for editing, tools to perform queries in the SQL language, an interface for the search of words.

General menu has items "File", “Вигляд”("View"), "Dictionary", Загальний вибIр (“General selection”(/"Total Choice"), „Вибірка”("Selection") and „?”("Help"). Each of the menu contains sub-items which realize the provided functions.

The following functions are realized:

- Browsing of the word list. This function is realized in the vocabulary zone (left part in Fig. 3.)

- Giving out of full word-inflexion paradigm of the word chosen from a vocabulary and its basic grammatical features. This information is displayed in the zone of lexicographical information (right part in Fig. 3.).

- Outputting and browsing of the vocabulary portion (as a part of speech, on the number of paradigmatic class, on any query (composed in the SQL-language). The possibility of outputting of the vocabulary is provided through the general menu and using the tools of functional area of DGD interface. In the general menu there is the item «General selection» with subparagraphs «All», «All with deleted», «Only deleted», «Only active», «Only inactive», «Deleted and inactive», through which you can view the appropriate groups

of words. The item of general menu «Selection» with subparagraphs «All», «Noun», «Adjective», «Numeral», «Pronoun», «Verb», «Participle», «Invariable», «Homonym», «Proper name» provides an opportunity of outputting and browsing of appropriate groups of words (as a part of speech) from the vocabulary. In addition to the opportunities provided in General menu, the selection of groups of words from the vocabulary can be done by the paradigmatical class number, as well as any query in SQL. This opportunity gives a button «PK» in the functional area and an edit box to the right of it; the button is intended to fulfill the query for the output of the group of words belonging to a given paradigmatical class. Button «SQL» is intended to implement the SQL-query, which is written in the text box to the right of the button «I»; button «I» is intended to test the query text.

- Output of all lexical homonymes, proper names, etc.

Such an opportunity is provided through a general menu. The choice of subparagraph «Homonym» of item «Selection» in the general menu initiates the output of lexical homonyms. The proper names are given through a subparagraph «Proper name» of the same menu.

- Displaying of the quantitative characteristics of the selected group of words (how many words has any selected paradigmatical class or parts of speech, how many homonyms there are in vocabulary, etc.).

- Search for words in the vocabulary. This function is provided by means of an interface to search for the word, which consists of the edit box for entering a search word and the button «Пошук» («Search»).

- Setting of direct or inverse sorting on the vocabulary. In the Fig. 3 the inverse sorting is represented.

- Entering a new word and editing the existing ones in the vocabulary, deleting words from the vocabulary.

Dialog box, which is used for these functions is as follows (Fig. 4.).

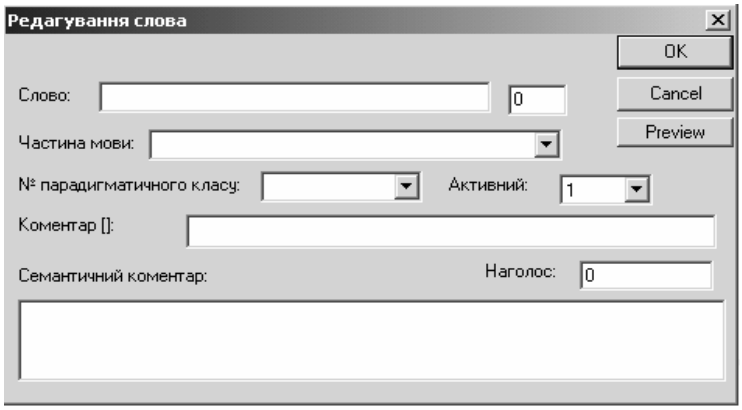


Fig.4. Dialog box for input and/or editing of a word

Deleting words from the vocabulary (the word is marked as deleted but remains in the LDB physically) is carried out by pressing the appropriate button on the panel to perform basic functions.

Assignment of other buttons on the panel: "P" - the "Paradigm" (the default is always active), the "T" - the function "Transcription" (in this version of this feature is not implemented).

The following buttons are designed to perform such functions: "Enter a new word", "Copy the selected word from the vocabulary" "Delete the selected word from the vocabulary", "Write the paradigm of a chosen word or a selected group of words to a text file", "Go to the editing of paradigmatic classes".

The next important functions of the DGD are the implementation of operations with paradigmatical classes:

• entering of a new paradigmatic class, editing, deleting of the existing paradigmatic classes (input of differential characteristics for a new PC, input and editing of quasi-flexions, etc). The entrance to the edit mode is realized by pressing the button «Paradigmatic classes» on the tools panel. The operations with paradigmatic classes (add, edit, delete) is carried out in the dialog box specially designed for this purpose. (Fig.5)

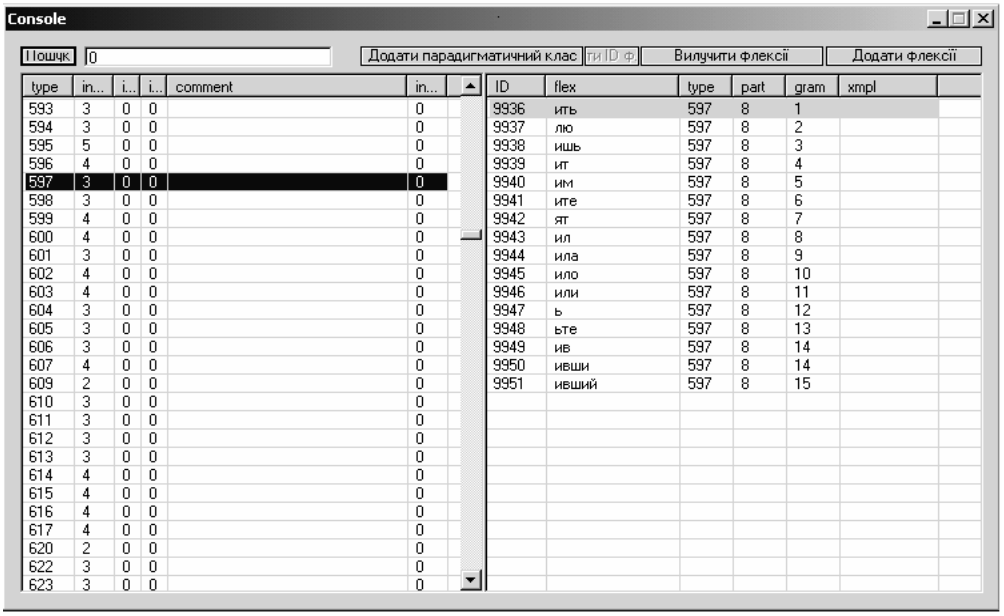


Fig.5. Dialog box for the work with paradigmatic classes

The dialog box (Fig.5) has three areas: the functional one, the paradigmatic classes area (left) and the quasi-flexions zone (right). The functional area is at the top of the window and contains controls for the functions with the paradigmatic classes:

The button «Search» is situated near the right text box to carry out the search of a paradigmatic class from the list of paradigmatic classes;

The button «Add paradigmatic class» is used to enter a new paradigmatic class. Pressing this button activates a dialog box to enter a new paradigmatic class (Fig 6.);

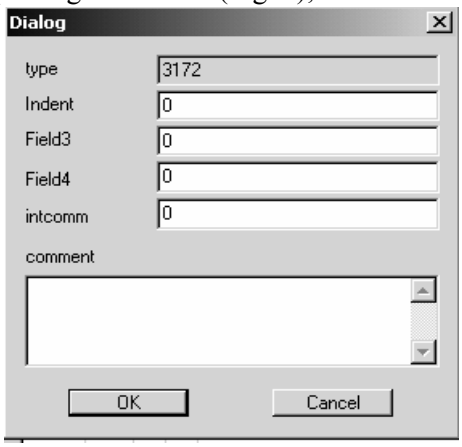


Fig. 6. Dialog box for entering a new paradigmatic class

The buttons «Delete quasi-flexion» and «Add quasi-flexion» provide the implementation of the appropriate functions of the selected paradigmatic class.

- Building vocabulary of quasi-stems (dictionary of quasi-stems is used in programs of morphological analysis).

Study of paradigmatical effects in the morphological word-inflexion system

Created grammatical LDB of Russian was used to study word-inflexion effects. In particular, it performe a study of the types of paradigm incompleteness and types of variability of the word-inflexion paradigm (WIP).

To indicate paradigms with missing word-inflexion forms we use the notion (concept) of word-inflexion paradigm incompleteness.

Paradigm incompleteness is characterized by the agency of incompleteness parameter *def*. This parameter is a list of numbers representing the paradigmatic forms that are not present in the paradigm.

Investigation of word-inflexion paradigm incompleteness types is implemented in the full scope of the digital grammatical dictionary created in the Ukrainian Lingua-information Fund of NASU.

The result of this investigation is the parametrization of the vocabulary according to the type of incompleteness and to the type of variability paradigms.

Types of paradigm incompleteness and types of variability have been identified by analyzing the table of quasi-flexions, which describes the paradigmatical classes in the database structure.

Here are the main results of this analysis.

There are 2015 paradigmatic classes (PC) in table of quasi-flexion of DGD. Among them, the 1207 PCs have a certain type of incompleteness, and 808 classes have the full paradigm (i.e., no defects: def = 0).

Distribution of paradigmatic classes for different grammatical classes, indicating the quantity of PCs with paradigm incompleteness and quantity of PCs with a full paradigm are showed in Table 1.

Table 1

Grammatical class	Quantity of PCs	Quantity of PCs with incompleteness	Quantity of PCs with full word-inflexion paradigm (def=0)
Noun	818	188	630
Adjective	132	44	88
Pronoun	37	5	32
Pronoun-adjective	31	2	29
Verb of imperfect aspect (нсв)	359	359	0
Participle	6	2	4
Verb of perfect aspect (св)	607	607	0
Cardinal number	25	0	25
Total	2015	1207	808

In general there are identified 99 types of paradigm incompleteness (TPI), among them: 15 TPIs of nouns; 8 – TPIs of adjectives, 4 – TPIs of pronouns, 2 – TPIs of pronouns-adjectives, 45 – TPIs of imperfective verb, 1 –TPIs of participle, 24 – TPIs of perfective verb. 808 paradigmatic classes are without a defect or have a full paradigm.

Types of paradigm incompleteness for the grammatical class of nouns are given in Table 2.

Table 2

#	Code of TPI	Parameter of incompleteness	Quantity PC with current TPI	Word example	Quantity of word with current TPI
1	2	3	4	5	6
d0		{0}	630	абажур	54086
d1	2	{1,2,3,4,5,6}	66	грабли	1110
d2	22	{1,2,3,4,5,6,7,9,10,11,12}	1	щец	3
d3	23	{1,2,3,4,5,6,8}	1	бразды	1
d4	24	{1,2,3,4,5,6,8,9,10,11,12}	1	тары-бары	2
d5	25	{1,2,3,4,5,6,8,9,11,12}	1	полсуток	1
d6	72	{2,3,5,6,7,8,9,10,11,12}	7	пламень	34
d7	76	{2,3,6,7,8,9,10,11,12}	1	полымя	1
d8	77	{2,5,6,7,8,9,10,11,12}	1	теля	1
d9	87	{3,4,5,6,7,8,9,10,11,12}	1	ведомо	1
d10	88	{3,4,5,7,8,9,10,11,12}	1	полдороги	1
d11	89	{3,5,7,8,9,10,11,12}	2	полслова	2
d12	91	{7,8,9,10,11,12}	92	богочеловек	8865
d13	92	{7,9,10,11,12}	1	зло	1
d14	93	{8}	7	башка	40
d15	94	{8,10}	5	брюзга	23

Parameter of incompleteness (column 3) contains a number of grammatical meanings, for which there is no word in the paradigm. For example, paradigm of lexeme «грабли» has only forms of plural (singular forms are not available in any case). In the corresponding row of Table 2 for the word «грабли» the numbers of grammatical forms with missing word-forms - 1, 2, 3, 4, 5, 6 are indicated.

Within each grammatical class of words for each type of paradigm incompleteness there are defined variants of paradigm fullness (VPF), among them some variative paradigms are revealed and the types of paradigm variability ascertained (TPV).

We have identified 253 variants of paradigm fullness (VPF), such as: 58 VPFs of nouns (42 types of variability(TPV)), 20 VPFs of adjectives (among them 4 VPFs without variability, 9 types of variability for adjectives with paradigm incompleteness, 7 types of variability for adjectives with a full paradigm). For other grammatical classes only the types of paradigm fullness were defined. Their variability has not been reviewed yet. Regarding the TPFs for the remaining grammatical classes such data are received: 9 VPFs for pronouns-nouns, 8 – VPFs for pronouns-adjectives, 3 – VPFs for participle, 93 – VPFs for imperfective verbs, 59 – VPFs for perfective verbs and 3 VPFs for cardinal numerals.

Examples of variants of paradigm fullness for grammatical class of nouns (Table 3) are given below. Variative paradigms are marked in bold.

Table 3

Variants of paradigm fullness and types of paradigm variability of nouns

№	Code of VPF	Description of VPF/ TPV												Quantity of paradigmatic forms	Example
		1	2	3	4	5	6	7	8	9	10	11	12		
1	2	3												4	5
1	39	0	0	0	0	0	0	0	1	0	0	0	0	1	щец
2	40	0	0	0	0	0	0	1	0	0	0	0	0	1	тары-бары
3	41	0	0	0	0	0	0	1	0	0	1	0	0	2	полсуток
4	56	1	0	0	1	0	0	0	0	0	0	0	0	2	пламень
5	68	1	1	0	0	0	0	0	0	0	0	0	0	2	ведомо
6	60	1	0	0	1	1	0	0	0	0	0	0	0	3	полымя
7	61	1	0	1	1	0	0	0	0	0	0	0	0	3	теля
8	69	1	1	0	0	0	1	0	0	0	0	0	0	3	полдороги
9	70	1	1	0	1	0	1	0	0	0	0	0	0	4	полслова
10	42	0	0	0	0	0	0	1	0	0	1	1	1	5	бразды
11	74	1	1	1	1	1	1	0	0	0	0	0	0	5	земля
12	43	0	0	0	0	0	0	1	1	1	1	1	1	6	зубы
13	**44**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**2**	**1**	**7**	**двери**
14	**45**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**2**	**1**	**1**	**1**	**1**	**7**	**грабли**
15	75	1	1	1	1	1	1	0	1	0	0	0	0	7	зло
16	**179**	**1**	**1**	**1**	**1**	**1**	**2**	**0**	**0**	**0**	**0**	**0**	**0**	**7**	**быт**
17	**193**	**1**	**1**	**1**	**2**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**7**	**Овен**
18	**220**	**1**	**2**	**1**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**7**	**мозг**
19	**190**	**1**	**1**	**1**	**1**	**2**	**2**	**0**	**0**	**0**	**0**	**0**	**0**	**8**	**Кия**
20	**224**	**1**	**2**	**1**	**1**	**1**	**2**	**0**	**0**	**0**	**0**	**0**	**0**	**8**	**ход**
21	**46**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**2**	**2**	**1**	**2**	**2**	**10**	**бубны**
22	79	1	1	1	1	1	1	1	0	1	0	1	1	10	брюзга
23	**229**	**1**	**2**	**2**	**1**	**2**	**2**	**0**	**0**	**0**	**0**	**0**	**0**	**10**	**Дулёво**
24	80	1	1	1	1	1	1	1	0	1	1	1	1	11	башка
25	92	1	1	1	1	1	1	1	1	1	1	1	1	12	майя
26	**155**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**13**	**зверь**
27	**156**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**13**	**чучело**
28	**159**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**13**	**пария**
29	**172**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**13**	**барин**
30	**180**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**13**	**горб**
31	**188**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**13**	**дитя**
32	**194**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**13**	**коса**
33	**221**	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**13**	**лишек**
34	**167**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**2**	**1**	**1**	**14**	**земгал**

№	Code of VPF	Description of VPF/ TPV												Quantity of paradigmatic forms	Example
		1	2	3	4	5	6	7	8	9	10	11	12		
1	2	3												4	5
35	**176**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**2**	**1**	**1**	**14**	**соболь**
36	**181**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**2**	**1**	**14**	**кость**
37	**189**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**14**	**деревце**
38	**214**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**14**	**азотобактер**
39	**218**	**1**	**1**	**2**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**14**	**Ия**
40	**225**	**1**	**2**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**14**	**снег**
41	**226**	**1**	**2**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**14**	**кишмиш**
42	**158**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**2**	**2**	**15**	**блоха**
43	**183**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**1**	**1**	**2**	**1**	**1**	**15**	**гроб**
44	**228**	**1**	**2**	**2**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**15**	**степь**
45	**169**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**1**	**2**	**2**	**16**	**сотня**
46	**184**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**2**	**1**	**2**	**1**	**1**	**16**	**год**
47	**215**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**1**	**2**	**1**	**2**	**2**	**16**	**серьга**
48	**230**	**1**	**2**	**2**	**1**	**2**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**16**	**кий**
49	**177**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**2**	**2**	**2**	**2**	**17**	**арба**
50	**182**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**2**	**2**	**1**	**2**	**2**	**17**	**тень**
51	**216**	**1**	**1**	**1**	**2**	**1**	**1**	**1**	**2**	**2**	**1**	**2**	**2**	**17**	**доска**
52	**232**	**1**	**2**	**2**	**1**	**2**	**3**	**1**	**1**	**1**	**1**	**1**	**1**	**17**	**метромост**
53	**235**	**1**	**2**	**2**	**2**	**2**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**17**	**Ланца**
54	**178**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**2**	**2**	**2**	**2**	**18**	**дембель**
55	**185**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**19**	**ветер**
56	**231**	**1**	**2**	**2**	**1**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**22**	**шприц**
57	**250**	**1**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**23**	**чуваш**
58	**253**	**1**	**4**	**2**	**1**	**2**	**4**	**2**	**2**	**2**	**2**	**2**	**2**	**26**	**мхи**

The description of VPF / TPV (column 3) is given as follows: for each grammatical meaning of noun (1-12) the quantity of words available in the word-inflexion paradigm is indicated. For example, the highlighted gray line in Table 3 (the type of VPF – 44) means that in the grammatical meanings (GMs) 1-6 word-inflexion forms are not available; the GMs # 7, 8, 9, 10 and 12 have one word-form each, and in the 11th GM there are 2 word-forms. Below, the Fig. 7 shows the word-inflexion paradigm of lexem “двери”, which is a representative of this TPV:

двери – существительное pluralia tantum

(двустворчатая дверь; дверь вообще)

Падеж	Единственное число	Множественное число
Именительный		Две́ри
Родительный		Дверей
Дательный		Дверям
Винительный		Две́ри
Творительный		дверьми́, дверя́ми
Предложный		Дверях

мн. <ж 2е>%

Fig. 7. Example of word-inflexion paradigm (TD = 2 and TV = 44)

As a result of this study, each paradigmatical class has received values of the TPI (types of paradigm incompleteness) and the TPV (types of paradigm variability) parameters. In such a way a new classification has been obtained. It can be represented as a scheme (Fig. 8).

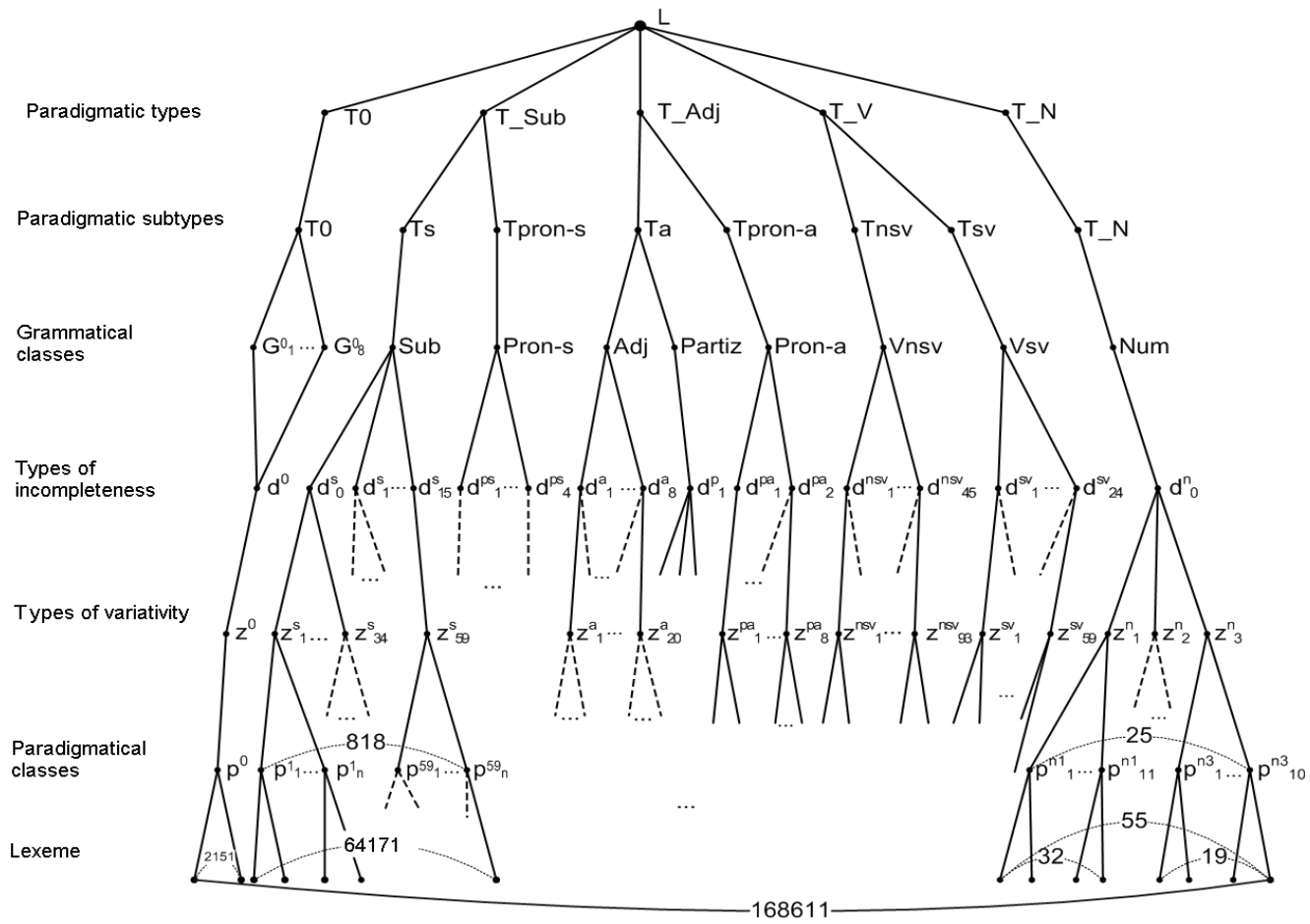


Fig. 8. Word-inflexion classification of the vocabulary of Russian, taking into account the parametrization of TPI and TPV

In this classification the additional parameterization of the vocabulary by the TPI and TPV is taken into account. This gives the researcher an additional possibility regarding the quantitative charasteristics of the types of paradigm incompleteness and of paradigm variability.

We believe that such a classification will be useful also for the morphological analysis in its algorithms, and that will ultimately contribute to improvement of its results especially in the disambiguation process.

Conclusion

The created digital grammatical dictionaries of Russian (Грязнухина Т.А., Любченко Т.П., Рабулец А.Г., 2002), the DGD of Ukrainian (Широков В.А., Рабулец А.Г., Шевченко И.В. , 2005), DEDs of German, English, and the corresponding software tools have been tested on large lexical arrays. They include; Russian – about 170 thousand units, German – more than 52 thousand (nouns, adjectives, verbs), English – about 20 thousand (nouns, verbs). For Spanish a computer experiment was based on verbs.

In conclusion I want to express my gratitude to prof. PhD V. Shirokov, T. Gryaznukhina, A. Rabulets and I. Shevchenko for usefull collaboration.

References

Грязнухина Т.А., Любченко Т.П., Рабулец А.Г. (2002). Электронная версия грамматического словаря русского языка (А.А.Зализняк) как инструмент автоматического морфологического анализа русского текста // Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных», Санкт-Петербург, март 2002. – С. 63-70.

Грязнухина Т.А., Любченко Т.П., Шевченко И.В. (2005). Новая версия электронного грамматического словаря русского языка с учетом акцентуации // Слово и словарь = Vocabulum et vocabularium: сб. науч. тр. по лексикографии / отв. ред. Рычкова Л.В. – Гродно: ГрГУ, 2005. – С. 188–193.

