THE EXPERIENCE OF DEVELOPING SOFTWARE FOR TYPOLOGICAL DATABASES (ON THE EXAMPLE OF DB "LANGUAGES OF THE WORLD") [1]

Vladimir Polyakov

Abstract: In the present article we will discuss the experience of creating software for the typological database "Languages of the World". The DB "Languages of the World" is one of the biggest typological computer resources. We have done a review of the software connected with the DB "Languages of the World". The following questions are discussed: compatibility of the versions, choice of the best structure of the data, development of the content in newer versions of the DB, creation of bilingual version, correct citing. The main lessons learnt from the project by the workgroup, are:

Long development and creation of different versions of the product during its life cycle (over 20 years), providing its livability against the background of changing of operational systems and paradigms of programming makes us seriously think about a technology of providing for compatibility between different versions of the product, documenting of the code, preserving the key participants of the workgroup.

The structure of the DB is a secondary moment in the relation to the content. In the end, choice of a certain structure of data presentation in a certain realization of the product is a question of comfortable programming. Besides, choice of the structure of the data is in many situation defined by the environment of data storage, dates and budget of the product.

Planning a long life cycle of a linguistic resource for scientific purposes must foresee tools of fixation and archiving the inevitable changes of the content. Lack of such tools or links to the contents without invariant binding lowers the quality and the value of the received scientific results.

The creation of the bilingual version of the product demanded thorough elaboration of the terminological part of the DB, as well as linkage of the languages to the international system of coding. Along with it, the specificity of Russian scientific linguistic school and a more detailed description of the languages of Eurasia in the DB "Languages of the World" did not allow us to withdraw these contradictions completely.

The main scientific results received for the past 5 years with the use of the DB, are enumerated. The perspectives of its future development and use are studied.

Keywords: *language typology, linguistic database*

[1] The research was supported by Russian Scientific Foundation of Humanities (grant □ 10-04-12125□)

***ACM Classification Keywords**: J.5 Arts and Humanities - Linguistics*

1. Introduction

The success of informational technologies, abundance of linguistic information on different aspects of linguistics, presence of hard-to-solve scientific problems in this field promoted first the appearance of, first as shy attempts, and then wider spread of computer linguistic resources. Today using computer technologies in the sphere of linguistics is not rare, and it is even becoming a necessary element of both the process of studying and the process of learning. Computer technologies give the researcher and the teacher an incontestable competitive advantage.

We can consider the creation of text corpora for national languages a start point of using computers in linguistics [Francis & Kucera, 1967]. It promoted using quantitative methods (statistical, distributive) and later studies with the use of the technologies of the intellectual data analysis. The necessity of solving the problems of computer linguistics conditioned the growth of interest for the sphere of corpora researches, creation of representative corpora in different aspects of the theory and the practice of text and speech. The appearance of the net Internet makes these resources even more available, and allow to conduct researches in the remote regime [Bultreebank, Penn_treebank, PDT, EXMARaLDA]. Later researches on creating computer linguistic resources were conducted also in the sphere of linguistic semantics. The most striking examples of such resources are the projects WordNet [Miller, George A., 1995], [WordNet Online], Roget’s Thesaurus [Roget, 1852], [Roget_Tes], FrameNet [Fillmore, Collin, 2010], [FrameNet], VerbNet [Kipper-Schuler, 2005], [VerbNet].

By the present time computer methods have been introduced into almost all spheres of linguistics. In the field of comparative linguistics lexical-statistical DB appeared historically first [Starostin]. Later there also appeared typological DB: Ethnologue - Languages of the World [Ethnologue], The World Atlas of Language Structures - WALS [Haspelmath et al., 2005], [WALS], UNESCO Atlas of the World's Languages in danger [Moseley, Christopher (ed.), 2010], Database “Languages of the World” of Institute of Linguistics of Russian Academy of Sciences [Polyakov, Solovyev, 2006], [DB_JM]. As opposed to researches in the sphere of corpus linguistics, which were first initiated by the problems of creating frequency dictionaries, and then by actual applied tasks (such as POS-tagging, morphological analysis, syntactic analysis, reference, etc.), in the linguistic typology the creation of DBs was initiated by the needs of fundamental scientific researches and partially by the educational sphere. A lot of discussion problems of typology, comparative linguistics, historical and areal linguistics, that had not been solved in the precomputer epoch, received a chance for solution in the conditions of large-scale use of databases and computer quantitative methods [Wichmann, Saunders, 2007], [Bayrasheva, Solovyev, 2008] [Solovyev, 2010].