Зализняк А.А. (2003). Грамматический словарь русского языка: Словоизменение. – М.: Русский язык, 1978. – 878с.

Кнут Д.Э. (1977). Искусство программирования. – М.: Мир, 1977.– т.2.: Получисленные алгоритмы.

Любченко Т.П. (2006). Морфологічна модель словозміни флективної мови та електронний граматичний словник // Біоніка інтелекту: Наук.-техн. журнал. – 2006. – № 1 (64). – С. 72–77.

Любченко Т.П. (2007). Програмно-технологічні аспекти створення граматичних лексикографічних систем // Проблеми програмування. – 2007. – № 3. – С. 61–75.

Любченко Т.П. (2008). Моделирование морфологии естественного флективного языка // Бионика интеллекта: Науч.-техн. журнал. – 2008. – № 1 (68). – С. 52–64.

Любченко Т.П. (2008a). Один из подходов к построению классификации русской лексики // Прикладна лінгвістика та лінгвістичні технології: MegaLing'2007: Збірник наук. праць.- К., Видавництво «Довіра» – 2008. – С. 211–232.

Петровский А.Б. (1995). Метрические пространства мультимножеств // Доклады Академии наук. 1995, Т.344, №2, С. 175-177.

Петровский А.Б. (2000). Комбинаторика мультимножеств // Доклады Академии наук. 2000, Т.370, №6, С. 750-753.

Петровский А.Б. (2003). Пространства множеств и мультимножеств. – М.: Едиториал УРСС, 2003.

Петровский А.Б. (2004). Многокритериальное принятие решений по противоречивым данным: подход теории мультимножеств // Информационные технологии и вычислительные системы. 2004, №2, С. 56-66.

Широков В.А. та ін. (2005). Корпусна лінгвістика: Монографія / Широков В.А., Бугаков О.В., Грязнухіна Т.О., Костишин О.М., Кригін М.Ю., Любченко Т.П., Рабулець О.Г., Сидоренко О.О., Сидорчук Н.М., Шевченко І.В., Шипнівська О.О., Якименко К.М.; Український мовно-інформаційний фонд НАН України - К. : Довіра, 2005. – 472 с.

Широков В.А., Рабулец А.Г., Шевченко И.В. Электронный грамматический словарь украинского языка // Труды международной конференции «MegeLing'2005. Прикладная лингвистика в поиске новых путей". 27 июня – 2 июля 2005 г. Меганом, Крым, Украина. – С. 124-129.

A Frequency Dictionary of Finnish Word Building

Konstantin Tyschenko, Bogdan Rudyj
Linguistic Museum,
Taras Shevchenko National University Kyiv, Ukraine
cris-evol@ukr.net

Abstract
A new type of dictionaries is described in the paper – "a word-building frequency etymological learner's dictionary". It can also be called a table dictionary or a multi-dimensional dictionary with spatial visualization. The new type of dictionaries has been realized so far only for the Finnish language and published in 2004. New spatial organization of the dictionary's material gives a student a non-traditional tool of individual work for building his own vocabulary while learning the language. It can also stimulate new theoretical studies of the language lexicology.

Keywords: *dictionary, frequency, word-building, Finnish language, etymology, lexico-semantic field*

The idea to combine frequency and word-building dictionaries in a single coordinate space emerged in 1998 during the teaching Finnish a joint group of students of Kyiv Universities. It was the copy of the first 135 pages (5000 words) of the frequency dictionary of Finnish (compiled by Pauli Saukkonen's group[1]) favoured by Mr. Timo Rantakaulio (University of Helsinki) that enabled realization of the idea. Already during the first attempts to form word-building clusters a very important peculiarity of the Finnish lexical system became apparent. It turned out that the special "hermeticism" of the Finnish vocabulary[2] is largely compensated with its etymological self-sufficiency – continuity and fullness of word-building clusters which result from spontaneous original effusion of the Finnish vocabulary and which were brought to logical end by efforts of generations of Finnish linguists and personalities of culture[3]. Both mentioned traits are also characteristic of such an ancient European language as Cymric, the original word-building dictionary of which became another inducement of our work[4].

Pilot prints of the dictionary's previous stages comprised Finnish-Ukrainian lexical correspondences grouped alphabetically for educational reasons within frequency "steps" of the first 350-700-1100 frequent words[5]. Later this initial pattern, useful for building student's individual vocabulary, was enlarged for convenience to round numbers: 500-1000-1500-2000-3000-4000 words – and used as a horizontal parameter of a single descriptive table of Finnish word building[6]. This prototyping 4000-words dictionary was shown to Prof. Ilkka Savijärvi, Mrs. Varpu Pöntynen and other Finnish professors on language courses in Savonlinna in 1999. At the time, we obtained the copy of pages 136÷237 (frequent words from 5001 to 10419) of P.Saukkonen's dictionary and an up-to-date bilingual dictionary of Mr. H.Särkkä[7], that enabled extension of our work and its orientation to international reader through English translation.

The actual version of the dictionary contains on its 238 pages (119 broadsides) 10417 frequent Finnish words 8 in 1250 word-building clusters grouped in 1095 lexico-semantic fields. All Finnish words are provided with English translation and are divided for educational reasons into 10 columns (1000 words a column) sorted descending by frequency. Didactic sense of such division consists in that it splits the immense educational task of memorization of the Finnish lexical system as a whole into 10 portions – "steps" 1000 words in each. Mastering of these "steps" guarantees to a student an optimal way to recognition of ever larger percentage of words of an ordinary Finnish text 9: 65-74-79-82-84% for columns 1÷5 and approximately 86-87-88-89-90% for columns 6÷10. First 5000 frequent words (columns 1÷5) are to be found always on left pages, whereas the next 5000 words (columns 6÷10) – on right pages. Every pair of

pages – left and right – form one ***broadside***, a single spatial unity, alphabet-frequency two-dimensional space, onto which the word-building clusters are projected. Words of the ***first column*** represent the first thousand of the most frequent Finnish words, which constitute two thirds of an average Finnish text. All of them have very high10 though different frequency, ranging from f=23,796 down to f=58 in a compound 400,000-words text. The ***second column*** contains the words from #1001 to #2000 according to P.Saukkonen's list with frequency f=58÷25; likewise, the third column – f=25÷15, the forth – f=15÷10, the fifth – f=10÷7, the sixth – f=7÷6, the seventh – f=6÷5, the eighth – f=5÷4, the ninth – f=4÷3, the tenth – f=3. It is obvious that the limits of groups of words with identical frequency do not completely coincide formally with the didactically motivated division of the dictionary into ***thousands*** of words. Besides, beginning with f=5 every group of words with identical frequency exceeds one thousand members, – an excess, considering alphabetical sorting within each group, resulting in temporary advantage for those alphabetically higher words and creating an illusion of absence of one of the columns 7÷10 on pp. 11÷191. Nevertheless, finally the proposed division embraces all the necessary words up to f=3, that makes it quite applicable for pedagogical purposes.

Still for educational reasons we preserved distinct formatting for lower and higher half-thousands inside two initial thousands: in the first column the words from #1 to #500 are formatted **CAPITAL BOLD UNDERLINED**, and in the second column the words from #1001 to #1500 are formatted ***CAPITAL BOLD ITALIC***.

The most important innovation of the book has become integration of common base words belonging to different frequency columns within one horizontal band in a specially created alphabet-frequency coordinate space of the dictionary. This placed the dictionary on an essentially new level, laying foundation to a new generation of ***dictionaries-tables*** – spatially arranged dictionaries with several simultaneous but independent criteria of classification of lexical material. As to the structure of word-building clusters, along vertical axis they are sorted alphabetically inside as well as outside, whereas along horizontal axis – by descending frequency of words. Moreover, for fullness/completeness of a word-building cluster sometimes a word beyond P.Saukkonen's list is added between back-slashes (e.g., /AAPINEN/). Solid horizontal lines through both sides of every broadside separate the ***lexico-semantic fields***. Dashed horizontal lines separate word-building clusters from one another within one lexico-semantic field. At the end of every word set corresponding to certain letter, lexical internationalisms and proper names (~13% of the dictionary) are gathered.

Use of up-to-date software and programming during the work on the book led processing of the dictionary data to a more thorough level. First, it allowed us to formalize consequently the process of nestling and formatting the lexical material[11]. Besides, the correctness of clusterization was verified with the help of new three-volume etymological dictionary of the Finnish language[12], on the basis of which indication of the source language at each cluster's keyword became possible. Second, after the English translation for every Finnish word was done[13], the spelling check for English words was performed automatically[14]. Third, automation of ulterior calculations facilitated obtaining theoretically important quantitative characteristics of the word-building system of Finnish frequent words vocabulary. It developed that within the specified (by the dictionary's framework) lexical array there are only 818 unproductive words (they are grouped into 115 bands on grey). Quantity of lexico-semantic fields with 2 members is 331, with 3 members – 225, with 4 members – 163, with 5 members – 102, with 6 members – 64, with 7 members – 53, with 8 members – 37, with 10 members – 25, with 12 members – 20, with 16 members – 10, with 20 members or more – 75.

The discovered quantitative distribution shows that by its lexical productivity parameter the Finnish word-building system subdues to the well-known lexico-statistical law of Estoup-Zipf, being a new sphere of its manifestation[15]. Theoretical interest of primary importance have also general quantitative characteristics of the dictionary. For reasons of convenience three categories of words can be discerned, namely: productive, unproductive and international (comprising proper names and abbreviations), numbering ~8221, 818 and ~1378 respectively. It turned out that ~69% of the productive words are "mounted" out of 1250 base morphemes and some 50 affixes, whereas ~31% of them being the result of word composition[16]. This makes clear the importance of suggestion to a student to memorize foremost that initial thousand of productive words ("branches" of the Finnish lexical tree), from which then grow out 7000 "leaves" – derivations of the Finnish basic dictionary.

Other possibilities to apply the new matrix approach not only to study the word building, but also to optimize the description of other language aspects are demonstrated in the book's supplement with a few concise tables of objectively complicated Finnish declension and conjugation, – it is their maximal demonstrativeness and easiness to survey that makes them didactically valuable.

Thus, the reader takes in his hands a new type of dictionary: a dictionary-table or a multi-dimensional dictionary with spatial visualization. Its full name should be: "A descriptive word-building frequency Finnish-English learner's dictionary". New spatial organization of the dictionary's material gives a student a non-traditional tool of individual work for building his own vocabulary while learning Finnish. It can also stimulate new theoretical studies of Finnish lexicology.

References

1 Pauli Saukkonen, Marjatta Haipus, Antero Niemikorpi, Helena Sulkala. Suomen kielen taajuussanasto. A Frequency Dictionary of Finnish. – Porvoo, Helsinki, Juva: WSOY, 1979.

2 "The Finns, much as they treasure their own language, cannot carry it beyond their borders and expect to be understood. «Meillä on maailman yksityisen kieli – We have the world's most private language», a Finn once told me half proudly, half ruefully". (S.Andersen. Foreword / M.-H.Aaltio. Finnish for foreigners. – Helsingissä: Otava. – 19662. – P.9).

3 Cf. Lauri Hakulinen. Suomen kielen rakenne ja kehitys. Toinen osa. Sanasto- ja lauseoppia. – Helsinki, 1946. – Mostly chapters III, IV.

4 Gareth Jones. Gwreiddiadur Cymraeg. Welsh Roots & Branches. – Biwch, Powys. 1994. – 334 p.

5 К.М.Тищенко. Фінсько-український алфавітно-ранговий словник 1100 слів. Препринт. – Серія оптимізованих дидактичних посібників. Вип. 18.– Київ: ЛНМКУ, 1999. – 15 с.

6 Kostäntyn Tyščenko. Suomen kielen sananjohdon taajuuskirja. (1. – 4. tuhatta). Ennakkopainos. – Серія оптимізованих дидактичних посібників. Вип. 17. – Київ: ЛНМКУ, 1999. – 43 с.

7 Heikki Särkkä. Englanti-Suomi-Englanti yleiskielen käyttösanakirja. English-Finnish-English General Dictionary. – Helsingissä: Otava. – 19995. – 928 s.

8 The number is explained by the necessity to include all words with minimal frequency f=3.

9 Cf. P.Saukkonen e.a. Op. cit. – P. 38.

10 Ibidem. – P. 41-59.

11 Microsoft® Office 2000 (Excel, Word & Access). Microsoft® Visual Basic 6.0.

12 Suomen sanojen alkuperä. Etymologinen sanakirja. Osat 1-3. – Helsinki, 1992-2000.

13 H.Särkkä. Op. cit.; Suomen kielen perussanakirja. Osat 1-3. / R.Haarala päätoimittaja jm. – Helsinki: Oy Edita Ab, 19964; Otavan Tietosanakirja. – Helsinki: Otava, 1997; Tiekartta Suomi & Tietovihko. – Helsinki: WSOY, 1999.

14 Microsoft® Word 2000: "Automatic spelling check" function.

15 К. Тищенко, Б.Рудий. Закон Есту-Зіпфа у фінському словотворі. (У друці).

16 Cf. the ratio supposed by L.Hakulinen. Op. cit. – § 9.A.1.

III. Linguistic and Mathematical Foundations

Many-volume Contrastive Grammar of Bulgarian and Polish[1]

Violetta Koseska-Toszewa
Warsaw
amaz1312@gmail.com

Abstract
Bulgarian-Polish Contrastive Grammar (BPCG) is the world's first, and until now only, extensive attempt at semantic juxtaposition with a gradually developed interlanguage. BPCG consists of 9 volume issued in 12 books. The authors decided to lead the description in the Grammar in the direction from content to form. A semantic interlanguage allowed for developing two grammars of equal rank: a grammar of contemporary Bulgarian and a grammar of contemporary Polish. The analysis of semantic categories applied in BPCG provides a coherent contrastive description, independent of the fact whether the described languages possess grammatical exponents of meanings. BPCG is part of the contemporary trend of theoretical contrastive studies based on the logical theory of quantification, the contemporary theory of processes known as „Petri nets", and the theory of logical predicate-argument structures. Our studies eliminate strict divisions into grammatical and lexical levels, and thanks to this introduce many new observations of the examined phenomena. We have selected here universal semantic language categories important for description of the language that have not been elaborated until now, namely basic language categories, such as time, modality, definedness/undefinedness and semantic case, which have not been described exhaustively until now in academic grammars of Polish and Bulgarian. The order of description in this synthesis has not been determined based on the order of the existing BPCG volumes, but based on the generally accepted order of elements in the semantic structure of a sentence. The outermost element in the semantic structures of the sentence is its modal characteristics. The subsequent elements are time, quantifiers and their order in the semantic structure of the sentence, and predicate-argument positions. Hence the Synthesis is not a brief summary of the issues analysed in the volumes of BPCG. Its is a description of selected semantic categories, ordered according to the semantic order in the semantic structure of Polish and Bulgarian sentences.

0.0 History. Bulgarian-Polish Contrastive Grammar [Polsko-bułgarska gramatyka konfrontatywna] (from now, BPCG) consists of 12 volumes. Volume 6 contains four separate parts – monographs (BPCG 1988 – 2008). The whole issue of that academic grammar is not available in Poland. Its first four volumes have been printed in Bulgarian in Sofia, and the remaining 4 have been published in Polish. The last, 9th volume of the Grammar, devoted to word formation, has been submitted for printing. Two different places of publication make it more difficult for the BPCG to reach the reader, and the bilingual publication does not facilitate reception of the study. In 1976-1981, the group of the former South Slavic Languages Laboratory of the Institute of Slavistics, Polish Academy of Sciences, consisting of: Kazimierz Feleszko, Violetta Koseska-Toszewa, Małgorzata Korytkowska, Jolanta Mindak and Irena Sawicka, completed work on the Project of two contrastive grammars: Bulgarian-Polish and Serbo-Kroatian - Polish. This was the first Project in the world (at that time, not only for Slavic languages) of a semantic contrastive grammar based on logical and semantic studies of the juxtaposed Slavic languages (see Studia 1984). Bulgarian scientific experts occupied with criticism of the theoretical and methodological approach chosen by the Polish team included, among others, Svetomir Ivančev and Ruselina Nicolova, while on the part of

[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

Polish scientists such a function was performed by: Kazimierz Feleszko, Roman Laskowski, Kazimierz Polański and Janusz Siatkowski. Together with scientific editors, they were looking after the correctness and theoretical uniformity of the grammar (the Polish scientific editor of Volume 1 of BPCG was Violetta Koseska-Toszewa, and her Bulgarian counterparts were Svetomir Ivančev, and Jordan Penčev).

1.0 Grammar studies. Another publication issued simultaneously with preparation of the consecutive volumes of the BPCG was the series *Bulgarian-Polish Grammar Studies* [Studia gramatyczne bułgarsko-polskie], Vol. 1-7. (*Studia*, 1986–2003), devoted to theoretical and methodological discussions on grammar problems. As the seventh (last) volume of *Bulgarian-Polish Grammar Studies* was meant to be a kind of guide to the *Bulgarian-Polish Contrastive Grammar*, it contains two different language versions: Polish and Bulgarian (Косеска-Тошева, Балтова 2004).

2.0. Grammar (BPCG). The grammar begins with a volume devoted to the phonetics and phonology of both languages, which is followed by volumes devoted to selected semantic categories and the means for expressing them in Bulgarian and Polish. Today we can say that two grammars of equal rank have been developed: the grammar of contemporary Bulgarian and the grammar of contemporary Polish, joined with a semantic interlanguage and describing the following semantic categories in both languages:

[1.] definedness/undefinedness (V. Koseska-Toszewa, G. Gargov);
[2.] quantity (L. Krumova, R. Roszko);
[3.] degree (M. Čoroleeva, A. Petrova-Wasilewicz);
[4.] communicating person (I. Gugulanova, P. Barakova, M. Szymański);
[5.] case: selected types of predicate-argument positions in both languages (M. Korytkowska).
[6.] modality (V. Koseska-Toszewa, V. Maldžieva, J. Penčev) — description theory.
[6 a] imperceptive modality (M. Korytkowska, R. Roszko);
[6b] hypothetical, irreal, optative and imperative modalities (V. Maldžieva);
[6c] interrogative modality (M. Korytkowska);
[7.] semantic category of time (V. Koseska-Toszewa);
[8.] semantic category of aspect (S. Karolak).

2.1. Semantic juxtaposition. For linguists interested in general and theoretic problems, a novelty is **semantic juxtaposition** of two languages performed for the first time in the world using an **interlanguage** (*i.e., a system of notions based on logical and mathematical theories, which is a starting point for the juxtaposition of the languages under examination*).

The employed method of juxtaposing languages based on an interlanguage developed in line with the progressing research guarantees obtaining reliable, comparable results of research for arbitrary juxtaposed languages. This method of analysing semantic categories assures a consistent contrastive description. It should be stressed that BPCG is the first grammar to treat certain issues previously overlooked in academic grammars of Polish and Bulgarian. For Polish, this includes, among others, an exhaustive study of the semantic category of time and aspect, of diverse modal categories – of which next to nothing was previously known, of the definedness / undefinedness category, as well as the categories of quantity and communicating person. The case with Bulgarian, which was treated equivalently to Polish using the interlanguage, is similar. Hence an attempt to juxtapose an analytic language with a synthetic one has succeeded.

3.0 Semantic theories. The semantic volumes of BPCG are based on logical and mathematical theories: quantification theory, contemporary theory of processes known as Petri nets, theory of logical predicate-argument structures useful for describing natural language. We understand semantics like in the works with a "direct approach to semantics" by B. Russell and H. Rasiowa, and in the later ones — on the situation semantics by Barwise and Perry (Rasiowa 1975, Russell 1967, Barwise, Perry 1983, Cooper 1996). Regardless of the trend to which a given theory of linguistic meaning belongs, such a theory should take into consideration information, understood broadly and intuitively as the ability of certain fragments of reality to change the state of human consciousness. According to a more precise definition, the notion of information is identified with „an abstract quantity which may be stored in certain objects, sent between certain objects, processed in certain objects and used for controlling certain objects, whereby objects are understood to mean living bodies, technical devices, or systems of such objects " (Mazurkiewicz 1970). Each semantic theory

should also take into consideration the relation between *knowledge and information* consisting in the fact that all processes which change our knowledge are carriers of information. As to the notion of *knowledge,* maybe we should reconcile ourselves with B. Russell's thesis that this notion is imprecise, and merges with what we mean by „probable opinion". The character of the above relations is most often shown on semantically and structurally „simple" examples, such as: *W Warszawie pada deszcz* [It is raining in Warsaw].

An analysis of such a sentence looks as follows: assuming that the speaker uttering that sentence (S) is not consciously lying, the recipient (R) will first learn something about the reality (about objects and about systems of such objects). Secondly, the recipient (R) will also learn something about the consciousness of the speaker uttering the sentence (S) – namely, that the speaker uttering the sentence (Z) claims that it is raining in Warsaw. The second type of knowledge provided by the uttered sentence is very essential for the act of communication. If the recipient (R) fails to accept the second type of knowledge, he/she will not be able to treat the uttered sentence (S) as a source of knowledge – which means that the act of communication will not be performed. Besides information, contemporary linguistic semantic theories recognise also another notion, which we will term here the classification ability of natural language. The ability of a language to distinguish between (classify) states of the real world is most often used in the meaning theories in the opposite direction, i.e. the language fragments themselves are classified. The classification reduces to distinguishing two classes of expressions:

1. class of language expressions referring to states of the real world, and
2. class of expressions referring to our consciousness.

Classes of linguistically discernible states of the material world form the domain of reality fragments relevant to the language. These are objects, their properties and relations between or among them. Following traditional logic, properties and relations form the notions of predicates, whereby properties represent unary predicates, binary relations – binary predicates, etc. In turn, classes of linguistically indiscernible states of consciousness consist of all the notions that belong to the given class for each type of classification. Traditionally, following Locke (Locke 1948), they are commonly termed "ideas". Hence semantic theories can also be understood as theories of classification capabilities of natural language.

3.1. The differences between meaning theories with regard to understanding and interpretation of the essence of meaning allow us to distinguish three groups of such theories:

3.1.1 In the first group we can place the semantic theories whose proponents stress the discernment ability of the language, while seeing the capability of the language to distinguish between states of the real word as its secondary ability, derived from the former one. Here these theories will be referred to a semantic mental theories. Following Locke, it has been generally accepted that words can replace real word objects only because ideas themselves can replace (symbolise) real objects. The basic methodological difficulty in this type of theories is the inherent impossibility of classifying the notions of individual consciousness. The second, equally complicated problem, is the issue of the relation between ideas and the real world, which reduces to the assumption that it is only the ideas rather than words that can mean something. The drawback of those theories is that they reduce reasoning to unending regress consisting in the fact that ideas symbolise ideas which symbolise ideas which... etc. Eventually, a level of reasoning is postulated where ideas directly symbolise real world objects. As a result, we come back to the issue of the relation between ideas and the real world. It is such methodological difficulties that make many researchers abandon semantic mental theories in contemporary linguistics.

3.1.2 The second group of semantic theories consists of theories interested in the external capabilities of natural language to denote the real world. **Words group around the way in which they describe the real world rather than around the idea that they express.** Those semantic theories are commonly termed **theories of direct relation to semantics.** The said direct relation to semantics consists in the assumption that the relation between the language and the real world **is not a problem.** The central point of the theory is the **language-based classification** of the real world, which obligatorily implies a **certain classification of consciousness phenomena (ideas).** In the composed sentence: *Kazik powiedział mi przed chwilą, że Marta z reguły śpi do dwunastej* [Kazik has just told me that Marta as a rule sleeps until noon.], the clause: *że Marta z reguły śpi do dwunastej* [that Marta as a rule sleeps until noon] does not reflect external objects but the state of Kazik that consists in his thinking about the fact that „Marta z reguły śpi do dwunastej" [Marta as

a rule sleeps until noon]. Hence a direct relation to semantics allows us to take into consideration also „ideas". The scholars subscribing to that semantic trend were also interested in the known but nontrivial fact that the same language forms (words, expression constructions, sentences) can be carriers of quite different information. So while in other semantic schools the meaning of a sentence is defined using two abstract objects: (with Frege, these are *truth* and *falsity*), the second trend of semantic theories discussed here introduces into the definition of the meaning of a sentence also the notion of a *situation*, defining the meaning of a sentence as a set of abstract situations (Barwise, Perry 1983).

3.1.2.3 A standard and already classic example of theories with direct relation to semantics are Bertrand Russell's denotation and description theories (Russell 1967), used in the second volume of the Bulgarian-Polish contrastive grammar (Koseska, Gargov 1990), and the model theory of first order predicate logic, which, as an extensional one, is applied to natural language with substantial limitations, see (Rasiowa 1975).

3.2 The third trend of semantic theories. With great simplification, one can say that the third trend of semantic theories developed around Frege's works and his criticism of the proponents of the direct relation to semantics. Frege charged the direct approach to the semantics of natural languages with **lack of strict separation of *denotation, expression* and *sense.*** (Frege 1892). According to Frege, the meanings of names like e.g. „Evening Star" and „Morning Star" cannot be separated for they denote the same object (denotation), namely the planet Venus. Frege noted also that natural language contains sensible expressions which do not denote anything from the real world. These observations have led to extending semantic theories of the third class with the class of notions as a necessary one: *the class of senses*. In Frege's opinion, the coherence of semantic theories of natural language was determined not only by the classes of ideas and objects, but also by the class of their relations and structure: the class of senses. Hence Frege initiated one of the directions of logical semantics, later known as *intensional logic*. That direction has also emerged in linguistics as a direction of formal semantics of natural languages; its fullest presentation can be found in the works of R. Montague (Montague 1975).

3.3. There is no doubt that when using the two descriptive expressions „Evening Star" and „Morning Star" we are speaking of Venus in two different ways, and that the above expressions differ with respect to their meaning. And since the expressions have different meanings, then - as G. Rylle says – "Venus, the planet described with these two expressions, cannot be the meaning of those expressions". To support this thesis, G. Rylle quotes works of John Stuart Mill, "who acknowledges this openly, and takes it into consideration" (Rylle 1967). The expression we have selected as an example here – „Evening Star" and „Morning Star" – widely discussed in the literature, are, in our opinion, *classifiers of various states of our consciousness*.

3.4. The above reasoning is especially important for us since the Bulgarian-Polish contrastive grammar adopts a description methodology close to the second trend of the meaning theory for natural languages presented in this chapter, see (Koseska, Gargov 1990). One of the reasons motivating the Grammar's authors to do so were. B. Russell's denotation and description theories. In turn, situation semantics, as J. Barwise and J. Perry acknowledge in their works, is close both to the ideas of B. Russell and A. Mostowski and to the intuition of linguists, especially those occupied with a functional grammar, see (Barwise, Perry 1983).

3.5. We think that the extensionality principle, which says "that the truth or falsity of any statement regarding a theorem "P" depends only on either the truth or falsity of "P" itself, and that the truth or falsity of any theorem containing a propositional function depends only on the extension of that function, i.e. on the series of values for which that propositional function is true" (Rasiowa 1975), is not the only important rule. And neither do we neglect the old Karnap's idea, which emphasises the *extensionality/intensionality* dichotomy. "Take, – B. Russell writes, – for example, "A believes that P". It is obvious that a man can believe in some theorems and not believe in others, so the truth of "A believes that P" does not only depend on the truth or falsity of P". (...) (B. Russell 1959: 128, 129).

The Grammar is based on situation semantics as a theory not only consistent with B. Russell's one, but also close to Petri net theory, which is a theory with direct approach to the natural language semantics. In BPCG, Petri net theory is used as a basis for describing such semantic categories of the language, as definedness / undefinedness, temporality and modality, and hence all other semantic categories that can be described using the notions of state, event, discrete process, and quantification of states and events.

4.0 Description direction. The authors have decided to develop the description in the direction from the content to the form (from semantic structures to formal structures). The converse approach, in the

form → content direction, still very frequent in linguistic papers, does not describe the problems precisely and exhaustively enough, since forms and formal structures are as a rule ambiguous in each natural language. This requires a strict separation of the language form from its meaning. Our experience showed that contrastive description of the form → content type would not be fully valuable for it would reduce to describing one language with help of another. The latter approach is used in most of the known contrastive grammars, which describe one language (most often, the foreign one) using another (the native one). However, the juxtaposed languages cannot be treated equally without trying to develop a semantic interlanguage.

4.1. Interlanguage. As I have already mentioned, the *Bulgarian-Polish Contrastive Grammar* is the world's first, and up to now only, extensive attempt at semantic juxtaposition with a gradually developed interlanguage. Though development of an interlanguage different form both juxtaposed languages has been postulated, this was a theoretical requirement difficult to meet in practice due to the need of isolating the basic semantic categories comprising the structure of the interlanguage. The interlanguage should consist of empirical notions discovered in the course of simultaneous studies of at least two languages. The task of constructing the interlanguage would be impossible to perform if only formal structures of both languages were examined. The interlanguage emerges as a product of theoretical contrastive studies, and represents a system of notions taken from selected mutually consistent theories describing the juxtaposed languages. Along with the progressing studies, the interlanguage develops gradually, and is enriched with new notions. We think that the most important rule in its development is the requirement that the interlanguage be developed based on theories not leading to a contradiction. For example, when developing basic semantic units used for describing the linguistic definedness/undefinedness category in the interlanguage, one can employ either reference theory or defined description and quantification theory. However, both theories cannot be used simultaneously, for this leads to internal contradiction in the notion system of the interlanguage. Already the second volume of BPCG (Koseska, Gargov 1990) clearly implies that a description that takes as a starting point the Bulgarian formal means of the language is quite different from a description that takes as a starting point the formal means of Polish. This is determined even by the more extended morphological plane of the means expressing the notions of definedness and undefinedness in Bulgarian as compared to Polish (see also Koseska, Mazurkiewicz 1988). The interlanguage we know from Volume 2 of BPCG, especially with respect to the notions related to quantification of time, is developed further in Volume 7. The interlanguage related to juxtaposing Polish and Bulgarian in the area of the semantic definedness/undefinedness category (Koseska, Gargov 1990) is based on the assumption on the quantificational character of that category. The basic notions, such as uniqueness (of an element or a set) could be written down using the language construction of the iota operator, existentiality – using an existential quantificational expression, universality – using a universal quantificational expression, etc. Volume 2 of BPCG (Koseska, Gargov 1990) undertakes the first attempt to implement the conception of a language juxtaposition using an interlanguage. In the subsequent volumes of the BPCG, the interlanguage is extended with notions related to modality and the semantic category of time, thanks to including the contemporary theory of processes known as Petri nets (Mazurkiewicz 1986, Koseska, Mazurkiewicz 1988).

5.0. Cognitive approach. If, for example, we want to describe the content of universal quantification in the semantic structure of a Bulgarian sentence, we should take into consideration the language phenomena visible on the morphological level: definite article, but we should also indicate universally quantifying lexems, both on the nominal phrase level and on the verbal phrase level, e.g. Bulgarian *всеки* 'each'*, винаги* 'always'. A strict separation of morphological, syntactic and lexical levels would prevent a comprehensive description of semantic phenomena. Hence it is worth stressing that our studies of theoretical semantics eliminate strict divisions into grammatical and lexical levels, bringing a lot of new observations regarding the examined phenomena. Such an approach is termed cognitive here. On the one hand, we understand cognitive studies as theoretical semantic studies which allow us to take into consideration language means from various levels: grammatical and lexical ones, seen as a single whole. On the other hand, when necessary, we use broader language situations, where the phenomena we are interested in are understood by language users in an unambiguous way. Such situations always take into consideration also the states of language users and their attitude to the communicated contents.

6.0. Semantic category of time. Examples. The examples selected will only concern description of the semantic category of time in both languages (Koseska 2006?). In Volume 7 of BPCG, the semantic category of time is described using the net model instead of the linear one. Net theory was first adapted to description of temporal and modal phenomena in natural language by A. Mazurkiewicz (Mazurkiewicz 1986), and later by V. Koseska-Toszewa and A. Mazurkiewicz (Koseska, Mazurkiewicz 1988). Petri nets (Petri 1962) are a tool independent of the existing natural languages, and so indifferent with respect to them. Their simplicity (they are based on just three primary notions: of a state, event and their mutual succession), combined with a considerable expressive power, predestines them for the role of a theory forming the tertium comparationis (interlanguage) in contrastive studies of natural languages. We adopt the notions of state and event as fundamental units of time description. The two notions are distinguished based on the time spread of states and the momentary character of events. States continue, while events can only happen. An abstract counterpart of the above distinction is the difference between a section of the real line (state) and a point lying on that line (event). The notion of a process is represented in nets by a configuration of states and events joined by the precedence-succession relation. By way of example, the meanings of praeterite forms in Polish and Bulgarian can be written down in the net notation as follows:

1. Event which has occurred before the speech state.

In Bulgarian, such temporal content is expressed by the aorist of perfective verbs, and in Polish — by praeteritum of perfective verbs:

Щъркелът се върна в гнездото си., Bocian wrócił do gniazda [The stork has returned to the nest.] –

(Net paraphrase: The event „the stork's return to the nest" has occurred before the speech state).

2. Unique configuration of a state and an event.

In Bulgarian, this type of temporal content is expressed by the aorist of imperfective verbs, and in Polish – by the praeteritum of imperfective verbs. In both languages, those verbal forms can be only accompanied by unique quantifying expressions, see sentences of the type:

Той точно тогава боледува от грип., On właśnie wtedy chorował na grypę. [He was sick with flu just then.]

(Here we have to do with a unique configuration of a state and an event).

3. Multiple occurrences of the same combination of a state and an event.

These contents are expressed by Bulgarian aorist of imperfective verbs and Polish praeteritum of imperfective verbs. In this case, the language situation is associated with a quantitative rather than scope quantification, see:

Тази седмица той ходи пеша няколко пъти до центъра на града., W tym tygodniu on kilka razy chodził pieszo do centrum miasta. [This week he has gone on foot to the city centre several times.]

4. Combination of states and events which have occurred before the speech state.

The Bulgarian imperfectum form of imperfective verbs emphasises states continuing in the past, while the aorist of imperfective verbs emphasizes events which occurred in the past and broke the described states. These subtle differences in meaning are revealed by differences in quantification characteristic for the Bulgarian aorist, which when formed from either perfective or imperfective verbs acts as a placeholder for the unique quantifier only; while the imperfectum of imperfective verbs acts as a placeholder for all types of scope quantification. Polish separates those meanings by selecting quantifying expressions which occur together with the praeteritum form of imperfective verbs. See examples of the type:

Той понякога намираше време за разходка., On od czasu do czasu znajdował czas na spacer. [He found time for a walk from time to time.]

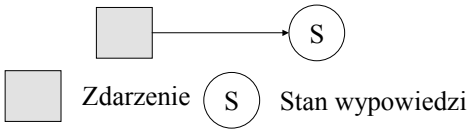
5. A state continuing before the speech state connected with an event and a state coexisting with the speech state.

In Bulgarian, such content is expressed by the perfectum form. In Polish, this meaning is expressed by the praeteritum form, which often occurs next to the praesens form. The past state expressed is not finished before the speech state, as is the case when aorist of perfective verbs is used. The result following the past state is still valid during the speech state:

Той е боледувал от грип (и сега още кашля)., On chorował na grypę (i teraz wciąż kaszle). [He has been sick with flu (and is still coughing now).]

Below we present schemata of nets showing the difference in the meanings of Bulgarian aorist and perfectum, together with Polish counterparts.

Aoryst od dokonanych

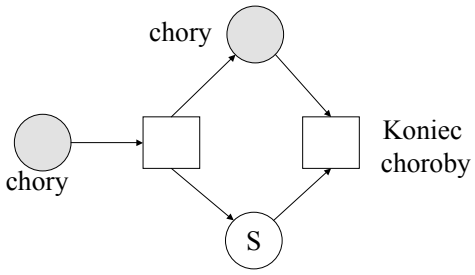


(zachorowanie)

Toj se razbolja ot grip
On zachorował na grypę

Legend:
Aoryst od dokonanych – Aorist of perfective verbs
Zdarzenie – Event
zachorowanie – falling sick
Stan wypowiedzi – Speech state
On zachorował na grypę – He has fallen sick with flu.

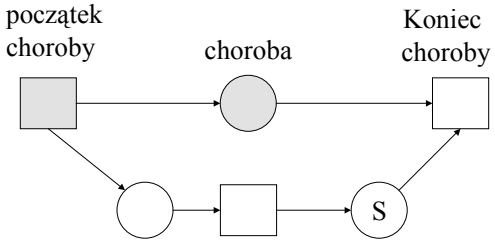
Perfectum od niedokonanych



Toj e boleduval ot grip
On chorował (i wciąż jeszcze choruje) na grypę

Legend:
Perfectum od niedokonanych – Perfectum of imperfective verbs
Początek choroby - start of sickness
Koniec choroby- end of sickness
choroba – sickness
On zachorował na grypę (i nadal choruje)– He has fallen sick with flu (and is still sick).

Perfectum od dokonanych



Toj se e razboljal ot grip
On zachorował na grypę (i nadal choruje)

Legend:
Perfectum od dokonanych – Perfectum of perfective verbs
Początek choroby - start of sickness
Koniec choroby- end of sickness
choroba – sickness
On zachorował na grypę – He has fallen sick with flu.

A net description facilitates understanding of temporal and modal phenomena in both Bulgarian and Polish unknown in the subject literature until now. Thanks to this, the myth of a relatively simple system of temporal meanings (tenses) in Polish has been rejected. Up to now, nobody has described quantificational meanings of time in Polish. However, as the Bulgarian material shows, without that description, we would be unable to distinguish between the uses of aorist of imperfective verbs and imperfectum of imperfective verbs. Despite the identical information on the aspects of the aorist and imperfectum forms of imperfective verbs, the two forms provide different temporal information. In both cases, we have a combination of states and events which have occurred before the speech state. However, using the imperfectum form, we emphasise states continuing in the past, while in case of the aorist of imperfective verbs we emphasise events which occurred in the past and broke the described states. These subtle differences in meanings are emphasised by the quantificational differences characteristic for the aorist. As we have already mentioned, the aorist, independently of the verb aspect, i.e. both of imperfective and perfective verbs, is a placeholder for the unique quantifier only, while the imperfectum of imperfective verbs is a placeholder for all kinds of scope quantification. The above fact explains the different distribution of both the forms in Bulgarian (see Koseska, Mazurkiewicz 1988, Koseska, Gargov 1990). As the aorist of both aspects of a verb can only express uniqueness, it is a self-contained, independent carrier of that quantificational meaning. In turn, the imperfectum of imperfective verbs is not a form independently expressing quantification. The imperfectum can be a placeholder for both universal and existential quantification. Though this is rare, we can also encounter it in contexts with uniquely quantified temporal information. The imperfectum always expresses quantification of states, and never of events, see: *Тя седеше пред прозореца.*, where the imperfectum may, depending on the completion of quantification, express universal quantification: *Тя винаги седеше пред прозореца., Ona zawsze siedziała przy oknie*. [She always sat at the window.]. It can also express existential quantification, like in the sentence: Тя понякога седеше пред прозореца., Ona czasem siedziała przy oknie.[She sometimes sat at the window.] As I have mentioned above, the imperfectum of imperfective verbs is exceptionally encountered also in a meaning analogous to praesens forms in contexts of the type: *В точно този момент, той я обичаше., W tej właśnie chwili on ją kochał*. [At just that moment, he loved her.] In that case, unique quantification refers to a past state continuing during the situation chosen as the only one (just at that moment...).

From the comparison of the uses of the Bulgarian aoryst and imperfectum, we can see that the aorist of both perfective and imperfective verbs expresses only quantificational uniqueness of events and states, while the imperfectum of imperfective and perfective verbs can express both existentiality and universality of events and states, but also (though very rarely) uniqueness of states, like the praesens form. It is worth stressing that Polish more often than Bulgarian copes with expressing temporal meanings using lexical means which are quantifiying expressions (and hence not only using verbal forms). When we want to express the temporal meaning of resultative perfectum, we need two Polish verbal forms rather than one, like in Bulgarian. Though all elements of temporality can be expressed in both languages, it is worth noting that some temporal meanings would not have been noticed in Polish without its juxtaposition with Bulgarian. We should emphasise the immense importance of the definedness / undefinedness opposition for understanding the semantic category of time expressed by quantification of time, and the fact that in Polish it can concern aspect and time, while in Bulgarian first of all time (see Koseska, Korytkowska, Roszko 2007).

7.0. Synthesis of Bulgarian-Polish Contrastive Grammar, or Polish -Bulgarian Contrastive Grammar. The synthesis of the Grammar is to make the results of the many year's work of a numerous international team of its authors, comprising linguists and logicians dealing with natural language problems, to the people interested in this subject (Koseska, Korytkowska, Roszko 2007). The swap of the languages in

the title of the synthesis is not accidental – it emphasizes the equal status of the language material in both the languages described using the interlanguage.

We should stress that this is a presentation of a new approach to many important theoretical issues, as well as a presentation of many problems which up to now have not been studied at all, or have been studied to an insufficient extent only, due to the absence of a contrastive perspective of description for Polish. Such an approach is valuable in teaching the language, as well as in translations to Polish and from Polish. The specialists in Polish studies should be interested in a new, yet unknown description of Polish as seen from the perspective of another language. The Slavists will complement their knowledge of the semantics and the specifics of the systems of the two languages which belong to different Slavic groups (south Slavic and west Slavic). For linguists interested in general and theoretical problems, a novelty will be a semantic confrontation of two languages carried out for the first time using an interlanguage. The synthesis is not a collection of selected issues. We have selected here universal semantic language categories important for description of the language that have not been elaborated until now, namely basic language categories, such as time, modality, definedness/undefinedness and semantic case, which have not been described exhaustively until now in academic grammars of Polish and Bulgarian. *The order of description in this synthesis has not been determined based on the order of the existing BPCG volumes, but based on the generally accepted order of elements in the semantic structure of a sentence*. The outermost element in the semantic structures of a sentence is its modal characteristics. The subsequent elements are time, quantifiers and their order in the semantic structure of the sentence, and predicate-argument positions. Hence the Synthesis is not a brief summary of the issues analysed in the volumes of BPCG. Its is a description of selected semantic categories, ordered according to the semantic order in the semantic structure of Polish and Bulgarian sentences. The synthetic chapter on the interlanguage placed at the end of the volume is very important for understanding the theoretical conception of the grammar. The interlaguage is the language of notions used as basis for parallel description of the phenomena in both languages.

REFERENCES

Barwise Jon, Perry John (1983): *Situations and Attitudes*. — Bradford Books, MIT.

Cooper Robin (1996): The Role of Situations and Generalized Quantifiers. — [w:] Shalom Lappin (ed.): *The Handbook of Contemporary Semantic Theory*. — Oxford.

Frege Gottlob (1892) : *Über Sinn und Bedeutung*. — Zeitschrift. fűr Phil. und phil. Kritik, 100, 25–50.

BPCG (1988 – 2008)

BPCG-1 1988: I. Sawicka, T. Bojadzhiev, Bylgarsko-polska sypostavitelna gramatika,

tom 1. Fonetika i fonologija, Sofija.

BPCG-2 1990: V. Koseska-Toszewa, G. Gargov, Bylgarsko-polska sypostavitelna gramatika, tom 2. Semantichnata kategorija opredelenost/neopredelenost, Sofija.

BPCG-3 1994: L. Krumova-Cvetkova, R. Roshko; A. Petrova, M. Choroleeva, Bylgarsko-polska sypostavitelna gramatika, tom 3. Semantichnite kategorii kolichestvo i stepen, Sofia: 15--38.

BPCG-4 1993: I. Gugulanova, M. Shimanski, P. Barakova, Bylgarsko-polska sypostavitelna gramatika, tom 4. Semantichnata kategorija komunikant, Sofija.

BPCG-5-1 1992: M. Korytkowska, Gramatyka konfrontatywna bułgarsko-polska, t. 5. cz. 1. Typy pozycji predykatowo-argumentowych, Warszawa.

BPCG-6-1 1995: V. Koseska-Toszewa, V. Maldzieva, J. Penchev, Gramatyka konfrontatywna bułgarsko-polska, t. 6. cz. 1. Modalność. Teoretyczne problemy description , Warszawa 1996.

BPCG-6-2 1997: M. Korytkowska, R. Roszko, Gramatyka konfrontatywna bułgarsko-polska, t. 6. cz. 2. Modalność imperceptywna, Warszawa.

BPCG-6-3 2003: V. Maldzieva, Gramatyka konfrontatywna bułgarsko-polska, t. 6. cz. 3.

Modalność: hipotetyczność, irrealność, optatywność i~imperatywność, warunkowość, Warszawa.

BPCG-6-4 2004: M. Korytkowska, Modalność interogatywna --- pytania o

rozstrzygnięcie, Warszawa.

BPCG-7 tom 2006: V. Koseska-Toszewa, Gramatyka konfrontatywna bułgarsko-polska, t. 7. Semantyczna kategoria czasu, SOW, Warszawa.

BPCG-8 tom 2008: S. Karolak, Gramatyka konfrontatywna bułgarsko-polska, t.8. Semantyczna kategoria aspektu, SOW 2008

BPCG-9 tom (w przygotowywaniu do druku) : Ju. Bałtova, W. Małdziewa, Gramatyka konfrontatywna bułgarsko-polska, t. 9. Słowotwórstwo.

Koseska – Toszewa Violetta (2006): *Semantyczna kategoria czasu, Gramatyka konfrontatywna bułgarsko-polska* (BPCG), t. 7 — Warszawa: Slawistyczny Ośrodek Wydawniczy SOW.

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

Koseska – Toszewa Violetta, Mazurkiewicz Antoni (1988): Net Representation of Sentences in Natural Languages. — [in:] *Lecture Notes in Computer Science* 340. *Advances in Petri Nets 1988*, Sprinter-Verlag, 249–266.

Lock (1948): Collected Works. Vol. 1, 2, London.

Mazurkiewicz Antoni (1970) Problemy języków formalnych w automatycznym

przetwarzaniu informacji [w:] Problemy przetwarzania informacji, Warszawa
1970, s.~15--62.

Mazurkiewicz Antoni (1986): Zdarzenia i stany: elementy temporalności. — [w:] Studia gramatyczne bułgarsko-polskie, t. I, Temporalność, Wrocław, 7–21.

Mazurkiewicz Antoni (in print): A formal description of temporality (Petri net approach).

R. Montague 1974: Formal Philosophy. Selected Papers of R. Montague. New Haven, 1974.

PetriI Carl. A. (1962): Fundamentals of the Theory of Asynchronous Information Flow. — [in:] Proc. of IFIP'62 Congress, Amsterdam: North Holland Publ. Comp.

Projekt (1984): Projekt gramatyki konfrontatywnej bułgarsko-polskiej i serbskochorwacko-polskiej, Wstęp. – [w:] Studia polsko-południowosłowiańskie, Wrocław, red. Kazimierz Polański.

Rasiowa Helena (1975): Wstęp do matematyki współczesnej. – Warszawa: Wydawnictwo Naukowe PWN.

G. Rylle 1967: Teoria znaczenia [w:] Logika i język, red. J. Pelc, Warszawa, 1967, 485--535.

B. Russell 1959: Mój rozwój filozoficzny, Warszawa, 1959, s.~128, 129, 145.

Russell Bertrand (1967): *Denotowanie, Deskrypcje.* [w:] *Logika i język*, Warszawa.

Studia (1984): Studia konfrontatywne polsko-południowosłowiańskie, pod red.

Kazimierza Polańskiego, Wrocław.

Studia (1986–2003): *Studia gramatyczne bułgarsko-polskie*, t. I (1986) – t. III (1989) Wrocław; t. IV (1991) – t. VII (2003), pod red. Violetty KOSESKIEJ – TOSZEWEJ i in. – Warszawa.

Косеска-Тошева Виолета, Гаргов Георги (1990): Семантичната категория определеност/неопределеност, Българско-полска съпоставителна граматика (= BPCG), т. 2, София.

Косеска-Тошева Виолета, Балтова Юлия, ред. (2004): Българско-полски граматични студии, Справочник по академичната Българско-полска съпоставителна граматика. — София: Академично издателство Марин Дринов.

Заимов Йордан (1982): Супрасълски или Ретков сборник. Увод и коментар на старобългарския текст. – София.

Net-based Description of Modality in Natural Language (on the Example of Conditional Modality)[1]

Violetta Koseska[1], *Antoni Mazurkiewicz*[2]
[1]Institute of Slavic Studies of PAS
amaz1312@gmail.com
[2]Institute of Computer Science of PAS

Abstract
The intention of the present paper is to show how the Petri nets formalism can be applied for explaining not only temporal but also modal properties of sentences in natural languages. A special attention has been paid for distinguishing courses of actions with forking (that creates different, but coexistent courses) from branching (that creates different and mutually exclusive courses). It is argued that conditional sentences cannot be represented properly by means of logical implication; instead, for this representation the net description is proposed. Examples serve to show how Petri nets can be viewed as a universal tool (an intermediate language) for analyzing and comparing different natural languages.

0.0. **The semantic model of description** of modality in a natural language can be based on the basic notions of Petri net theory, see (Koseska, Mazurkiewicz 1988) and (GKBP, vol. VI-VII, 1990 – 2007). The net-based representation of time and modality is a significant extension of Reichenbach's conception of tenses, and hence it is rather a generalization than negation of that conception. In other words, each temporal situation expressed using Reichenbach's schemata can be represented using nets, while not every situation expressible with nets can be represented in Reichenbach's model (Koseska, Mazurkiewicz, 1988, 1991 and Mazurkiewicz 1986).

1.0. In the net-based representation of an utterance, we talk about **states**, **situations, events and histories.** *Local states* represent certain momentary properties of objects being the subject of utterance; *global states* consist of states of all such objects. *Events* cause a change in the state of some object or several objects, which gives the net-based description a dynamic character, varying over time. The course of events expressed by an utterance forms a *history,* representing mutually interrelated dependence among states and events. In the net-based approach to description of such processes, the paradigm of a *state* is its continuance. Each states continues for some specific time. Two different states following one another are separated by some *event* which begins the new state and ends the old one. The event, which represents a change, does not continue; it only occurs at a certain point of time. By a *situation* we shall mean here a certain fragment of reality which might encompass part of the past, the present and the future of the states of some objects. All utterances of a temporal character will refer to such situations.

1.1 Temporal utterances describe some situation, i.e. they talk about dependencies which appear in the temporal course of events and states. We will describe situations with help of Petri nets [...]. Analysis of an utterance must take into consideration the speaking subject's position with respect to the uttered situation. In Reichenbach's schemata, that position corresponds to a point on the timescale, in Petri nets it is a state of some object, namely the subject of utterance, from now on referred to as the observer. The position of the **observer** describing the situation will correspond to the so-called *moment of speech*. Due to the impreciseness of the latter term, in our description it has been replaced with the term „**state of utterance".** The state of utterance is a position occupied by the observer, i.e. the sender of information (or the *speaker* in terms of traditional grammar); hence the state of utterance determines all states possible in the present (the present situation), and indirectly also all states and events possible in the future and in the past. Knowing 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

(hypothetical or actual) course of the described events and states, we can draw conclusions regarding situations which take place in the discussed fragment of the changing reality.

1.2. It is worth adding that the above basic notions of the net-based description of time are also associated with other notions, which are important for the semantic interlanguage connected with the net theory:

Local state – state of certain special objects discussed in the utterance.

Global state – state of all objects determining the situation.

Accessible state – state reachable in the future with respect to the observer's situation. A global state consists of the states of all objects in some situation, as opposed to a local state, which refers to one or several objects of that situation. For example, a local state for the objects "doors, windows" will be "the doors are closed", while "the doors are closed, and the windows are open" will be a global state. We can say that a global state is a special case of a local state: namely, it is local state that encompasses, as we have mentioned in the foregoing, all objects of the discussed situation, in opposition to a local state, which encompasses either one or some of them. Events occur locally, i.e. they change local states. If we want to describe the real world in a natural language, we must refer in it to local states; modal phenomena in a natural language reflect effects of the local character of states. This implies the need for the description methods to take into consideration the local character of states. According to the principles of net-based description, a given local state can be assigned a set of global states –namely, all the states which are compliant with that local state in the given fragment of the described reality.

C**urrent state** is in turn a state containing the state of utterance, and other states coexisting with it.

P**ast state** is a state whose consequences include, among others, the state of utterance.

F**uture state** is a state being one of the possible consequences of the state of utterance;

P**resent state** means the same as current state.

State of utterance is the state of the information sender, and it determines the temporal position of the observer, i.e. the sender of information (or "speaker" in terms of traditional grammar). Hence the state of utterance determines all states possible in the present, and indirectly also all states and events possible in the future and in the past.

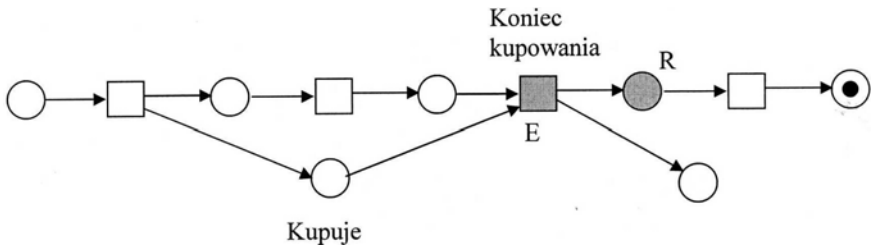
Permanent state is a state that neither has been initiated nor will be broken by any event. However, such states, describing constant laws of nature, are of no importance in the dynamic aspects of temporal and modal situations in a natural language which are described here.

1 3.Below we present a number of net schemata which show some temporal situations we are interested in. Following Reichenbach, we mark with the letter E the event or state under discussion, with the letter R – the reference state that our utterance refers to, and – customarily – we place a dot in the place denoting the state of utterance.

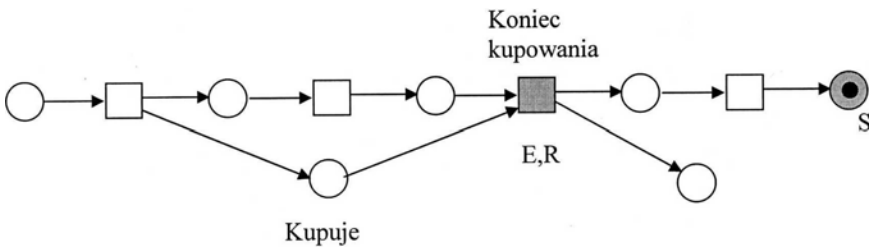
Legend:

Kupuje	She is shopping
Koniec kupowania	End of shopping

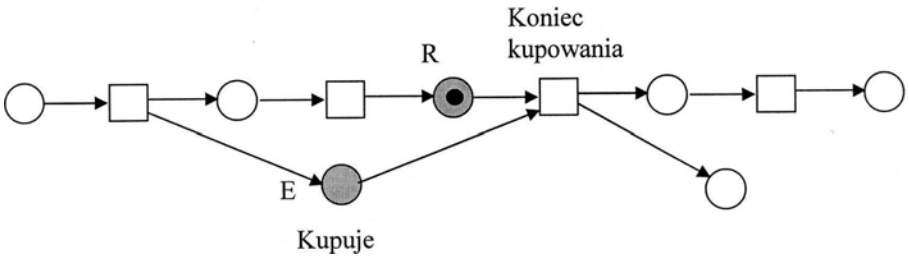
Ona zrobiła zakupy [She did the shopping]:



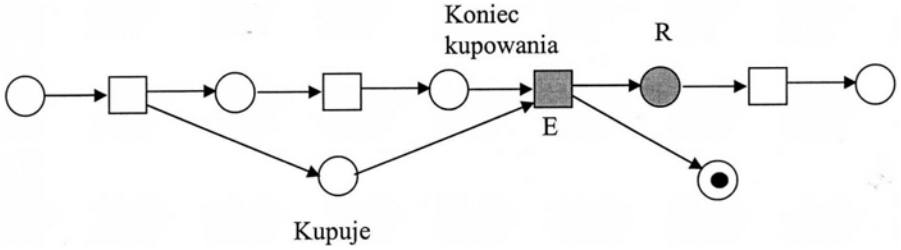
Ona skończyła robić zakupy [She has finished doing the shopping]:



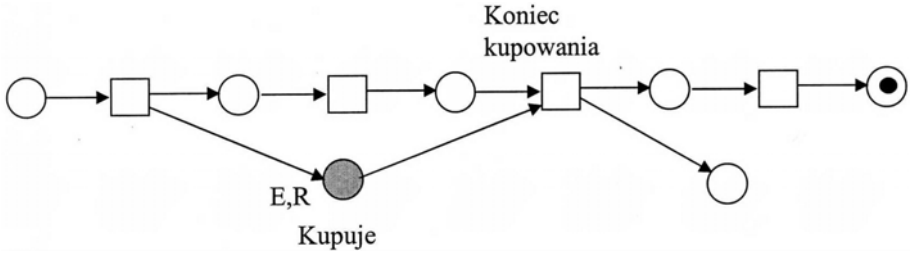
Ona robi zakupy [She is doing the shopping]:



Ona ma zrobione zakupy [She has the shopping done]:



Ona robiła zakupy [She was doing the shopping]



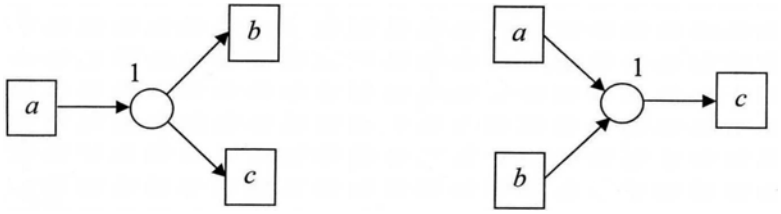
The figures show the nets representing the discussed situations.

In the first net, the analogue of the Polish *Ona zrobiła zakupy* is the Bulgarian Tja napazaruva. In the second net we refer to the event being the end of the state "kupuje" preceding the state of utterance. We have there the aorist of imperfective verbs. In the second net, the analogue of the Polish *Ona skończyła zakupy* is the Bulgarian Tja sv'arshi da pazaruva. In the third net, the analogue of the Polish *Ona robi zakupy* is the Bulgarian Tja pazaruva. In that net, our speaking and her doing of the shopping are concurrent. In the fourth net, the analogue of the Polish *Ona ma zrobione zakupy* is the Bulgarian Tja e napazaruvala. In that net, we refer to the state initiated by the event ending "kupowanie", and we have there the perfectum of perfective verbs. In the fifth net, the analogue of the Polish *Ona robiła zakupy* is the Bulgarian Tja e pazaruvala. In that net, we refer to the state "kupowanie", which was initiated before the state of utterance. We have there the perfectum of imperfective verbs. It should be noted that the first net contains an occurrence of Bulgarian aorist of perfective verbs used for denoting the event which occurred before the state of utterance. In the fourth net, in Bulgarian we have an occurrence of Bulgarian perfectum with the function of ascertainment. In the fifth net, perfectum has a resultative meaning; for details see Koseska – GKBP, vol. VII. As the net description implies, in such nets two different elements can occur simultaneously, and hence two different states can coexist. Moreover, two events can be executable independently, and events can be executable during the existence of some state.

Branchings and forks. The descriptions and schemata given above referred to a single history, and the differences followed from different mutual positioning of the observer's position, events and states. In reality, we sometimes speak about certain variants of the future and the past, and then the observer's position must be positioned in some way with respect to these variants. In the net-based description, the description of such situations follows in a natural way from the specifics of the net-based description itself.

2.1 **Branchings**. A net may contain *branching*; if they exist, they represent different possibilities of its course. The branchings may be unconditional in the sense that the choice of one or another exit event is not conditional upon anything. To explain the idea of net-based description of temporal situations, below we present simple nets illustrating basic cause-effect relationships. In both net schemata, there are three events, *a*, *b* and *c* connected with

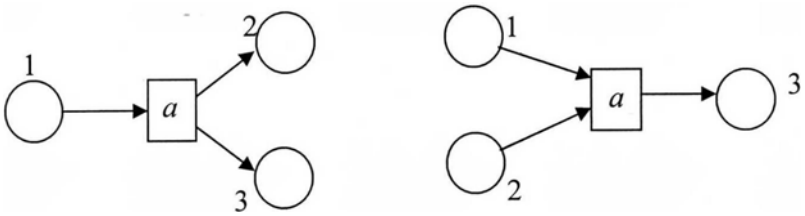
some common state 1. This state in Example (1) begins with event *a*, and ends with one of (mutually exclusive) events *b*, *c*. In Example (2), the state begins with one of (mutually exclusive) events *a, b,* and ends with event *c*.



(1) (2)

Schema (1) expresses the situation where occurrence of event *a* is a necessary condition for occurrence of one of events *b, c* – without occurrence of *a*, occurrence of any of them is not possible. However, the schema does not imply that occurrence of *a* causes occurrence of *b*, because we can have a course of events where occurrence of *a* will cause occurrence of the event *c*, excluding *b*. Schema (2) describes a situation where occurrence of event *a* is a sufficient condition for occurrence of event *c*; the schema also implies that another such condition is occurrence of event *b*, with *a* and *b* being mutually exclusive. This is because we can have a course of events where *c* is preceded by *b*, but event a does not occur, which means occurrence of event *a* is not a necessary condition for occurrence of event *c*.

2.2 **Forks**. Another schema of a situation, in some sense dual to branching, is the so-called *fork*, where one event starts (or ends) a number of co-existing states. Examples might include e.g. end of a railway journey, which ends both the state of travelling and the state of remaining in a railway car; or beginning of a sickness, which also begins the state of fever. The simplest examples of forks are presented in Schemata (3) and (4). In Schema (3), event *a* ends state 1 and begins two coexisting states 2 and 3, which represent the beginnings of two independently running histories. In Schema (4), event *a* begins state 3 and ends two coexisting e states 1 and 2, which representing the final states of two independently running, but not mutually exclusive, histories.



(3) (4)

In the situation presented in Schema (3), states 2 and 3 are consequences of state 1; state 1 is a necessary condition for occurrence of any of states 2 and 3. In Schema (4), occurrence of both states 1, 2 is a necessary condition for occurrence of state 3; if any of them fails to occur, event *a* cannot occur either, and hence state 3 cannot begin.

2.3. Summing up, by a fork in the net we mean a *situation* where one event can begin or end more than one state, and is characteristic for a parallel course of one or more components of the system, not colliding with each other.

By a branching in the net we mean a *situation* where one state can begin or end with more than one event. A branching is a characteristic feature of nets describing a situation with the possibility modality and corresponds to what has been characterized in logic using the functor „it is possible that". A branching represents a choice among a few mutually exclusive possibilities. A branching and a fork in the net are the main sources of the possibility modality in natural language sentences. Two or more events directly following some state are a phenomenon characteristic for a branching, while two or more states following one event are characteristic for a fork.

In the above examples, the states of utterance have not been marked, because the cause-effects relationships do no depend there on one or another position of that state, GKBP, 1995, Vol. VI, Part 1.

3.0. The problems connected with sentences involving the possibility modality require pointing out a certain fact which is important for understanding the essence of the theory we are using. Understanding "conflicts", or "branchings", as a synonym of "negation" would be a gross misunderstanding – and we have met with such remarks in the discussions among linguists over the net-based description method and its applications to studies of modality in a natural language. The essence of **conflict** is choice between two (or more) mutually exclusive possibilities, while negation is a logical functor, and as such has a static character; however, resolution of a conflict (choice) has a dynamic character (like each net-based description), and affects the subsequent course of things. This remark applies especially to the description of conditionality in a natural language.

In case of a conflict, the history presented with the net can run in different ways, depending on the circumstances or the choice made. Such a choice determines one of a few possible, but mutually exclusive, continuations of action. Using a net, we can describe future and past consequences of choices already made in the past or those which can be made in the future. The possibility of a branching in the net offers means for describing conditional sentences, see (Koseska, Mazurkiewicz 1988).

3.1. Most of the works describing the semantic structure of conditional sentences base it on the so-called conditional, and deliberate on the relationship between the conditional and logical implication, see: (Pelc 1986). The problems of implication and conditional have been the subject of numerous papers and discussions in logic, which concentrated mainly on the so-called implication paradox. The problem emerged when implication was read using the words *if...then,* see (Quine 1955). In a natural language, the above expression as a rule has a broad range of different meanings. The problem reduces to the question whether logical implication can be "read" as the conjunction *if...then.*

3. 2. Implication, conditional, or something else? Most of the authors involved in the discussion agree that the truth of a natural language sentence does not necessarily depend on the truth of the succedent and the antecedent, as is the case for logical implication. Ajdukiewicz points out the fact that in each natural language the meaning of the conjunction *if...then* is only close to, but different from the meaning associated by contemporary logic with the notion of implication, and that certain natural language sentences become false after replacing the implication symbol with the conjunction *if...then*. This refers to the theorems, writes Ajdukiewicz, which are connected with the fact that implication is true whenever either its antecedent is false or its succedent is true. This is because in logic implication is false only if the antecedent is true, and the succedent – false (Ajdukiewicz 1956).

3.3. The divergences between the meaning of implication and the meaning of a conditional from a natural language gave rise to the question whether it is appropriate to analyse reasoning in a natural language using the notion of implication. The problem of the implication paradox led to disputes on the limits of applicability of logical methods to natural language studies (Kotarbiński 1958). In connection with difficulties in interpreting a sentence of the type : *If p, then q,* the linguistic literature on that subject also put forward a question whether it is appropriate to analyse reasoning in a natural language using the notion of implication (Bogusławski 1986), (Banyś 1989). For example, A. Bogusławski refuses to equal conditional sentences with either material or strict implication in any sense or mode (Bogusławski 1986).

3.4. It is worth stressing that most of the scholars acknowledges that there is a "dynamic relationship" between *p* and *q,* i.e. the two components of a conditional. The most appropriate approach in the net-based description seems to be representation of the conditional „if – then" by the cause-effect relationship (Petri,62). Subject to the reservation that we do not list all meanings of *if...then, "*and indirectly connections between *p* and *q,* i.e. between events or states of things referred to, respectively, by conditional sentence *p,* or antecedent of the conditional, and by the main clause *q,* or its succedent", Pelc lists several ways of interpreting the conditional *if p, then q,* see for example:

1. Causal relationship, e.g. *If you eat too many carbohydrates, then you will grow fat*;
2. Sign relationship, e.g. *If he has rash, then he is sick with scarlet fever*.
3. Special cases of a universal relationship, one of which is formal logical implication (Pelc 1986: 272).

4.0. The **quantifier *only*** occurs in the semantic structure of an expression rather than in its surface structure, where it can be „incomplete". In our opinion, in a natural language we do not have to do with the expression *if p, then q,* but rather with an expression with the meaning: *only if p, then q* or *if only p, then q.* A proof for the fact that the „only" quantifier always occurs in the semantic structure of such a natural language sentence, though it does not necessarily appear in the surface structure of that type of sentence, is negation of the logical expression *if p, then q*, see the sentence: *If it's raining, I'll take an umbrella* and its logical negation: *It isn't raining and I'll take an umbrella.* The logical negation of *„If it's raining, then I'll take an umbrella" is „it is raining and I won't take an umbrella".* The truth of the above implication is guaranteed by the following situations:

(a) It's raining and I'm taking an umbrella
(b) It isn't raining and I'm not taking an umbrella
(c) It isn't raining and I'm taking an umbrella.

When in a natural language we say „ *If it's raining, then I'll take an umbrella*", our intention is to exclude possibility (c), i.e. we really have in mind the formulation „Only if it rains, then I'll take an umbrella". This sentence uttered in a natural language is true in the following situations:

It's raining and I'm taking an umbrella,

It isn't raining and I'm not taking an umbrella,

i.e. it describes the logical equivalence "p if and only if, when q". Let us note that, formally, "I'll take an umbrella only if it's raining" means the same as "if I take an umbrella, then it's raining", and it is true in the following situations:

I'll take an umbrella and it's raining,

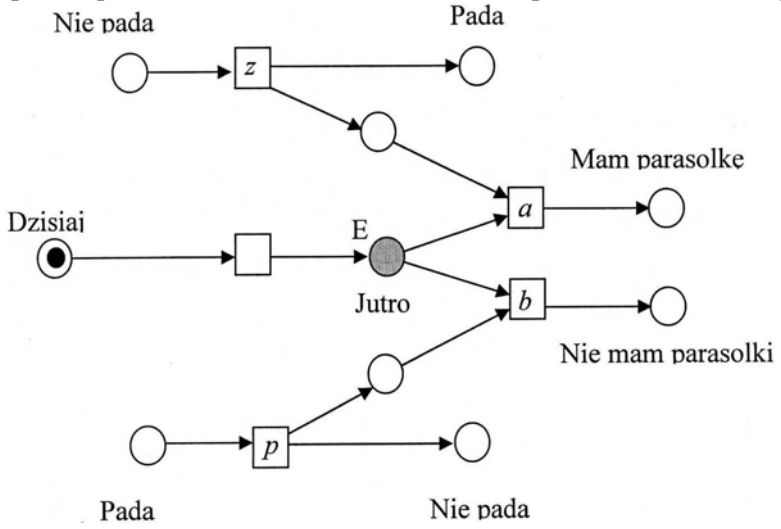
I won't take an umbrella and it's raining,

I won't take an umbrella and it isn't raining.

The (logically) correct equivalent formulation of "I'll take an umbrella only if it's raining" is "I'll take an umbrella when it's raining and only when it's raining". In reality, it is not the speaker's intention to talk about the situation of somebody who can see that it isn't raining, that the sun is shining, and who is nevertheless taking an umbrella. Accordingly, the natural situation is completed by the sentence: *It isn't raining and I'm not taking an umbrella*. However, it is also a negation of the semantic structure of the sentence: *if I take an umbrella, then it's raining*, i.e. *I'll take an umbrella only if it's raining*. The latter sentence can occur without the surface unique quantifier: *only*, see the sentence: „If it rains, I'll take an umbrella", when by default we have „*only*".

4.1. In Petri nets, a **history** describes a sequence of transformations of states through occurrence of events; in each history, the relation between states and events is a cause-effect relation (Mazurkiewicz 1986). Without defining precisely, what we understand by cause and effect, we can assume that we know how to understand a cause-effect relation. It is a temporal relation. The states representing a condition for event occurrence (appearing "before" the event) represent its **cause,** and the states following from them (appearing "after" the event) are their **effect.** In the net theory, connections of states and events are underlain by a relationship corresponding to the cause-effect relation rather than to the notion of implication. The cause for occurrence of some event in a Petri net is the occurrence of all states constituting the causes of that event, and the effect of an event is the occurrence of all states constituting the effects of that event, see schemas (3) and (4) concerning a necessary condition and a sufficient condition (Mazurkiewicz, Koseska, 1991.)

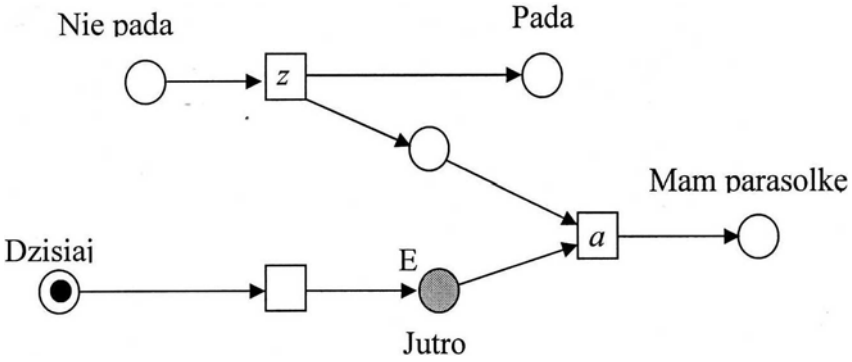
Example. Let us examine the sentence: *If it rains tomorrow, I'll take an umbrella*, and the net corresponding to that sentence, presented in the figure below. The example is rather expanded due to further considerations, which are outside the scope of the present problem area; it will then serve as an explanation for other language phenomena too.



Legend:

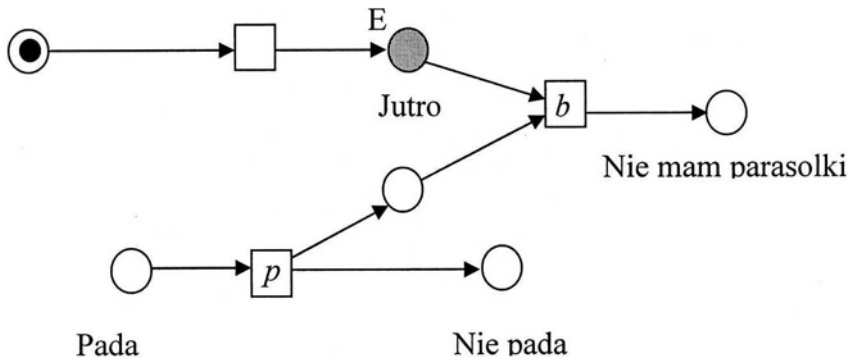
Nie pada	It isn't raining
Pada	It's raining
Dzisiaj	Today
Jutro	Tomorrow
Mam parasolkę	I have an umbrella
Nie ma prasolki	I don't have an umbrella

In the figure we have a net describing a situation which explains the meaning of the utterance *If it rains tomorrow, I'll take an umbrella*. The net describes two mutually exclusive histories:



History 1 („It's raining")

and



History 2 („It isn't raining")

In the net we have marked the state of utterance ("today"), the state of decision-making (E), as well as events *z* and *p* (it starts and stops raining), and *a* and *b* (I'm taking and not taking an umbrella). Moreover, the states "it's raining" and "it isn't raining", which according to the laws of logic exclude each other, are also marked; we assume that the states of raining and not raining change cyclically (alternate). Let us note that occurrence of event *z* (it starts raining) as well as of event *p* (it stops raining) is independent of the state of utterance: it can occur either before or after, or else during the state of utterance. The decision on taking an umbrella is made under the influence of those events; if event *z* occurs in the history, then I have an umbrella and it's raining; if *p* oocurs, then I don't have an umbrella and it isn't raining. Moreover, let us note that the change of the state *dziś* (today) to the state *jutro* (tomorrow) is effected by an event independent of the states *pada* (it's raining) and *nie pada* (it isn't raining). The above net shows both states and events appearing explicitly and those appearing implicitly, as well as possible relations among them. It is, in our opinion, a good representation of the semantic structure of the conditional sentence: *If it rains tomorrow, I'll take an umbrella.*

4.2. Nets describe transformations of states by events and their mutual relations, determined in the net theory by the cause-effect relation. The cause-effect relation is always a temporal one. Hence the net-based description of conditionality allow us to use the notions of state, event and cause-effect relationships. The net-based interpretation of conditionality refers to previous states, previous events, as well as states and events following the former as a result of the cause-effect relation. Logical implication is an indispensable tool of formal deduction, leading always from true premises to true conclusions, but it says nothing about cause-effect dependencies – and they are exactly what we have to do with in the conditional sentences of a natural language.

5.0. Nets and interlanguage. A interlanguage, necessary for juxtaposing different languages, in particular Polish and Bulgarian, within the time and modality categories is a semantic tool, see GKBP, 1990 – 2007. Petri Nets, through their universality and independence of natural languages, are a perfect candidate for an interlanguage. This theoretical tool reveals language phenomena sometimes overlooked by linguists. Proponents of the net-based description discover on the example of conditional modality more and more of new possibilities provided by the net-based description of natural language, which often fundamentally diverge from the tradition. The net-based description of modality and time allows us to capture the most

important semantic features of various types of modality, such as conditionality, imperceptiveness, or "hypotheticality". In this paper, we have captured conditionality through the constructions of net branching and the cause-effect law, connecting states and events. The modal and temporal problems concerning conditionality discussed here a novelty with respect to the problems connected with conditionality already discussed in the Polish–Bulgarian contrastive grammar, see (GKBP, 2007).

REFERENCES

Ajdukiewicz 1956: K. Ajdukiewicz, Okres warunkowy a implikacja materialna, [w:] Język i poznanie, t.~2: Wybór pism z lat 1945--1963, Warszawa.

Banyś 1989: W. Banyś, Theorie semantique et SI...ALORS. Aspects semantico\=logiques de la proposition conditionelle.

Bogusławski 1986: A. Bogusławski, Analiza zdań warunkowych a problem funkcji semiotycznych, [w:] Studia semiotyczne XIV--XV, Warszawa, s.~215-- --224.

BPCG (1990 – 2007)

BPCG-2 1990: V. Koseska-Toszewa, G. Gargov, Bylgarsko-polska sypostavitelna gramatika, tom 2. Semantichnata kategorija opredelenost/neopredelenost, Sofija.

BPCG-6-1 1995: V. Koseska-Toszewa, V. Maldzieva, J. Penchev, Gramatyka konfrontatywna bułgarsko-polska, t. 6. cz. 1. Modalność. Teoretyczne problemy opisu , Warszawa.

BPCG-6-2 1997: M. Korytkowska, R. Roszko, Gramatyka konfrontatywna bułgarsko-polska, t. 6. cz. 2. Modalność imperceptywna, Warszawa.

BPCG-7 tom 2006: V. Koseska-Toszewa, Gramatyka konfrontatywna bułgarsko-polska, t. 7. Semantyczna kategoria czasu, SOW, Warszawa.

Koseska – Toszewa, Korytkowska, Roszko 2007:

Koseska – Toszewa Violetta, Korytkowska Małgorzata, Roszko Roman *Polsko* -- bułgarska gramatyka konfrontatywna. — Warszawa: Wydawnictwo Akademickie Dialog 2007.

Koseska, Mazurkiewicz (1988):

Koseska – Toszewa Violetta, Mazurkiewicz Antoni: Net Representation of Sentences in Natural Languages. – [in:] Lecture Notes in Computer Science 340. Advances in Petri Nets 1988, Sprinter-Verlag, 249–266.

Косеска-Тошева, Гаргов (1990):

Косеска-Тошева Виолета, Гаргов Георги (1990): Семантичната категория определеност/неопределеност, Българско-полска съпоставителна граматика (= BPCG), т. 2, София.

Quine 1955: W. Quine, Mathematical Logic, Cambridge, Mass.

Kotarbiński 1957: T. Kotarbiński, Wykłady z dziejów logiki, Łódż.

Mazurkiewicz, Koseska 1991: A. Mazurkiewicz, V. Koseska-Toszewa, Sieciowe przedstawienie temporalności i modalności w zdaniach języka naturalnego, [w:] Studia gramatyczne bułgarsko\=polskie, t.~IV, Modalność a inne kategorie językowe, Warszawa, s.~7--25.

Mazurkiewicz 1986: A. Mazurkiewicz, Zdarzenia i stany: elementy temporalności, [w:] Studia gramatyczne bułgarsko\=polskie, t.~I, Temporalność, Wrocław, s.~7--21.

Pelc 1986: J. Pelc, Jeżeli, to\/, [w:] Studia semiotyczne XIV--XV, Wrocław, s.~271--287.

Petri 1962 C. A. Petri, Fundamentals of the Theory of Asynchronious Information Flow, [in:] Proc. of IFIP'62 Congress, North Holland Publ. Comp., Amsterdam.

Statistical methods for text analysis and comparison[1]

Maxim Krygin
Ukrainian Lingua-Information Fund
National Academy of Science of Ukraine
maxus@zeos.net

Abstract
The article presents an application of the statistical methods for the real language problems solution. The theoretical background is formulated for research and models. Several examples are given.

Keywords: *Statistical portrait, text analysis, transition probability, grammatical value, text corpora.*

1. Introduction.

The statistical methods in science can be generally used if the large amount of different data is under consideration or if some widespread phenomenon should be examined on the relatively small amount of data. Our methods allow us to get statistical characteristics of language phenomena for their recognition in the text or to make conclusion on presence and functioning in some text specimen. The objective of statistical natural language processing is to get the mathematical description of the language facts and phenomena and the connections between them to obtain some models of the natural language for solving various linguistic problems.

The first task is to define the language phenomena and the facts for analysis. Probably, this problem is one of the most difficult because the correct choice of objects for examination exerts the impact on the further research results. In our work we distinguish levels of language objects: character, phonetic, lexical, grammar, etc.

The main objective for research of the character system of language is the modelling language on the character level. It is not only simple calculation of distribution for characters and their combinations in the text. Statistical investigation of the phonetic system of language, for example, can show the degree of similarity between the phonetic system and the orthographic one.

The most difficult and varied are the statistical investigations of lexical, grammatical and syntax systems of the natural language. Such kind of research allows one not only to group words by their frequencies and obtain frequency dictionary but to examine compatibility of lexical units and grammatical forms, to facilitate the automatic separation of idioms, to realize disambiguation and parsing.

The main data source for our research is the text corpus. There we can explore grammatical and lexical ambiguity, find properties of the text which are constant for the author. Reducting the text to the form of phonetic transcription allows one to get a full statistical investigation of the phonetic system of language.

The statistical investigations are generally fulfilled in the text corpus of different genres. We suppose that the texts of the corpus are absolutely pure – they should not contain the mistakes and they are identical to originals. The texts must be previously marked according to the research purpose. As well for some investigations we use the digital Ukrainian Grammar Dictionary (UGD), which contains all of the existing word forms.

[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. Concept of statistical portrait

In this research we use the concept of the statistical portrait – a set of specially formed statistical descriptors of the text obtained by statistical processing according to some principles.

A set of statistical description consists of:

– Statistics of the character system including statistics of the phonetic (transcribed) text variation;

– Statistics of the lexical and grammatical systems;

– Syntax system statistics, etc.

Let us introduce some formal definitions.

Let T be the text in some natural language. We define a set of the mappings:

$\sigma_\alpha : T \rightarrow T_\alpha$, де $T_\alpha = \cup T_\alpha^i$, where

T_α^i – is a set of statistical parameters which presents a certain text property;

T_α – is statistical portrait of the text T, which is a union for sets of the statistical parameters for all the text properties under consideration.

2.1. Character system of language.

The lowest level of the text research represents the examination of its character system. The analysis of the character system was historically first in the language statistics.

Let us suppose U_M a generalized alphabet of some language M, e.g. the finite set of all characters used for graphical (written) representation of the texts written by the language M. This set is wider than the phonetic alphabet because in addition to letters representing the sounds of the language it contains a symbol of the space, punctuation symbols, hyphen, some special symbols (like digits), etc. So the U_M structure is the following:

$UM = BM \cup PM \cup SM \cup \pi$,

where B_M – is the alphabet of language M;

P_M – is a set of punctuation symbols: $P_M = \{p_1$='.', p_2 =',', p_3 =';', p_4 =':', p_5 ='...', p_6 ='!', p_7 ='?', p_8 ='-', p_9='"', p_{10} =')', p_{11} ='('$\}$;

S_M – is a set of special signs;

π is a space symbol.

For Ukrainian B_M is presented as:

$B_{УКР}$ = {b_1 = А, b_2 = Б, b_3 = В, b_4 = Г, b_5 = Ґ, b_6 = Д, b_7 = Е, b_8 = Є, b_9 = Ж, b_{10} = З, b_{11} =І, b_{12} =Ї, b_{13} = И, b_{14} = Й, b_{15} = К, b_{16} = Л, b_{17} = М, b_{18} = Н, b_{19} = О, b_{20} = П, b_{21} = Р, b_{22} = С, b_{23} = Т, b_{24} = У, b_{25} = Ф, b_{26} = Х, b_{27} = Ц, b_{28} = Ч, b_{29} = Ш, b_{30} = Щ, b_{31} = Ь, b_{32} = Ю, b_{33} = Я, b_{34} = ' };

Union of BM, hyphen and space symbols is identified as AM:

$A_M = B_M \cup \alpha \cup \{-\}$

We consider text as a linear sequence of symbols from the set A_M: $T = x_1, x_2, \ldots x_N$, N – is the text length, $x_i \in A_M$. In the text T we define the distribution of the text characters as probabilities to meet symbols in the text T, i.e. the number of the entries of some symbol in the text T divided to N – the full length of T. In the same way we define the distribution of the character combinations of the definite group consisting of j symbols as the number of entries of this character combination of j characters in the text T divided on N-j.

We define the transition probability of the character x to the character y ($x, y \in A_M$) as the number of entries of the combinations symbols x, y (x is the first symbol in combination, y is the second one) in the text T divided to the number of all the two-character combinations where the first symbol is x. In the same way we define n-rank transition probability from the character sequence $\{x_1, x_2, \ldots x_n\}$ to the character y as the number of entries of the symbol combinations $\{x_1, x_2, \ldots x_n, y$ (symbol y is fixed)$\}$ divided to the entry number of combinations $\{x_1, x_2, \ldots x_n, x$ (x vary over all A_M)$\}$.

The defined transition probabilities coincide with that of Markov Chain [1], i.e. our principal model of the text character system is Markov Chain, so we can predict the next symbol of the text if we know one or several of the previous symbols.

In the text corpora of language M we define the family of mappings $\sigma_{zn} : T \rightarrow T_{zn}$

$\sigma^1_{zn} : T \rightarrow T_{zn}{}^1 = \{(a); P(a), a \in A_M\}$

$\sigma^2_{zn} : T \rightarrow T_{zn}^2 = \{(a_1, a_2); P(a_1, a_2), a_1, a_2 \in A_M\}$

…

$\sigma^{10}_{zn} : T \rightarrow T_{zn}^{10} = \{(a_1, a_2, \ldots a_{10}); P(a_1, a_2, \ldots a_{10}), a_1, a_2, \ldots a_{10} \in A_M\}$

$\sigma^1_{mark} : T \rightarrow T_{mark}^1 = \{P(x_i = a_k, / x_{i-1} = a_m), a_k, a_m$ vary over all A_M, i is a serial number of the character x in the text}

$\sigma^2_{mark} : T \rightarrow T_{mark}^2 = \{P(x_i = a_k, / x_{i-1} = a_m, x_{i-2} = a_n), a_k, a_m, a_n$ vary over all A_M, i is a serial number of the character x in the text }

…

$\sigma^{10}_{mark} : T \rightarrow T_{mark}^{10} = \{P(x_i = a_k, / x_{i-1} = a_m, x_{i-2} = a_n, \ldots x_{i-10} = a_1), a_k, a_m, a_n, a_l$ vary over all A_M, i is a serial number of the character x in the text }.

$T_{zn} = \cup T_{zn}^i \cup T_{mark}^i$

T_{zn}^i is a distribution of i-symbol combinations in text T.

T_{mark}^i is a set of the i-rank transition probabilities that coincides with transition probabilities for i-connected Markov Chain.

T_{zn} – is the character level of the text statistical portrait.

The statistical investigations were fulfilled on the material of the Ukrainian text corpora [3] using the Ukrainian Grammar Dictionary (UGD)[5], which contains 3.2 million word forms.

The UGD has been built on the basis of grammatical database for the Ukrainian language which contains materials of the Ukrainian dictionaries. The left part of the UGD is a word register. Every word belongs to a certain Word-Inflection Class (WIC) and the word-inflection paradigm is automatically built according to the WIC [4]. UGD is used as a tool for analysis. Using UGD we can: a) build a chain of all the word forms for any word; b) find initial word form accurate within ambiguity if we have some word form. For some types of analysis such data of the UGD have been used as a full list of all the word forms presented there.

The data obtained are too large to show it in the article so we do not demonstrate it here. However, some conclusions on the results of the analysis can be presented.

Rank of combination	Number of unique combinations found in the UGD	Theoretically possible number of unique combinations	% presence of the theoretically possible
1 sign	35	35	100
2 signs	1037	1225	84,65
3 signs	13781	42875	32,14
4 signs	93774	1500625	6,25
5 signs	354544	52521875	0,68
6 signs	850049	1838265625	0,05
7 signs	1524535	64339296875	0,0024
8 signs	2222461	2251875390625	0,000099
9 signs	2740430	78815638671875	0,0000035
10 signs	2921764	2758547353515625	0,0000001

Table 1. Character combinations in the Ukrainian Grammar Dictionary (UGD)

Then the text generator was built on the basis of T_{zn}. The model of this generator reflects statistical laws of character distribution in Ukrainian:

G: $T_{zn} \rightarrow T^{gen}$, $G \in \{G_0, G_1, G_2, \ldots, G_n\}$,

G – is text generator, T^{gen} – is generated text.

The simple model of the text generator G_0 is represented by the discrete steady-flow process with independent values and equally probable states. It means that the first and every next symbol of the generated text are chosen in a probabilistic way independently and with equal probability. The next step presents the text generator G_1. It chooses the Ukrainian symbols according to the T_{zn}^1. The next text generators are based on the model of the Markov Chain. The generator G_2 proceeds according to the transition probabilities T_{mark}^1. The generated text looks like the following: "ПЕНИВІДЕНОВЖЕВСТ ВАПІЙТЬОМИТРЦЯ ЗАСІТОМ'Ю Д ПЕМИЙ ДЖКОМЕЙ А ПОДИРДКИ ГОБРОГА". Analogically, the generator G_3 uses T_{mark}^2 and the generator G_4 uses T_{mark}^3. The text "ВЖДЕНЕХИТИТТЯМ ВУ МАТНО ВИВ ЯКУ МОВАВРИЖА

ВОПІШКАЛА ЛЮБКАЗ СТУТІЙСЬ ДОРЧУ АРИЗНА" has been generated by $G3$, and the text "ДУМАЛИ ДЯКУСЬ НУТИ НЕ ПОДИНІ СЬОГОМУ КИНУ ХАЙ НЕЇ ТУТ СТАНА ПРИТИ У ВИЛИПА" has been generated by G_4. As one can see the last one already looks like the real Ukrainian text. So, the generators based on high levels of the Markov chains produce texts more and more similar to the real ones.

The results of the text generators' work (the length of the generated text is 4000 symbols)

Model of generator's work	General number of "words"	Average length of the "word"	Average length of the real words	Number of the real words	The real words, %
Discrete steady-flow process with independent values and equally probable states	253	14,83	0	0	0
Discrete steady-flow process with independent values and its states distributed according to the text symbols distributions	503	6,95	2,1	10	1,99
Homogeneous singly connected Markov chain	553	6,23	2,58	42	7,88
Homogeneous doubly connected Markov chain	522	6,66	3,04	95	18,2
Homogeneous triply connected Markov chain	511	6,82	3,42	225	44,03

2.2. Phonetic system

To obtain statistical information for the phonetic system of language, the original text T must be transcribed $T \rightarrow T_{tr} = at_1, at_2, \ldots at_N$, where $at_i \in AT_M$, AT_M – are symbols of the phonetic transcription for language M. For Ukrainian: $AT_{УКР}$ = {а, б, бт, б:, бт:, в, вт, в:, вт:, у, г, гт, г:, гт:, ґ, ґт, ґ:, ґт:, д, дп, д:, дп:, g, gn, g:, gn:, р, рп, р:, рп:, е, с, k, ж, жт, ж:, жт:, з, зп, з:, зп:, и, K, s, u, й, й:, к, кт, к:, кт:, л, лп, л:, лп:, м, мт, м:, мт:, н, нп, н:, нп:, о, d, п, пт, п:, пт:, р, рт, р:, рт:, с, сп, с:, сп:, т, тп, т:, тп:, у, ф, фт, ф:, фт:, х, хт, х:, хт:, ц, цп, ц:, цп:, ч, чт, ч:, чт:, ш, шт, ш:, шт:}. The analysis has been conducted on a sample that contains all the word forms fixed in the UGD. A conclusion is made that the degree of the correspondence for the Ukrainian orthography and phonetics is more than 76%. For more details one can address to the paper [2].

2.3. Lexical system

Let us consider the statistical analysis of higher level by analyzing the lexical and grammatical systems of language.

We define objects to deal with like we have done it for the character system. First we deal with the lexical system. Consider the text T as a sequence of word forms: $T=w_1\ w_2, \ldots, w_{i-1}, w_i, w_{i+1}, \ldots, w_N, w_i \in D_M$, N – is length of the texts (counted in words), D_M – is extended dictionary of language M, it contains all the word forms possible for this language. We suppose that the text does not contain words from other languages, special words like formulas and mistakes. Let us define the family of mappings $\sigma_{lex} : T \rightarrow T_{lex}$

$\sigma^1_{lex} : T \rightarrow T_{lex}{}^1 = \{(w); P(w), w \in D_M\}$

$\sigma^2_{lex} : T \rightarrow T_{lex}{}^2 = \{(w_1, w_2); P(w_1, w_2), w_1, w_2 \in D_M\}$

…

$\sigma^{10}_{lex} : T \rightarrow T_{lex}{}^{10} = \{(w_1, w_2, \ldots w_{10}); P(w_1, w_2, \ldots w_{10}), w_1, w_2, \ldots w_{10} \in D_M\}$

$\sigma^1_{l\_mark} : T \rightarrow T_{l\_mark}{}^1 = \{P(w_i=a_k,/w_{i-1}= a_m), a_k, a_m$ vary over all D_M, i – is serial number of word w in the text$\}$

$\sigma^2_{l\_mark} : T \rightarrow T_{l\_mark}{}^2 = \{P(w_i=a_k,/w_{i-1}=a_m, w_{i-2}=a_n), a_k, a_m, a_n$ vary over all D_M, i – is serial number of word w in the text$\}$

…

$\sigma^{10}_{l\_mark} : T \rightarrow T_{l\_mark}{}^{10} = \{P(w_i=a_k,/w_{i-1}=a_m, w_{i-2}=a_n, \ldots w_{i-10}=a_l), a_k, a_m, a_n, a_l$ vary over all D_M, i – is serial number of word w in the text$\}$

$T_{lex} = \cup\ T_{lex}{}^i \cup\ T_{l\_mark}{}^i$

This is the way the lexical part of statistical portrait T_{lex} of the text has been obtained.

2.4. Grammatical system

Let us consider the text T in the representation as a sequence of grammatical values of the text words: $T=g_1\ g_2,\ \ldots,\ g_{i-1},\ g_i,\ g_{i+1},\ \ldots,\ g_N,\ g_i \in GD_M$, N – is length of the texts (counted in words), GD_M – is a subset of word grammatical characteristics from the grammar dictionary of the language M.

Suppose $g_i=(p_i, f_i)$, where parameter p_i defines part of speech, f_i – is word form.

Let us define the family of mappings $\sigma_{gram} : T \rightarrow T_{gram}$

$\sigma^1_{gram} : T \rightarrow T_{gram}{}^1=\{(g);\ P(g),\ g \in GD_M\}$

$\sigma^2_{gram} : T \rightarrow T_{gram}{}^2=\{(g_1, g_2);\ P(g_1, g_2),\ g_1, g_2 \in GD_M\}$

…

$\sigma^{10}_{gram} : T \rightarrow T_{gram}{}^{10}=\{(g_1, g_2, \ldots g_{10});\ P(g_1, g_2, \ldots g_{10}),\ g_1, g_2, \ldots g_{10} \in GD_M\}$

$\sigma^1_{g\_mark} : T \rightarrow T_{g\_mark}{}^1=\{P(g_i=a_k,/g_{i-1}= a_m),\ a_k,\ a_m$ vary over all GD_M, i – is serial number of the word with grammar value g in the text$\}$

$\sigma^2_{g\_mark} : T \rightarrow T_{g\_mark}{}^2=\{P(g_i=a_k,/g_{i-1}=a_m,\ g_{i-2}=a_n),\ a_k,\ a_m,\ a_n$ vary over all D_M, i – is serial number of word with grammar value g in the text$\}$

…

$\sigma^{10}_{g\_mark} : T \rightarrow T_{g\_mark}{}^{10}=\{P(g_i=a_k,/g_{i-1}=a_m,\ g_{i-2}=a_n,\ \ldots\ g_{i-10}=a_l),\ a_k,\ a_m,\ a_n,\ a_l$ vary over all GD_M, i – is serial number of word with grammar value g in the text$\}$

$T_{gram}=\cup\ T_{gram}{}^i \cup\ T_{g\_mark}{}^i$

The same way we can define semantic, syntactic and other components of statistical portrait. Used in our research statistical portrait can be represented as union of character, lexical and grammatical components. $T_{stat} = T_{zn} \cup T_{lex} \cup T_{gram}$.

3. Applications for solution of the real language problems

Our approach to statistical analysis is based not on direct analysis of the texts. First we obtain statistical portraits of the texts and then those statistical portraits must be analyzed. Having a statistical portrait of the text, we can perform various analyses like defining the text language (or languages), formal comparison of the texts to reveal plagiarism, comparison of the text styles (to find the text's author), defining the text subject, etc. It is not necessary to get a full statistical portrait for every research. We can choose components of the statistical portrait according to a purpose of the text examination.

Now we give examples of applying statistical methods for solving some problems.

1. Analysis of the students' works texts to reveal plagiarism. This text examination is one of the easiest because we believe that student doesn't change initial text. If the initial text is changed by a student we suppose that it is a different text because text editing is the same difficult as writing a new one. It means that student not only has read it carefully and understood, but, probably, has added some his own ideas. To conduct analysis of students' works we must have a database of students' works. We compare the initial text with texts of all the works in the database that have the same subject. The statistical portrait for this examination is $T_{STAT}= T_{lex}{}^1 \cup T_{lex}{}^2 \cup T_{lex}{}^3$. The analysis here is a step-by-step comparison of sets $T_{lex}{}^i$ for the examined text and the texts from the students' works database. We can make a conclusion on whether the text was written by a student independently or was plagiarized on the estimations of coincidence of $T_{lex}{}^i$ for the examined text of student's work and the database texts. An example of such analysis is given on the figures below. *Figure 1* shows results of students' works analysis. The left part of the window is an examined text; the right part of the window is a text from the database where coincidences with the examined text were found. Such coincidences are marked blue.

2. Comparison of political parties programs. This research was realized for pre-election programs of political parties of Ukraine (elections 2002). Statistical portrait used here is $T_{STAT}= T_{lex}{}^1 \cup T_{lex}{}^2 \cup T_{lex}{}^3$. It is the same for previous research. We compare statistical portraits by pairs for every pairs of parties. Estimation of similarity measure based on similarity sets $T_{lex}{}^i$ is the result of political parties analysis. It is illustrated on the *Figure 2*.

3. Comparison of dictionaries texts. For this analysis we use the same statistical portraits as in previous.

4. Definition of the text language. Statistical portrait for analysis is: $T_{STAT} = T_{zn}^{1} \cup T_{zn}^{2} \cup T_{zn}^{3}$. To make this analysis you need a database of statistical portraits for the texts of the languages we want to recognize. Conclusion about the text language can be made by the largest coincidence of text portrait of the analyzed text and the one from the database.

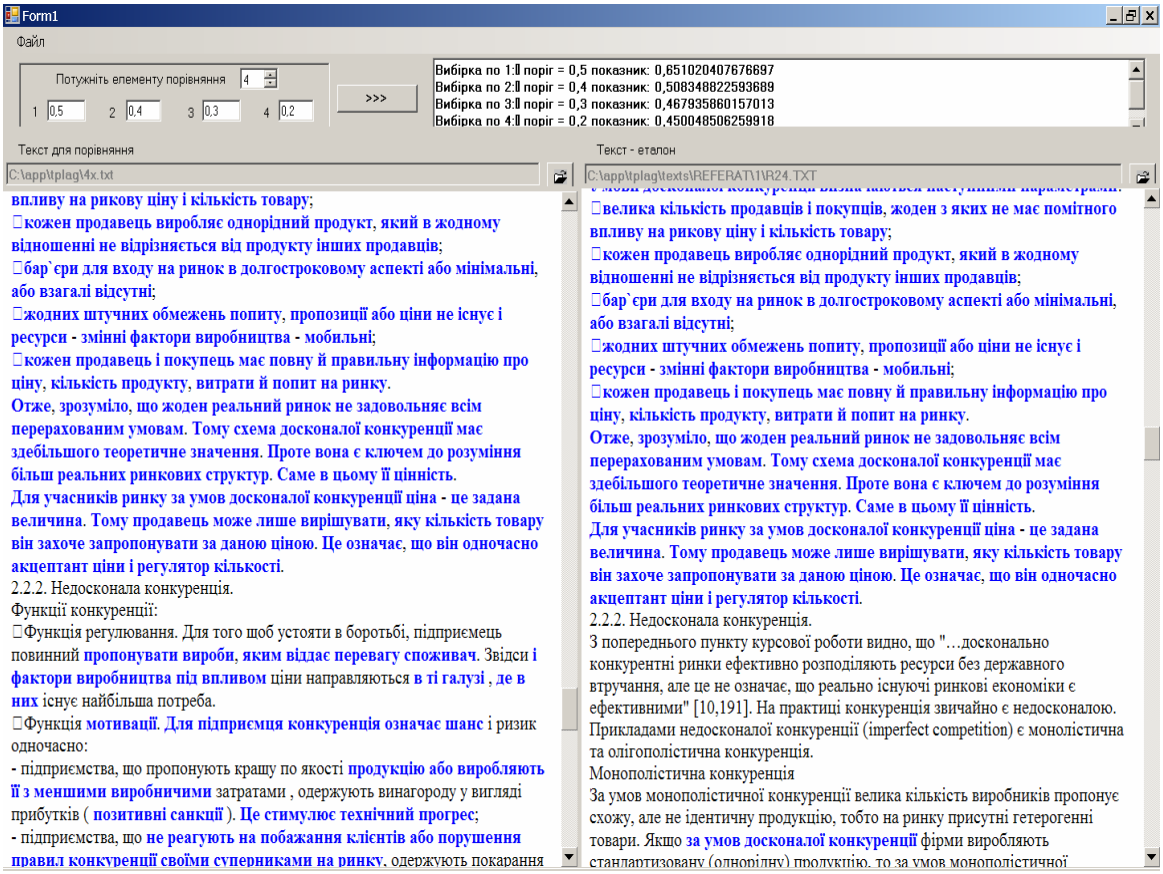


Figure 1. Result of students' works analysis.

	Аграрна	Народно-демократична	Конституція	Партія регіонів	Партія промисловців і підприємців	Рух	Соціалістична	Трудова Україна	СДПУ(О)	УРП Собор	БЮТ	Наша Україна	Комуністична
Аграрна	100	12	7	9	11	13	11	10	17	6	1	4	2
Народно-демократична	6	100	5	9	10	11	7	8	15	6	1	4	0
Конституція	2	3	100	3	3	5	4	3	7	2	0	2	0
Партія регіонів	4	8	5	100	10	9	7	7	13	5	1	4	1
Партія промисловців і підприємців	4	7	4	7	100	8	7	6	13	4	1	3	1
Рух	3	6	4	5	6	100	5	5	10	3	0	3	0
Соціалістична	3	4	4	5	6	6	100	5	11	2	0	2	1
Трудова Україна	4	6	4	6	7	8	6	100	12	4	1	4	0
СДПУ(О)	3	5	3	5	6	6	6	5	100	3	0	3	0
УРП Собор	5	10	6	9	10	12	7	8	14	100	1	4	0
БЮТ	4	6	5	6	7	8	7	7	12	6	100	5	0
Наша Україна	3	7	5	7	8	9	7	8	13	4	1	100	0
Комуністична	7	6	6	7	9	7	12	5	14	3	1	3	100

Figure 2.Results of comparison for programs of political parties.

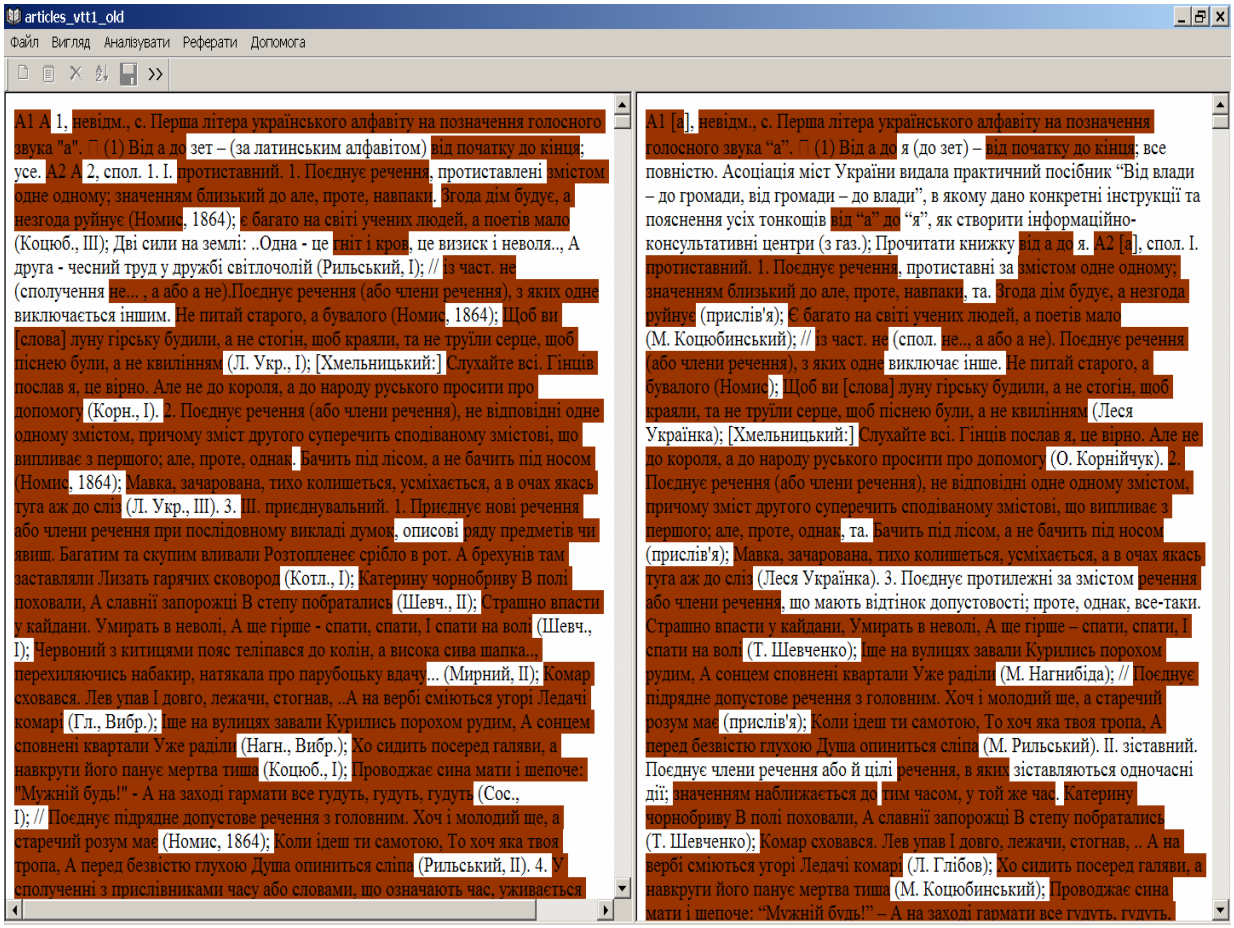


Figure 3. Comparison of the texts from the old 11-volume and the new 20-volume dictionaries of the Ukrainian language.

5. Analysis of the texts on closeness by meaning. Statistical portrait $T_{STAT} = T_{gram}{}^{i} \cup T_{g\_mark}{}^{i} \cup T_{lex}{}^{i} \cup T_{l\_mark}{}^{i}$, $i = 1, 2, 3$.

6. Definition of text's author or proof in arguable cases. $T_{STAT} = T_{gram}{}^{i} \cup T_{g\_mark}{}^{i} \cup T_{lex}{}^{1}$, $i = 1, 2, 3$.

5. Conclusion

The statistical methods can be successful for solution of a large range of the linguistic problems. Disadvantage of using statistical methods in linguistics is making approach to a problem. First you need making previous analysis of some phenomena on a big amount of text or on other linguistic data. The texts must be specially prepared before this analysis. Application of the statistical methods in linguistics can help to perform disambiguation and machine translation by application of special language models in questionable situations.

References

1. *Вентцель Е.С., Овчаров Л.А*. Теория случайных процессов и ее инженерные приложения. – М.: Наука, 1991. – С. 98-127.

2. *Кригін М.Ю*. Дослідження відповідності між фонетичними системами та графічним представленням лексики східнослов'янських мов // Бионика интеллекта / In print.

3. *Рабулець О.Г., Сидорчук Н.М*. Лінгвістичний корпус як середовище обробки та збереження інформаційних ресурсів // Прикладна лінгвістика та лінгвістичні технології. – К.: Довіра, 2008. – с. 316-328.

4. *Шевченко І.В.* Алгоритмічна словозмінна класифікація української лексики. // Мовознавство. – 1996. – № 4–5. – С. 40–44.

5. *Шевченко И.В., Рабулец А.Г., Широков В.А*. Электронный грамматический словарь украинского языка // Труды международной конференции «Megaling'2005. Прикладная лингвистика в поиске новых путей». 27 июня – 2 июля 2005 года. Меганом, Крым, Украина. – С. 124–129.

IV. Common Slavic Etymological Problems

Current Trends in the Reconstruction of Common Slavonic lexis

Tetiana Chernysh
Kyiv National Taras Shevchenko University
Kyiv, Ukraine
signum@irpen.kiev.ua

Abstract
The author focuses on features characteristic of the current reconstruction of the Common Slavonic vocabulary, such as interest in lexis and semantics as well as the use of combined comparative-typological approach and word family method.

Keywords: *the Slavonic languages, comparative linguistics, etymology, lexis, semantics, typology, word family, root, word, rerconstruction.*

Present-day Slavonic comparative linguistics is distinguished by its growing interest in lexis. Along with more traditional phonetic and morphological items, word seen as the unity of sound and sense now tends to be recognized as its legitimate object of study. As far as in 1986 G.A.Tsychun of Byelorussia described this state of affairs as "expansion of Slavonic word's rights" [Цыхун 1986: 211]. Last years have proved both the perspicacity of the remark and the fruitfulness of the tendency. Thus, it can safely be said that not only there is an urgent need (emphasized by W.Boryś in 1994 [Boryś 1994: 19]) for a historical lexicology of the Slavonic languages as a separate branch of linguistics; but also that such diachronous, or even (accordning to V.V.Nimchuk's definition) polychronous, linguistic discipline already exists, with maximally deep historical study of lexical systems as a whole being one of its tasks.

Correspondingly, there is a growing interest among etymologists in the meanings of Common Slavonic words they reconstruct. However, it already was M.Vasmer that regretfully wrote in the Afterword to his "Russian Etymological Dictionary": "If I were to start my work anew, I would give more attention to … semantics" [Фасмер 1: 14]. The reconstruction of Common Slavonic words based on their immediate and secondary reflections makes possible the reconstruction of their meaning as well. O.N.Trubachev pointed out that "reconstruction such as practiced in comparative linguistics has always been that of form. Yet for reconstruction of meaningful items to be real, it must recreate their meaning as well" [Трубачев 2004: 108]. The "root" approach in traditional etymology limited the possibility of such reconstruction since its goal was to identify words with the same root morpheme, and that put the problem of the underlying root's meaningful evolution outside its scope.

This approach reflected the general skeptic attitude of comparative linguists towards the feasibility of establishing meaning of words they reconstructed. For instance, A.Meillet argued that semantic reconstruction comparable with respect to its exactness to phonetic one was impossible [Мейе 1938: 385]. But, on the contrary, E.Benveniste, Meillet's pupil, already assessed this issue both realistically and with certain degree of optimism, distinguishing, as criteria for semantic reconstruction, an etymologist's personal choice as well as the reconstruction's verisimilitude from the viewpoint of some general considerations (or, as he put it, common sense) and parallels which the etymologist would be able to draw [Бенвенист 1973: 331-350]. Besides, he should take into account, as fully as possible, contexts in which cognate words occur and also meaningful relations among contextual variants (similar opinion was voiced in Russia by G.A.Klimov [Климов 1985: 16-23]).

It should be noted that the meaningful reconstruction of Common Slavonic word is not aimed at recreating its diffuse primary semantics which should encompass all possible results of meaning changes taking place in individual Slavonic languages; rather, its aim now is to reconstruct the hierarchically ordered system of meanings of a Common Slavonic word (including metaphors of that period). Naturally, this aim is based on understanding the semantics of ancient words as such hierarchical system. Methodologically

important in this respect is D.N.Shmelev's contention as to the diffuseness principle: "the word's semantic unity consists in certain relations between its separate meanings rather than in its having some "general meaning" allegedly subsuming individual ones" [Шмелев 1973: 76]. Commenting these words, O.N.Trubachev wrote: "Such understanding of diffuseness can be applied typologically to the reconstruction of ancient lexical and semantic entities, which, as is known, tend to be reconstructed in modern science on the basis of our knowledge about processes in living languages rather than some speculative ideas about ancient primeval simplicity" [Трубачев 1976: 166].

Thus, etymologists assume that words they reconstruct had already had in ancient times sets of meanings, and that these words had been later adopted by historical languages as their reflections with a developed semantic structure. In interpreting these, scholars take into consideration (wherever it is possible) their lexical combinability and contexts they occur in.

Establishing the semantic as well as derivational structure of reconstructed words, and, consequently, relationships of formal and meaningful derivation between them (cf. Trubachev's definition of etymology as science of historical word formation par excellence), – all this results in shifting the focus from the reconstruction of separate words and their meaning to that of the Common Slavonic vocabulary viewed as a system.

Studies in this field have been enhanced by drawing on principles, ideas and research procedures of some other branches of modern science of language, first of all cognitive and ethnolinguistics [Wojtyła-Świerzowska 1998; Mazurkiewicz 1988; Jakubowicz-MS; Jakubowicz 2000; Варбот 1998]; this author, too, will use this knowledge in her research [Черниш 2003; Черниш 2004]). The recreation of the inner form of reconstructed words elucidates hierarchical meaningful relations underlying word formative links, which in its turn makes it possible for a researcher to recreate those identifications and differentiations of objects and phenomena that ancient speakers had made in segmenting their language's conceptual space and assigning interpretation to its fragments. Viewing derivational relationships retrospectively, i.e. from the derived unit to the underlying one, enables us to discover those components of the underlying unit's semantic structure that do not belong to its significative nucleus (such as connotative-evaluative, implicative and associative semantic features) and so to reconstruct this structure more fully. This kind of integral approach to the modelling of historical lines of sense development of reconstructed units makes possible ethnolinguistic reconstruction, i.e. the recreation of a certain fragment of *Weltbild* as represented by Common Slavic. Owing to this, cognate languages lexis, traditionally recognized as a source for the historical study of culture of the ancestors of these languages speakers, acquires a new significance, in particular, with respect to reconstruction of forms of their material and spiritual life (cf. in this respect the reconstruction of the terms of ancient Slavonic agriculture in [Черниш 1991; Куркина 1998]).

Another prominent feature of contemporary comparative Slavonic lexicology and comparative Slavonic linguistics in general is the role played in them by the typological approach. V.K.Zhuravlev especially emphasizes that in present-day reconstruction of Common Slavonic, the purely genetic approach is often superseded by a unified genetic-typological one [Журавлев 1987: 493]. Also, it is because of the use of typological criteria that lexical-semantic reconstruction, as well as lexical and semantic issues as such came to be in the foreground of comparative Slavonic linguistics. And it is this change in the theory and methodology of comparative studies that has also influenced the shift in reconstruction from separate words to the whole lexical-semantic system.

An important typological aspect of comparative studies is typological verification used as a means of evaluating reconstruction. Typological data can play a crucial role in selecting the most probable hypothesis among several [Якобсон 1963, 102-104; Гамкрелидзе 1977, 195-200]). Another essential result of using typological approach in etymology consists in establishing historical models of meaning development, these models then operating as tools for assessing and verifying etymological versions.

Yet another factor important for current studies in comparative Slavonic lexicology and etymology is the introduction of the genetic word family approach in these studies. Arguably, the priority of introducing it, as well as elaborating its theory and practice belongs to the Ukrainian linguistics and, in particular, its outstanding representative O.S.Mel'nychuk, investigating these issues in the Indo-European context [Мельничук 1966; Мельничук 1968; Мельничук 1978; Черниш 2001]. Heuristic assets of the use of the word family method in the field of Slavonic studies were proved by, and further developed in, the work of other linguists in Ukraine and elsewhere [Варбот 1984; Варбот 1993; Варбот 1998; Коломієць 1992;

Козлова 1997; Шульгач 1998; Budziszewska 1983; Waniakowa 1996; Влајић-Поповић 1998; Шальтяните 1990; Калашников 1993; Куркина 1993; Іліаді 2001].

As is known, genetic (or etymological) word family is a system of cognate words, i.e. words sharing historically the same root (etymon), with this etymon being the system's core. The etymon belongs to a parent language underlying a languages family of a various degree of genetic community, i.e. Common Slavonic or Indo-European, and thus comprising languages that are more or less closely related. Correspondingly, the word family will include all known words in languages of a certain degree of genetic relationship, as well as their hypothetical cognates in the analogical word family belonging to, and reconstructed in, the parent language. Studying an etymological word family, one must establish and further analyze the set of cognate words it consists of and relations existing among them; also, it involves reconstructing the composition and structure of such a family in the parent language. The scope of such a study will be fairly wide since the word family structure includes various kinds of relationships, namely phonetic, morphological, lexical-grammatical, semantic, and word-formative, all of them closely intertwined and interacting. Studying the origin and history of a word within the framework of the word family as a systemic unity of a specific kind gives us a valuable insight into the history of the language seen as a developing natural (i.e. not artificially conceived) system. R.Kozlova arguably points out that "the problem of formation of Slavonic word can be successfully tackled … with the help of lexical material making up a word family, in which all items are interdependent and mutually conditioned, so that such a family is a microsystem in whose environment word is formed as a linguistic entity" [Козлова 1984, 18]. Because of that, etymological research carried out within word family framework makes possible a fuller and more reliable reconstruction. As O.S.Mel'nychuk emphasized, "…seen as a criterion of reliability of etymological conclusions, the volume of lexical material involved makes studies aimed at establishing composition and structure of large word families preferable to those dealing with sets of genetically unrelated words sharing some semantic feature and representing a common category" [Мельничук 1966: 265]. Besides, employing this approach yields, as its result, reconstruction of word-formative relations existing between Common Slavonic words, thus leading to integral recreation of Common Slavonic word families and corresponding fragments of Common Slavonic lexical-derivational system. Mel'nychuk also pointed out that study of genetically related word families opens a perspective for a genetic study of vocabulary of Common Slavic or another reconstructed language as a systemic whole [Мельничук 1978: 3]. Finally, such studies can provide a unified framework for dealing with issues relating both to the origin and further evolution of words, i.e. those two aspects of their history which are often treated separately, in etymological research, on one hand, and historical research, on the other.

The two approaches to the comparative study of Common Slavonic lexis, the comparative-typological and word family one, demonstrate a whole range of points of contact and fruitful cooperation; in particular, as this author's own research in this field [Черныш 1985; Черниш 1998; Черниш 2003] has hopefully shown, historical-etymological study of lexical material within word family framework implies, as one of its aspect, its historical-typological analysis. This aspect becomes even more important when the study focuses not on a single family but on several word families – one would say, a field of such families – whose etymons are either synonymous or close in meaning. It is to the latter kind of comparative study that many of my works are dedicated, whose object is families of Slavonic words with Common Slavonic etymons sharing the semantic feature 'thermal process'. The heuristic value inherent in this variety of word family approach was specially marked by O.N.Trubachev in his closing speech at the XII International congress of Slavonic studies at Krakow (1998). Afterwards, similar thoughts and themes were highlighted in S.M.Tolstaya's presentation "Semantic reconstruction and the problem of synonymy in the Common Slavonic lexis" at the next XIII Congress in Ljubljana [Толстая 2003: 550]. .

It is exactly this integral approach to understanding word family semantic structure that necessitates simultaneous study of several historical-etymological word families linked by the primary meaning of their etymons. However, in this case quest for a structured meaningful unity is carried out on a yet higher scale; observations and conclusions made here within a unified comparative-typological framework are yet more general and at the same time specific, which makes them all the more conclusive and reliable.

So, it can safely be said that the study of historical-etymological word families has given an ample evidence of its heuristic efficiency in obtaining knowledge relevant for reconstruction of Common Slavonic

word and its meaning. The exhaustive involvement of cognate words as realized within this approach makes possible an objective assessment of the range of Indo-European or Common Slavonic root morpheme self-reproduction in time and space, especially regarding its diachronic (or evolutional) polysemy, i.e. all its semantic variants ever existing. Being a result of the meaningful development of the underlying etymon, the word family semantic structure can be regarded as reproducing the latter's semantic structure. And this gives a special significance to the historical-typological study of word families with semantically related etymons. A combined use of comparative and typological methods makes the knowledge obtained in this way especially important, as it can be used both for making historical-typological generalizations and for typological verification of etymological hypotheses.

BIBLIOGRAPHY

Бенвенист 1973 – *Бенвенист Э.* Семантические проблемы реконструкции // Бенвенист Э. Общая лингвистика. – М.: Прогресс, 1973. – С. 331-349.

Варбот 1984 – *Варбот Ж.Ж.* О возможностях реконструкции этимологического гнезда на семантических основаниях // Международный симпозиум по проблемам этимологии, исторической лексикологии и лексикографии : Тезисы докладов. – М.: Наука, 1984. – С. 5-6.

Варбот 1993 – *Варбот Ж.Ж.* История славянского этимологического гнезда в праславянском словаре // Славянское языкознание : XI Международный съезд славистов. Доклады российской делегации.– М.: Наука, 1993. – С. 23-35.

Варбот 1998 – *Варбот Ж.Ж.* Славянские представления о скорости в свете этимологии (к реконструкции славянской картины мира) // Славянское языкознание : XII Международный съезд славистов. Доклады российской делегации.– М.: Наука, 1998. – С. 115-129.

Влајић-Поповић 1998 – *Влајић-Поповић Ј.* Семантика као критериум у формирању етимолошког гнезда // Prasłowiańszczyzna i jej rozpad. – Warszawa: Energeia, 1998. – S. 255-269.

Гамкрелидзе 1977 – *Гамкрелидзе Т.В.* Лингвистическая типология и индоевропейская реконструкция // Известия АН СССР. – СЛЯ. – 1977. – № 3. – С. 195-200.

Журавлев 1987 – *Журавлев В.К.* Наука о праславянском языке: эволюция идей, понятий и методов // Х.Бирнбаум. Праславянский язык: достижения и проблемы в его реконструкции. – М.: Прогресс, 1987. – С. 453-493.

Іліаді 2001 – *Іліаді О.І.* Етимологічне гніздо з коренем **ver-* у праслов'янській мові. – Кіровоград: ДЛАУ, 2001. – 162 с.

Калашников 1993 – *Калашников А.А.* Структурно-семантический анализ славянских словообразовательно-этимологических гнезд (гнезда глаголов с исходным значением "вязать, плести". Автореф. дис. … канд. филол. наук. – М., 1993. – 24 с.

Климов 1985 – *Климов Г.А.* К семантической реконструкции (по материалам кавказской этимологии) // Теория и практика этимологических исследований. – М.: Наука, 1985. – С. 16-23.

Козлова 1997 – *Козлова Р.М.* Структура праславянского слова : Праславянское слово в генетическом гнезде. – Гомель: ГГУ, 1997. – 412 с.

Коломієць 1992 – *Коломієць В.Т.* Етимологічне гніздо *(до)точити* "приєднати" в слов'янських мовах // Мовознавство. – 1992. – № 1. – С. 41-45.

Куркина 1993 – *Куркина Л.В.* О некоторых фрагментах этимологического гнезда слав. **pelti, peljǫ* // Принципы составления этимологических и исторических словарей языков разных семей : Тезисы докладов. – М., 1993. – С. 22-25.

Куркина 1998 – *Куркина Л.В.* Термины подсечно-огневого земледелия в составе праславянского словаря // Prasłowiańszczyzna i jej rozpad. – Warszawa: Energeia, 1998. – S. 207-221.

Мейе 1938 – *Мейе А.* Введение в сравнительное изучение индоевропейских языков. – М.–Л.: Государственное социально-экономическое издательство, 1938. – 510 с.

Мельничук 1966 – *Мельничук А. С.* Об одном из перспективных видов этимологического исследования // Проблемы славянских этимологических исследований в связи с общей проблематикой современной этимологии. Программа : Тез. докл. – М., 1966. – С. 11-12.

Мельничук 1968 – *Мельничук А. С.* Корень *kes- и его разновидности в лексике славянских и других индоевропейских языков // Этимология 1966. – М.: Наука, 1968. – С. 194-240.

Мельничук 1978 – *Мельничук А. С.* Этимологическое гнездо с корнем *u̯ei- в славянских и других индоевропейских языках. – К.: Наукова думка, 1978. – 16 с.

Толстая 2003 – *Толстая С.М.* Семантическая реконструкция и проблема синонимии в праславянской лексике // Славянское языкознание. XIII Международный съезд славистов. Любляна, 2003. Доклады российской делегации. – М.: РАН, 2003. – С. 549-563.

Трубачев 1976 – *Трубачев О.Н.* Этимологические исследования и лексическая семантика // Принципы и методы семантических исследований. – М.: Наука, 1976. – С. 147-180.

Трубачев 2004 – *Трубачев О.Н.* Реконструкция слов и их значений // Трубачев О.Н. Труды по этимологии: слово, история, культура. В двух томах. – М.: Языки славянской культуры, 2004. – Т. 1. – С. 107-122.

Фасмер – *Фасмер М.* Этимологический словарь русского языка. – М.: Прогресс, 1964-1973. – Т. 1-4.

Цыхун 1986 – *Цыхун Г.А.* К реконструкции праславянской метафоры // Этимология 1984. – М.: Наука, 1986. – С. 211-216.

Черниш 1991 – *Черниш Т.О.* Реконструкція внутрішньої форми лексичних реліктів праслов'янської доби в галузі підсічно-вогняного господарства // Мовознавство. – 1991. – № 6. – С. 25-30.

Черниш 2003 – *Черниш Т.О.* Слов'янська лексика в історико-етимологічному висвітленні : Гніздовий підхід. – К.: б.в., 2003. – 480 с.

Черниш 1998 – *Черниш Т.О.* Компаративно-зіставне дослідження слов'янської лексики в контексті етимологічних гнізд із близькозначними коренями // Мовознавство. (XII Міжнародний з'їзд славістів. Доповіді української делегації). – 1998 – № 2-3. – С. 168-179.

Черниш 2001 – *Черниш Т.О.* Про теоретико-методологічні засади застосування гніздового методу в етимологічних дослідженнях О.С.Мельничука // Мовознавство. – 2001. – № 6. – С. 50-57.

Черниш 2004 – *Черниш Т.О.* Внутрішня форма мовних одиниць і проблема мовного образу світу // О.О.Потебня й актуальні питання мови та культури. Зб. наукових праць. – К.: Видавничий Дім Дмитра Бураго, 2004. – С. 83-88.

Черныш 1985 – *Черныш Т. А.* Этимологические гнезда корней с исходным значением горения в славянских языках (**gŏr-/žĕr-*, **žĕg-/žĭg-*, **(pel-)/pŏl-*) : Дис. ... канд. филол. наук. – Киев, 1985. – 234 с.

Шальтяните 1990 – *Шальтяните А.П.* Семантика группы словообразовательно-этимологических гнезд в русском языке (на материале гнезд глаголов со значением "драть, дергать"). – Автореф. дис. … канд. филол. наук. – М., 1990. – 24 с.

Шмелев 1973 – *Шмелев Д.Н.* Проблемы семантического анализа лексики (на материале русского языка). – М.: Наука, 1973. – 218 с.

Якобсон 1963 – *Якобсон Р.* Типологические исследования и их вклад в сравнительно-историческое языкознание // Новое в лингвистике. – 1963. – Вып. 3. – С. 95-106.

Boryś 1994 – *Boryś W.* O możliwościach odtwarzania historii słownictwa prasłowiańskiego // Uwarunkowania i przyczyny zmian językowych. – Warszawa: SOW, 1994. – S.19-25.

Budziszewska 1983 – *Budziszewska W.* Rdzeń **reu̯-/*rou̯-* w językach słowiańskich // Studia lingvistica memoriae Zdislav Stieber dedicata - Wrocław etc.: Ossolineum, 1983. – S. 165-171.

Jakubowicz 2000 – *Jakubowicz M.* Oblicza miłości. Porównanie językowych obrazów miłości tkwiących w etymologii i frazeologii // Acta Universitatis Wratislaviensis. – 2229. – Język a kultura. – T. XIV. – Wrocław, 2000. – S. 233-244.

Jakubowicz MS – *Jakubowicz M.* O kognitywizmie w etymologii i etymologii w kognitywizmie (рукопис).

Mazurkiewicz 1988 – *Mazurkiewicz M.* Eymologia a konotacja semantyczna // Konotacja. Praca zbiorowa pod red. J.Bartmińskiego. – Lublin: UMSC, 1988. – S. 99-112.

Waniakowa 1996 – *Waniakowa W.* Derywaty z prasłowiańskim rdzeniem *tuch-* i *tǫch-* w języku polskim i innych językach słowiańskich na tle indoeuropejskim // Studia z Filologii Polskiej i Słowiańskiej. – T. 33. – 1996. – S. 225-241.

Wojtyła-Świerzowska 1998 – *Wojtyła-Świerzowska M.* Kognitywizm w etymologii // Rocznik Slawistyczny. – T. LI. – 1998. – S. 17-30.

Problems of Creation of Etymological Dictionary of Suffixes of Ukrainian Language

Vasyl Luchick
Institute of Linguistics, National Academy of Sciences, Ukraine

Abstract
''Etymological dictionary of suffixes of Ukrainian language'' will become the first edition of such type in a Slavic linguistics. In regard of its preparation several problems existed: 1) lack of experience in this sphere; 2) not equal lexicograpgical study of suffixes in Slavic languages; 3)difficulties in reconstruction of the primary meaning; 4) determining the invariants of morphemes;5) defining sources of suffixes.

Key words: *suffix, dictionary, etymological, meaning, origin, Ukrainian language.*

Etymology as a study connected with reconstruction of the primary (veritable) meaning of a word was originated in ancient Greek linguistics according to well known discussion about the character of names. But as a scientific method and separate linguistic branch it has been formed only after the appearing of comparative-historical linguistics which achieved great results during less than two centuries of its development.The main acquisition of it was the creation of etymological national dictionaries, related or prehistoric (for example, of primitive Slavonic language) the first of which was in Europe ''Етимологічний словник романських мов'' (1853) Ф.Діца and in Slavic studies – "Етимологічний словник слов'янських мов" (1886) Ф.Міклошича which was published in German language. ''Етимологічний словник української мови'' in two volumes Я.Рудницький published in Winnipeg during 1962-1982 and then came similar dictionary under the editorship of О.С.Мельничук in seven volumes (as for today 5 volumes are published).

Traditional for such dictionaries reconstructuion of the primary bases and meaning lately was accompanied with the attention to clarification of old structure of the word including historical changes, motivative connections and functions of its structural elements: ablaut, allothesis, infixes, determinatives, afixes etc. Such units were explained only opportunely according to etymologizing of bases which come to reconstruction of roots with primary material semantics. Beside the roots, the origin of preffixes was specially defined at the beginning of a word and demanded explanation of their nature. They also often preserve etymological connections with the primary roots, that gives reasons to observe them in one row with full meaning morphemes. The example can be general Slavic prefix of Indo-European origin *у-(в-),*that is used in a dimensional meaning (compare ukrainian *усувати,* рос. *убывать, убежать,* д.-р. *оуходити* and others) and together with proper preposition is studied in a separate dictionary article as a reflex *і.-є. *аu'' віддалятися, зникати''. [Фасмер,с.142].*However much numerous and important in the case of forming old and modern bases suffixes have not been the object of systematical etymological explanation in a Slavic lexicography. Clarification of their origin and primary meaning is very important for understanding the nature of the word, its internal and external structure.

First attempt to constitute the origin of suffixes was made in Slavic lexicography by a well-known polish comparativist Ф.Славський who gave substantial ''Нарис праслов'янського словотвору'' in a preface to the first volume of '' Праслов'янського словника"(1974).[Slawski,s.43-141].To tell the truth, the author confines with analyzing suffixal morphemes only of verbs and nouns and word buiding construction of adjectives and compound words with etymological commentaries of formants planned to accomplish in one of the next volumes.

More full etymological analysis of suffixes than in preface of Ф.Славський for the next 35 years in a Slavic linguistics was not made by anyone. At the same time this essay of word building of primitive Slavonic language does not cover etymology of all suffixal morphemes of a basic language and leaves

without attention all the units which were formed in separate Slavic languages.This concerns Ukrainian language in which scores suffixes of post primitive Slavonic language origin function.

There is another situation in synchronic morphemics and word building. Under the influence of structuralism which was predominant abroad in the middle of XX century there was a rapid growth of studying modern morphemics and morphological world building. In Ukraine there are several monograph researches devoted to these problems І.І.Ковалика, В.О.Горпинича, А.Д.Зверєва, Т.М.Возного, Н.Ф.Клименко, В.В.Грещука and others. As a result in a native lexicography several dictionaries in which suffixes were separated or interpreted with other morphemes in the structure of the word were published.

Thus the development of Ukrainian and in total slavic lexicography from morphemical to wordbuiding and explanatory dictionaries of affixal morphemes approached linguists for the necessity to create ''Етимологічний словник суфіксів української мови'' which together with traditional etymological dictionary will fill the system of comparative- historical analysis of basic morphemes of the Ukrainian language: roots, prefixes, suffixes.This idea was originated in the O.O.Potebnija Institute of linguistics of NAS of Ukraine in the department of general Slavic problems and East Slavic languages which began to work on this lexicographical edition.

Creation of '' Етимологічного словника суфіксів української мови'' is connected with overcoming of several problems. Firstly, there was no such dictionary in Slavic and world lexicography and to create something new was always complicated because of objective necessity to overcome unknown problems. Experience of competent lexicographers affirms that the best dictionaries in their first editions cannot be blameless. Т.Ф.Єфремова on this feature indicates using expression of the author who published the world known dictionary in three volumes ''Материалы для словаря древнерусского языка по письменным памятникам'' І.І.Срезневського who said that good and full dictionary cannot be composed from the first time. None of the books blanks and oversights conscious or unconscious are not so possible and fixed as in the dictionary and the most satisfactory dictionary during some period of time loses its dignity and needs to be fixed. [Єфремова, с.4].

Secondly, composing of the Etymological Dictionary of Suffixes of the Ukrainian Language as composing traditional etymological dictionaries provides attraction of the akin material from other Slavic and even Indo-European languages.As О.С.Мельничук mentioned that each etymological dictionary of the separate Slavic language marks definite link in a single indissoluble chain of general Slavic etymology which is in the structure of etymology of Indo-European languages and wider comparative-historical linguistics [Мельничук, с.9]. Taking into consideration that in different Slavic languages there is irregular usage of suffixes, the full value usage of definite material becomes complicated. That can have efect on the quality of separate dictionary articles.

Thirdly, suffixes are connective morphemes which are joined to root or widened with affixes bases in a derivational or grammatical function and that's why they are much more abstracted than roots with material meaning. As classificators of the words, suffixes in contrast to roots with individual lexical meaning present generalized types of marching the derivational or grammatical semantics of which is very abstract that makes difficult to determine their primary semantics. In contrast to preffixes which often remain etymological connections with prepositions and primary roots, suffixes almost lose such connections with morphemes of full meaning. To form the word building semantics is complicated.

Fourthly, there is a problem in establishment structure and differentiation of suffixes.This is connected with the fact that in a theoretical word buiding the question about distinguishing the word and wordform is still open, about correlation polysemantic and omonyms morphemes, about segmentation words into affixes about міжморфемні interlayers in reference to them several terms are used: *формативи, інтерфікси, асемантеми, структеми, субморфи.* If we take into consideration that morpheme is an elementary minimal language unit which is formally indivisible in the boundary of one word but it is divisible in semantics.That means it is bilateral unit which has the meaning and material expression and can modify as in the first also in the second case.[УМЕ, с.371-372] then dismembering functioning integral suffixes for important and unimportant components is unreasonably.The same is in the etymological dictionary unreasonably to oppose to separate morphemes phonetical variants as *–ак/-як, -уч(ий)/-юч(ий), -б-/-бл-,* that occured in the earlier mentioned dictionary of affixal morphemes Н.Ф.Клименко, Є.А.Карпіловської and others. [Клименко, с.65-68]. In particular the last morph is a phonetical variant of suffix *-б-* and it was formed as a result of interaction between labial with archaic suffix *-j-:незганьблений, вирізьблений, вирізблювати.* There are no so many reasons to distinguish as separate suffixal units vowel or consonant

complex that arise as a result of changing phonetical position of one or another affix. For example, given in ''Словнику афіксальних морфем української мови''as separate suffix alomorph *–ель-* [Клименко,с.65] emerges as a phonetical variant of the noun formant *-л(о)* in case of interaction with the next formant : compare *сідельце<сідло<псл.*sedьlo.* For this dictionary the contrasting is reasonable for its type and mission.''This is part-valency dictionary of affixal morphemes of modern ukrainian literary language which is composed with the help of computer" [Клименко, с.5] and for computer translation. For etymological interpretation of suffixes this mechanical stamping of the word structure not taking into account the integrity of morpheme as invariant similar in meaning morphs is unnecessary.

Fifthly, it is not always possible to define the interposition or the tendency of expansion the suffixal model from one language into another. In a native linguistics there is a common thought that ukrainian language borrowed from Russian suffix *–чанин,*which was originated on the base of *–анин/ -янин* as a result of decomposing of the generating base ending with *–к-,-ч-,-ц-.* But this conclusion is neither proved by historical facts nor by special researches of the specialists. This suffix is well known in the names of citizens from Old Russian and Old Ukrainian leaflets and belongs to the most productive in the katajkonimichnij system of Ukrainian language: compare *Трубежскъ-трубчане(1223), Львовъ-львовчане(1407), Глухів-глухівчани, Суми-сумчани, Стеців-стецівчани* [див.Горпинич,с.86-87]. As to the Russian language it is less productive and as unnatural for it caused serious opposition of well-known linguists С.Сергеєв-Ценський, Ф.Гладков та ін.[див.Потиха, с.191-192]. Thats why it is unfoundedly to say that suffix *-чан(и), -чанин* was borrowed from Russian language into Ukrainian. В.О.Горпинич as the most competent analyst of katajkonimichnich systems of East Slavic languages said that in Ukrainian language as in Russian and Byelorussian suffix *–чан-и (-чанин)* came in the store of general system of old Russian geographical names word building. [Горпинич,с.87].

Contrary to existing difficulties part of which is hard even to foresee the work, which has already begun, deserves approval. This edition will be useful for linguists, lecturers, students, teachers and for all who is interested in the etymology, history, morphemics and word building of the Ukrainian language. In particular the dictionary will be a good base for preparation of the academic edition ''Historical word building of the Ukrainian language" which will end the set of works in different language levels.

ЛІТЕРАТУРА

1. *Горпинич В.О*. Назви жителів в українській мові.–К.: ВШ, 1979. –158 с.
2. *Єфремова Т.Ф*.Толковый словарь словообразовательных единиц русского языка.– М.:Рус. яз., 1996. – 638 с.
3. *Клеменко Н.В. Карпіловська Є.А., Карпіловський В.С., Недозим Т.І*. Словник афіксальних морфем української мови. – К.,1988. – 434 с.
4. *Клименко Н.Ф. Карпіловська Є.А., Кислюк Л.П*. Шкільний словотвірний словник сучасної української мови. – К.: Наук.думка, 2005.
5. *Лучик В.В*. Автохтонні гідроніми Середнього Дніпро-Бузького межиріччя. –Кіровоград,1996. – 235 с.
6. *Мельничук О.С*. Вступ // Етимологічний словник української мови : В 7т./ Відп. Ред. О.С. Мельничук.– К.: Наук.думка, 1982.–Т.1–С.7-13.
7. *Онишкевич М.Й*. Словник бойківських говірок: У 2 ч.– К.: Наук.думка, 1984.–Ч.2.–515 с.
8. *Полюга Л.М*. Морфемний словник.– К.: Рад. школа, 1983.– 464 с.Р
9. *Потиха З.А*. Современное русское словообразование.– М.: Просвещение,1970.–384 с.
10. *Сікорська З.С*. Українсько-російський словотворчий словник. –К.: Рад. школа, 1985.–188 с.
11. Українська мова: енциклопедія.–Вид. 2-ге, випр.і доп.–К.: Укр. енц.ім. М.П.Бажана, 2004.–821 с.
12. *Фасмер М*. Этимологический словарь русского языка: В 4 т. / Пер. с нем. и доп. О.Н.Трубачева.–2-е изд., стер.–М.: Прогресс, 1987.–Т.4.–864 с.
13. *Цыганенко Г.П*. Словарь служебных морфем русского языка. –К.: Рад. школа, 1982.– 240 с.
14. *Яценко І.Т*. Морфемний аналіз : словник-довідник: У 2 т. / За ред.Н.Ф.Клименко .–К.: ВШ, 1980.–Т.1.–356 с.; 1981.–Т.2.–352 с.
15. An Indo-European comparative dictionary / By Stuart E. Mann .–Hamburg, 1984-1987.–1683 p.
16. *Slawski F*. Zarys slowotworstwa praslowianskiego // Slownik praslowianski / Pod red. F. Slawskiego. – Wroclaw-Warszawa-Krakow-Gdansk: PAN, 1974.–T.1.–487 s.

To the Problem of Irregular Phonetic Phenomena in Language (Delabialization *l'u- › *li-)

Viktor Shulhach
Institute of Ukrainian Language, National Academy of Sciences, Ukraine

Abstract
This article is devoted to one of the cases of irregular phonetic changes in language – delabialization ***l'u- › li-*** *(on the example of the Slavic anthroponymic vocabulary with the root* ****L'ub-****). Subject to this phenomenon the author restores a significant piece of the Pre-Slavic lexical fund. We emphasize that 36 pre-lexemes from the reconstructed list are absent in the "Etymological Dictionary of Slavic languages" by O.N. Trubachova.*

Keywords: *anthroponym, Pre-Slavic archetype, etymology*

Delabialization, or change of the initial **l'u- › *li-*, is referred to the number of so-called regular phonetic phenomena. According to historical and modern grammars, it is not known to all Slavic languages. If this regular phenomenon is at the appellate and at the proprietary levels in Czech, than, for example, in Bulgarian it is fixed sporadically in the eastern and western dialects[1], and it is reflected in some written sources[2]. Such a situation is also in the Polish, Lower and Upper Lusatian languages, in several dialects of Russian and Ukrainian[3]. However, as it is evidenced by onomastics, especially anthroponymy, the geography of this phenomenon is much broader. In this case, naturally, the correction for the possible cases of the interlanguage influence should be made (for example, for Czech and Slovak), and it is also important to take into account the fact that the class of anthroponyms is mobile in the space, that is capable of "migrations".

The following is a Slavic anthroponymy with anlaut *Li- ‹ L'u-* (a lexical-word system of derivatives with the root **L'ub-*). The material is given under the relevant Pre-Slavic archetypes.

***Bogol'ubъ:** укр. *Боголіб* (Літ. ЖС 2007, № 6, 92).

***Dol'uba:** укр. *Доліба* (КПУ Хм. 3, 368), пол. *Doliba* (SN II, 462) ~ славян. **Dol'ubъ* [‹ *Taliub,* 791 г. – антропоним альпийских славян (Kronsteiner 210 – с реконструкцией)], ст.-блр. (производное) *Dolubow*, XVI ст. – название поместья (ПКГЭ II, 399).

***Dol'ubanъ:** пол. *Doliban* (SN II, 462).

***Dol'ubičь:** хорв. *Dolibić* (Leksik 135).

***Dol'ubosъ:** пол. *Dolibós* (SN II, 462).

***Dol'ubьcь:** укр. *Долібець* (РР 102).

***L'uba, *L'ubo, *L'ubъ(jь):** ст.-укр. *Занко Либа*, 1552 г. (Тупиков 226), укр. *Ліб, Ліба* (Горпинич, Тимченко 163), блр. *Ліба* (Бірыла 251), *Либо* (г. Минск), *Либов* (г. Могилев), русск. *Либ* (САМ 132), *Либо* (г. Калуга), болг. *Либа* (Илчев 302), *Либо* (Заимов 139), хорв. *Lib, Libo, Liby* (Leksik 365, 366), пол. *Lib, Liba* (SN V, 587), чеш. *Líba, Libo, Libý* (ЧП), слвц. *Liba, Libo* (TZ Bratislava 207), луж. *Liba* (Jordan 141), *Libo* (Wenzel II/1, 249). Ср. также славян. (производные) *Λίμποβον* – топоним на территории Греции (Гильфердинг 287 – как *Любово*), *Либы*, XVI ст. – топоним в Жемайтии (Спрогис 167).

***L'ubahъ:** македон. (производное) *Либаово* (варианты *Либахово, Либаво*) – топоним (Симовски 2, 61), пол. *Libach* (SN V, 587).

***L'ub'aha, *L'ub'ahь:** болг. *Либяха* (Заимов 139), (производное) *Либяхово* – старое название с. Илинден в Гоцеделчевско.

***L'ubajь:** русск. *Либаев* (Pamięć 2, 320).

***L'ubakъ:** русск. *Либаков* (ЖПТ), чеш. *Libák* (ЧП), (производное) *Libákovice* – топоним (Profous II, 576), луж. *Libak* (Wenzel II/1, 247).

***L'ub'akъ:** хорв. *Libjak* (Leksik 365), чеш. *Libiak* (ЧП), слвц. *Libiak* (TZ Bratislava 207).

***L'ubanъ, *L'ubanь:** укр. (производное) *Либановка* – топоним в бывшей Волинській губ. (Vasmer RGN V, 133), русск. *Либанов* (СКТ 451), блр. *Либанов* (г. Могилев), болг. *Либàн* (Заимов 139), (производные) *Либан* (варианты *Любян, Любаново*) – топоним (*Иванов Долна* Струма 151), *Либàнов дол*, *Либàнов рът* – микротопонимы, которые производят из антропонима **Любàн* (Заимов Панагюрско 127), макед. (производное) *Либаново* – топоним (Симовски 1, 120), хорв. *Liban* (Leksik 365) ~ *Ljuban* (Leksik 379), слвн. (производное) *Libanja* – топоним (Im. m. 260), пол. *Liban, Libań* (SN V, 587), чеш. *Libánek* (ЧП), (производные) *Libáň* (2) – топонимы (Profous II, 576).

***L'ubatъ(jь):** болг. *Либат* – личное имя (Заимов 139).

***L'ubava:** пол. *Libawa* (SN V, 587).

***L'ubexъ:** болг. (производное) *Либèхово* – ороним (Иванов Долна Струма 151), чеш. *Liběchov* (ЧП).

***L'ubenъ:** болг. *Либен* (Илчев 302), (производные) *Либенов трап* (Ковачев Троянско 166), *Либеновото* (Ковачев Габровско 124), *Либенов рът* (Заимов Пирдопско 204) – микротопонимы, чеш. *Líbenek* (ЧП), (производные) *Libeň* (3) – топонимы (Profous II, 581).

***L'ubešь:** чеш. *Libeš, Liběšov* (ЧП), (производное) *Liběšov* – топоним (Profous II, 585).

***L'ubezьnъ(jь):** чеш. *Líbezný* (ЧП).

***L'ubějь:** ст.-русск. *Лембейко Якушовъ, Гридка Лембіевъ*, 1500 г. (Тупиков 225, 620), болг. *Либей*, 1445 г. – имя князя (Морошкин 111).

***L'ubędь:** пол. (производное) *Libiąż* – ойконим (NMP VI, 99). См. также: (Казлова II, 72-73).

***L'ubęga:** укр. *Лібега* (Рівне 416).

***L'ubica:** болг. *Либица* (Заимов 139; Илчев 302), слвц. *Libica* (TZ Bratislava 207).

***L'ubičь:** укр. *Лібич* (Богдан 163), русск. *Либич* (ЖПТ), (производное) *Либичова Гатка* – микротопоним (Ковалев 222), *Лимбичи-Сельга* – ойконим в бывшей Олонецкой губ. (Vasmer RGN V, 139), схв. *Либић* (Речник XI, 409) ~ хорв. *Ljubić* (Leksik 379), ст.-пол. *Libicze*, 1473 р. ~ *Lubicz*, 1443 р. (SSNO III, 299), пол. *Libicz* (SN V, 588), чеш. *Libič* (ЧП), слвц. *Libič* (TZ Bratislava 207).

***L'ubixъ:** пол. (производное) *Libichowa* – топоним, известный с 1262 г. (NMP VI, 99-100).

***L'ubikъ:** русск. *Либиков* (САЛ 500), блр. *Лібік* (Бірыла 251), слвц. *Libíková* (TZ Bratislava 207), пол. *Libik* (SN V, 588), луж. *Libik* (Wenzel II/1, 247).

***L'ubimirъ:** лит. *Libimĩrskas* (LPŽ II, 74) ‹ славян.

***L'ubinъ:** пол. *Libin* (SN V, 588). Ср. также славян. (производное) *Λιμπίνοβον* – топоним на территории Греции (Гильфердинг 288 – как *Любиново*).

***L'ubisъ:** пол. *Libis* (SN V, 587).

***L'ubiša, *L'ubišь:** макед. (производное) *Либишево* – топоним (Симовски 2, 28), хорв. *Libiš* (Leksik 365), пол. *Libisz* (SN V, 588), (производное) *Libiszów* – ойконим (NMP VI, 100), чеш. *Libiš* (ЧП), луж. *Libiš, Libiša* (Wenzel II/1, 249).

***L'ubivьcь:** ст.-русск. *Либивец*, 1547-1548 гг. (Греков 86), русск. (производное) *Либивцево* – ойконим в бывшей Псковской губ. (Списки Псков. 548).

***L'ublь ‹ *L'ubjь:** хорв. *Libl* (Leksik 365), слвн. (производные) *Libelj, Libeliče* – топонимы (Im. m. 260), пол. *Libel* (SN V, 587), чеш. *Líbl* (ЧП).

***L'ubočajь:** пол. *Liboczaj* (SN V, 589).

***L'ubogodъ:** чеш. *Libohod* (Moldanová 106, статья Líba).

***L'ubogostь:** чеш. *Libhost* (Jungmann II, 315), (производное) *Libohošt'* – топоним (Profous II, 596). См. еще: (ЭССЯ 15, 177).

***L'uboxъ, *L'uboxa:** укр. (производное) *Лібухова* – ойконим в Львовской обл. [1589 г. – *Lubochow,* 1785 г. – *Libuchowa* (Makarski 144)], макед. (производное) *Либохово* – топоним (Симовски 1, 341), пол. *Libocha* (SN V, 589). Ср. еще славян. (производное) *Λιμπόχοβον* – топоним на территории Греции, который восстанавливают как *Любохово* (Гильфердинг 289).

***L'uboměričь:** чеш. (производное) *Libоměřice* – топоним (Profous II, 598).

***L'ubomirъ:** чеш. *Libomír* (Kott I, 913).

***L'ubomyslъ:** чеш. *Libomysl* (Kott I, 913), (производное) *Libomyšl* – топоним (Profous II, 598).

***L'ubonъ, *L'ubonь:** пол. *Libon, Liboń* (SN V, 589), чеш. (производное) *Libonice* – топоним (Profous II, 598).

***L'uborada, *L'uboradъ:** пол. *Liberada* (SN V, 587), *Liberadz* – ойконим, производный от личного имени **Luborad* (NMP VI, 99).

***L'ubosějь:** русск. (производное) *Либосеевка* (вариант *Любосеевка*) – гидроним в бас. Оки (Смолицкая 199).

***L'ubosikъ:** пол. *Libosik* (SN V, 589).

***L'uboslavъ:** чеш., слвц. *Liboslav* – мужское имя (Kott I, 913; Majtán, Považaj 160).

***L'ubostь:** пол. *Libostka* (SN V, 589). Сюда же славян. (производное) *'Αλῆ μπόστοβα* – топоним (с греческой протезой) на территории Греции, производный, по мнению, И. Заимова, от исчезнувшего имени Любост (Заимов Нови 113).

***L'ubosvarъ:** чеш. *Libosvár* (ЧП), слвц. *Libošvár* (TZ Trnava 124).

***L'ubosъ(jь):** славян. *Либосый* (Pamięć 2, 523).

***L'ubošь:** пол. *Libosz* (SN V, 589), чеш. *Liboš* – мужское имя (Knappová 123), *Libośka* (ЧП).

***L'ubota, *L'ubotъ:** чеш. *Libotovski* (ЧП), (производные) *Libotov, Libotyně* – топонимы (Profous II, 600). Сюда же славян. *Али ботуш* – топоним на территории Греции, который связывают с личным именем *Ботух* + *-jь* (Заимов Нови 113).

***L'ubovidъ:** славян. *Λιμποβίζια* – топоним на территории Греции, из **Любовижа* (Добрев 54).

***L'ubožerъ:** чеш. *Libožery* – топоним (Profous II, 598).

***L'ubožęda:** укр. (производное) *Либожада* – гидроним в бас. Уборти п. Припяти п. Днепра (СГУ 313).

***L'ubőtъ:** слвц. *Libutka* (TZ Nitra 119).

***L'ubra, L'ubrъ:** д.-русск. *Либиаръ*, 944 г. – один из русских послов, подписавший договор с греками (ПСРЛ I, 20), укр. *Лібра* (ВЧ 177), *Лібер* (РІ ІФ 2, 46), ст.-русск. *Ыван Либор*, 1550 г. (Мацулевич 18), русск. *Либрович* (САЛ 501), *Либеров* (КЕКН 2, 172), хорв. *Libar* (Leksik 365), пол. *Libr, Libra* (SN V, 589), чеш. (производное) *Libeř* – топоним (Profous II, 582).

***L'ubričь:** хорв. *Librić* (Leksik 366), слвн. *Librič* (ZSSP 334), славян. *Libericz*, 1388 р., вариант *Luberitz*, 1424 г. (Körner 94, 96).

***L'ubrikъ:** укр. *Лібрик* (Богдан 163).

***L'ubrъko, *L'ubrъkъ:** ст.-укр. Wстапъ *Либєрко*, 1649 г. (Реєстр 267), пол. *Liberek* (SN V, 587).

***L'ubrykъ:** пол. *Libryk* (SN V, 589).

***L'ubuxa, *L'ubuxъ:** пол. *Libucha* (SN V, 589), (производное) *Libusza* – ойконим, мотивированный личным именем **Lubuch* (NMP VI, 101).

***L'ubujь:** ст.-русск. (производное) *Лимбуева Лахта,* 1500 р. – топоним в Карелии (Сергий 93) ‹ **Любуева Лахта*.

***L'ubunъ:** укр. *Либун* (СКТ 451), чеш. (производное) *Libouň* – топоним (Profous II, 601).

***L'ubusъ:** укр. *Лібус* (Рівне 416), пол. *Libus* (SN V, 590), чеш. *Libus* (ЧП).

***L'ubusь:** русск. *Либусь* (САМ 133), пол. *Libuś* (SN V, 590).

***L'ubuša, *L'ubušь:** болг. *Либуша* (Илчев 302), хорв. *Libušek* (Leksik 366), пол. *Libusz* (SN V, 590), чеш., слвц. *Libuša* (Kott I, 914; Majtán, Považaj 160), луж. *Libuš, Libuša* (Wenzel II/1, 249).

***L'ubъka, *L'ubъkъ:** русск. *Либкин* (Лет. ЖС 2004, № 1-13, 103), хорв. *Libek* (Leksik 365), пол. *Libek* (SN V, 587), чеш. *Libek* – мужское имя (Knappová 123), (производное) *Libkov* – топоним (Profous II, 591), луж. *Lib(e)k* (Wenzel II/1, 247).

***L'ubъša, *L'ubъšь:** пол. *Libsz* (SN V, 589), чеш. *Libša* – мужское имя (Knappová 123), луж. *Libš* (Wenzel II/1, 249).

***L'ubьcь:** укр. *Лібець* (ВЧ 177), русск. *Либец* (КС Зап. 352), хорв. *Libac* (Leksik 365) ~ *Ljubac* (Leksik 379), чеш. (производное) *Libeč* – топоним (Profous II, 578).

***Nal'ubъ:** блр. (производное) *Nalibówka* – топоним в бывшей Минской губ. (Vasmer RGN VI, 83).

***Nel'uba:** чеш. *Neliba* (ЧП).

***Obl'ubъ:** укр. (производное) *Олибів* – ойконим в Ровенской обл.

***Perl'ubъ:** чеш. (производное) *Prelibsko* – топоним (Profous III, 465).

*__Pol’ubinъ:__ ст.-русск. *Русинъ Рудаковъ* с. Полибина, 1616 г. (Тупиков 700), Федор Полибин, 1668 г. (ОАРП 206).

*__Sъl’ubőta:__ русск. (производное) *Слибуты* – ойконим в бывшей Псковской губ. (Vasmer RGN VIII, 338).

*__Sъl’ubьnikъ:__ луж. *Slibnik* (Wenzel II/2, 99).

*__Tul’uba:__ пол. *Tuliba* (SN IX, 617), ст.-укр. (производное) *Тулибле*, 1552 г. – ойконим (АЮЗР VII/1, 99).

*__Vyl’ubičь:__ хорв. *Vilibić* (Leksik 716).

*__Vьsel’ubъ:__ чеш. (производные) *Všeliby* (2), *Všelibice* – топонимы (Profous IV, 643, 644).

*__Zal’ubuxъ:__ укр. (производное) *Залибухів* – название бывшего хутора в Любешовском р-не Волынской обл.

*__Zal’ubъ:__ укр. (производное) *Залибовка* – топоним в бывшейй Волынской губ. (Vasmer RGN III, 402), пол. *Zalibowski* (SN X, 416).

Commented above, the actual material from the Slavic anthroponymy can be used as an important additional tool for reconstructing the lexical Pre-Slavic fund (in this case, reconstructing a fragment of lexical-word microsystem with the root **L’ub-*). Thus, 36 potential Pre-Slavic archetypes from the reconstructed list are absent in the “Etymological Dictionary of Slavic languages” by O.N. Trubachova. There are **L’ubičь, *L’ubikъ, *L’ubišь, *L’ubogostь, *L’uboradъ, *L’ubota, *L’ubъša, *L’ubra, *L’ubьcь, *Nel’uba*, etc. among them. Thus, in the process of selection and the linguistic expertise of the lexical material we must take into account the fact of irregular phonetic transformations and do not neglect the facts of onomastics.

References

[1]*Младенов Ст*. История на българския език. София, 1979. С. 96.

[2]*Стефова Л.* Един ранен пример за перехода ’у › и в български // Старобългаристика. 1995. XIX. С. 73-77 (с литературой).

[3]*Gruszczyńska M.* Oboczność ’u // i w języku polskim // Zeszyty naukowe Uniwersytetu Jagiellońskiego. Prace językoznawcze. Kraków, 1968. Z. 21. S. 121-136.

sources

АЮЗР – Архив Юго-Западной России, изд. Временною комиссиею для разбора древних актов [...]. Киев, 1859-1914. Ч. I-VIII.

Бірыла – *Бірыла М.В*. Беларуская антрапанімія. 2: Прозвішчы, утвораныя ад апелятыўнай лексікі. Мінск, 1969.

Богдан – *Богдан Ф*. Словник українських прізвищ у Канаді. Вінніпег; Ванкувер, 1974.

ВЧ – Всі Чернівці. 1999: Телефонний довідник. Чернівці, 1999.

Гильфердинг – *Гильфердинг А*. История сербов и болгар // Гильфердинг А. Собр. сочинений. СПб., 1868. Т. I. С. 281-291.

Горпинич, Тимченко – *Горпинич В.О., Тимченко Т.В*. Прізвища правобережного Степу: Словник. Дніпропетровськ, 2005.

Греков – *Греков Б.Д*. Монастырское хозяйство XVI-XVII веков. Л., 1924.

Добрев – *Добрев И*. Славянска топонимия в Пелопонес като извор за историята на българския език // Съпоставително езикознание. София, 1987. Т. XII. № 3. С. 45-57.

ЖПТ – Жертвы политического террора в СССР. In: http://lists.mono.ru.

Заимов – *Заимов Й*. Български именник. София, 1988.

Заимов Нови – *Заимов Й*. Нови български географски имена от Гърция // Славистичен зборник. Посвещава се на IX международен конгрес по славистика в Киев, 1983. София, 1985. С. 113-120.

Заимов Панагюрско – *Заимов Й*. Местните имена в Панагюрско. София, 1977.

Заимов Пирдопско – *Заимов Й*. Местните имена в Пирдопско. София, 1959.

Заіка – *Заіка З.М*. З мікратапаніміі Століншчыны // Беларуская анамастіка. Мінск, 1992. С. 140-149.

Иванов Долна Струма – *Иванов Й.Н.* Местните имена между Долна Струма и Долна Места. София, 1982.

Илчев – *Илчев С.* Речник на личните и фамилни имена у българите. София, 1969.

Казлова – *Казлова Р.М.* Беларуская і славянская гідранімія: Праславянскі фонд. Гомель, 2002. Т. II.

КЕКН – Кто есть кто в Нижегородской области. Нижний Новгород, 2000. Вып. 2.

Ковалев – *Ковалев Г.Ф.* Микротопонимия Воронежской области: Словарь. Воронеж, 2007.

Ковачев Габровско – *Ковачев Н.П.* Местните названия в Габровско. София, 1965.

Ковачев Троянско – *Ковачев Н.П.* Топонимията на Троянско. София, 1969.

КПУ Хм. – Книга пам'яті України: Хмельницька область. Хмельницьк, 1995. Т. 3.

КС Зап. – Книга скорботи України: Запорізька область. Запоріжжя, 2000.

Лет. ЖС – Летопись журнальных статей. М., 2004.

Літ. ЖС – Літопис журнальних статей. К., 2007.

Мацулевич – *Мацулевич Л.А.* Фрески барабанов Знаменского собора // Сб. Новгородского об-ва любителей древности. Новгород, 1911. Вып. V. С. 1-45.

Морошкин – *Морошкин М.* Славянский именослов, или Собрание славянских личных имен в алфавитном порядке. СПб., 1867.

ОАРП – Описи архива Разрядного приказа XVII в. / Подг. текста и вступ. статья В.П. Петрова. С.-Петербург, 2001.

ПКГЭ – Писцовая книга Гродненской экономии с прибавлениями, изд. Виленскою комиссиею для разбора древних актов. Вильна, 1891-1892. Ч. I-II.

ПСРЛ – Полное собрание русских летописей. М., 1962. Т. 1.

Реєстр – Реєстр Війська Запорозького 1649 р.: Транслітерація тексту / Підгот. до друку: О.В. Тодійчук, В.В. Страшко, Р.І. Осташ та ін. К., 1995.

Речник – Речник српскохрватског књижевног и народног језика. Београд, 1959-2001. Књ. I-XXVI.

РІ ІФ – Реабілітовані історією: Івано-Франківська область. Івано-Франківськ, 2001. Т. 2.

Рівне – Рівне: Щорічний телефонний довідник. Рівне, 2006.

РР – Романів і вся Романівщина: Телефонно-інформаційний довідник [...]. Житомир, 2003.

САЛ – Список абонентов Ленинградской городской телефонной сети. 1965. Л., 1965.

САМ – Список абонентов Московской городской телефонной сети. М., 1939.

СГУ – Словник гідронімів України / Ред. колегія: А.П. Непокупний, О.С. Стрижак, К.К. Цілуйко. К., 1979.

Сергий – *Сергий*, архимандрит. Черты церковно-приходского и монастырского быта в писцовой книге Водской пятины 1500 года (в связи с общими условиями жизни). СПб., 1905 (Приложения).

Симовски – *Симовски Т.Х.* Населените места во Егејска Македонија. Скопје, 1998. Д. 1-2.

СКТ – Справочник квартирных телефонов г. Киева / Сост. Д.М. Циолек. К., 1976.

Смолицкая – *Смолицкая Г.П.* Гидронимия бассейна Оки (Список рек и озер) / Под ред. О.Н. Трубачева. М., 1976.

Списки Псков. – Списки населенных мест Российской империи, изд. Центральным статистическим комитетом МВД. СПб., 1885. Т. XXXIV: Псковская губерния.

Спрогис – Географический словарь древней Жомойтской земли XVI столетия, сост. по 40 актовым книгам Россиенского земского суда / Сост. И.Я. Спрогис. Вильна, 1888.

Тупиков – *Тупиков Н.М.* Словарь древнерусских личных собственных имен // Записки Отделения рус. и славян. археологии имп. Русского археол. общества. 1903. Т. 6. С. 86-914.

ЧП – Чоловічі прізвища громадян Чеської Республіки за даними Міністерства внутрішніх справ ЧР станом на 1.04. 2004 // www.mvcr.cz.

ЭССЯ – Этимологический словарь славянских языков: Праслав. лекс. фонд / Под ред. О.Н. Трубачева. М., 1988. Вып. 15.

Im. m. – Imenik mesta. Pregled svih mesta i opština, narodnih odbora srezova i pošta u Jugoslaviji. Beograd, 1956.

Jordan – *Jordan H.* Delnjołužicke swójbne mjena // Časopis Maćicy Serbskeje. 1892. Lět. XLV. S. 138-143.

Jungmann – *Jungmann J.* Slovník česko-německý. Praha, 1835-1839. D. I-V.

Knappová – *Knappová M*. Jak se bude vaše dítě jmenovat? Praha, 2001.

Körner – *Körner S*. Die patronymischen Ortsnamen im Altsorbischen. Berlin, 1972.

Kott – Česko-německý slovník zvlášté grammaticko-fraseologický / Sest. Fr.Št. Kott. V Praze, 1878-1884. D. I-IV.

Kronsteiner – *Kronsteiner O*. Die alpenslawischen Personennamen. Wien, 1981.

Leksik – Leksik prezimena Socijalističke Republike Hrvatske. Zagreb, 1976.

LPŽ – Lietuvių pavardžių žodynas. Vilnius, 1985-1989. T. I-II.

Majtán, Považaj – *Majtán M., Považaj M*. Vyberte si meno pre svoje diet'a. Bratislava, 1998.

Makarski – *Makarski W*. Nazwy miejscowości dawnej ziemi Przemyskiej. Lublin, 1999.

Moldanová – *Moldanová D.M*. Naše příjmení. Praha, 2004.

NMP – Nazwy miejscowe Polski: Historia. Pohodzenie. Zmiany / Pod red. K. Rymuta. Kraków, 2005. T. VI.

Pamięć – Pamięć. Память / Составители: Я. Пшимановский, Х. Прокопчук, Р. Мурани. Пер. с польск. под ред. К. Козакевич. Варшава, 1987. Ч. 1-2.

Profous – *Profous A., Svoboda J*. Místní jmená v Čechách: Jejich vznik, původní význam a změny. Praha, 1947-1957. D. I-IV; 1960. D. V. Napsali J. Svoboda, V. Šmilauer a další.

SN – Słownik nazwisk współcześnie w Polsce używanych / Wydał K. Rymut. Kraków, 1992-1994. T. I-X.

SSNO – Słownik staropolskich nazw osobowych / Pod red. W. Taszyckiego. Wrocław etc., 1965-1985. T. I-VII.

TZ Bratislava – Bratislava. Telefónny zoznam. 2002-2003. Bratislava, 2002.

TZ Nitra – Nitra. Telefónny zoznam. 1988-1989. Bratislava, [1988].

TZ Trnava – Trnava. Telefónny zoznam. 2004-2005. Trnava, 2004.

Vasmer RGN – Russisches geographisches Namenbuch / Begr. von M. Vasmer. Wiesbaden, 1962-1980. Bd I-X.

Wenzel – *Wenzel W*. Studien zu sorbischen Personennamen. Bautzen, 1991-1992. T. I-II/1, 2.

ZSSP – Začasni slovar slovenskih priimkov / Odg. red. F. Bezlaj. Ljubljana, 1974.

Authors

Tetiana Chernysh: professor, Kyiv National Taras Shevchenko University Kyiv, Ukraine

Ivan Derzhanski: associate professor, Department of Mathematical Linguistics, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria.

Ludmila Dimitrova: associate professor, Department of Mathematical Linguistics, 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

Violetta Koseska-Toszewa: professor, head of Department of Semantics, Institute of Slavic Stdies, Polish Academy of Sciences, Warsaw, Poland

Natalia Kotsyba: assistant professor, Institute of Slavic Studies Polish Academy of Sciences, Warsaw, Poland

Simon Krek: researcher at the Department of Knowledge Technologies at the Jožef Stefan Institute, Ljubljana, Slovenia

Maxim Krygin: researcher of Ukrainian Lingua-Information Fund National Academy of Science of Ukraine, Kyiv, Ukraine

Vasyl Luchick: professor, Institute of Linguistics, National Academy of Sciences, Ukraine

Tetyana Lyubchenko: researcher of Ukrainian Lingua-Information Fund National Academy of Science of Ukraine, Kyiv, Ukraine

Daniela Majchráková: Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences

Eugene Martynov: professor, head of Laboratory of Grid Calculations in Physics; Bogolyubov Institute for Theoretical Physics of NAS of Ukraine, Kyiv, Ukraine

Antoni Mazurkiewicz: professor, Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

Irina Ostapova: researcher of Ukrainian Lingua-Information Fund National Academy of Science of Ukraine, Kyiv, Ukraine

Alexander Rabulets: senior researcher of Ukrainian Lingua-Information Fund National Academy of Science of Ukraine, Kyiv, Ukraine

Peter Rashkov: Jacobs University Bremen, Germany

Roman Roszko: docent dr habil, Department of Semantics, Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, Poland

Bogdan Rudyj: Linguistic Museum, Taras Shevchenko National University Kyiv, Ukraine

Igor Shevchenko: senior researcher of Ukrainian Lingua-Information Fund National Academy of Science of Ukraine, Kyiv, Ukraine

Viktor Shulhach: Institute of Ukrainian Language, National Academy of Sciences, Ukraine

Volodymyr Shyrokov: professor, director of Ukrainian Lingua-Information Fund National Academy of Science of Ukraine, Kyiv, Ukraine

Konstantin Tyschenko: professor, Linguistic Museum, Taras Shevchenko National University Kyiv, Ukraine