In the present article we will discuss the experience of creating software for the typological DB “Languages of the World”. The DB “Languages of the World” is one of the biggest typological computer resources. There are resources that exceed the DB in the number of covered languages, but so far there are no resources that have a comparable level of detail of description of such big language communities. The article [Polyakov, Solovyev, Wichmann, Belyaev, 2009] gives a detailed comparison of the DB “Languages of the World” and WALS. The project DB was started in the eighties of the XX-th century in Institute of Linguistics RAS and at the present time is positioned as an infrastructural scientific project. The DB has several versions, including Windows and Web-version. Around the kernel of the DB there were created a lot of programs for research purposes. Interesting scientific results were received and new quantitative methods of scientific researched were worked out. The focus of the present article is on the problems that arise during the creation of software for such a long-term scientific computer project.

2. History of the project and a brief characteristic of the DB

In the 80-s in Institute of Linguistics of Russian Academy of Sciences (IL RAS) a decision about the launch of work on creation of the database Jazyki mira ("Languages of the World") was made. The encyclopedic issue of the same name [Jazyki mira, 1993…2006] is used as the source of information. The work was initiated by corresponding member of the Academy of Science Yartseva V.N. and was conducted at the department of applied linguistics under the surveillance of Novikov A.I. The following people took part in the development of the conception and the structure of the DB: Zotova A.K., Ryabtseva N.K., Rogova N., Romanova O.I. – analysis of abstracts, Vinogradov V.A., Zhurinskaya M.A., Testelets Ya.I., Yaroslavtseva E.I. – authors of the model, Skokan U.P., Novikov A.I., Nesterova N.N. – computer formalization of the model.

The first version of the DB was realized by programmer Skokan U.P. in DBMS Clipper (MS DOS). The registration certificate of Unitary Entreprise Scientific Technical Centre "Informregistr" № 7706 from November, 26, 2001 was received for the DB. A number of publications were made. In 2005 Yaroslavtseva E.I. defended a doctoral thesis on the topic "The computer database "Languages of the World" and its possible applications" [Yaroslavtseva, 2005].

In 2002 a Windows-version of the DB "Languages of the World" was created (project director – Polyakov V.N., programmer – Logunov V.). In 2005 the first variant of the Web-version was presented, and in March, 2006, the base was published on the Internet at www.dblang.ru (project director – Polyakov V.N., programmers – Goncharov E., Shcherbinin T., Khanukaev R.). A curriculum of the optional course "DB "Languages of the World" and new possibilities of typological and comparative researches" was worked out (authors: Polyakov V.N., Solovyev V.D.), it was read at the department of Theoretical and Applied Linguistics (philological department of Moscow State University) and at the department "Linguistics" of South Ural State University in 2006. In 2008 work on Reference and educational version of the DB with a more developed interface and new possibilities of search and navigation began (project director – Polyakov V.N., programmers – Belyaev O., Anisimov I.). This project is planned to summarize the experience gained for the past five years of researches and operation of the DB [Belyaev, 2008].

The work on creation and further development of the DB was repeatedly supported by grants of Russian Humanitarian Scientific Fond and Russian Fond of Fundamental Researches, partially is was financed by Moscow State Linguistic University, in 2006 and 2010 it became part of financing of Baudouin de Courtenay Russian Scientific-Educational Centre of Linguistics of Kazan State University (manager of scientific educational centre – Solovyev V.D.)

The Data Base "Languages of the World" (Jaziki Mira) has the following quantitative characteristics.

- contains more than 3800 features

- the number of languages is 313 Eurasian languages

- contains the description of the following spheres of language: phonetics, morphology, syntax.

- representation of data: binary

In Data Base "Languages of the World" the following language families and unities are represented: Austroasian, Austronesian, Altaic, Afroasian, Indoeuropean, Caucasian, Paleoasian, Sinotibetic, Uralic, Hurrito-Urartean. DB contains the description of languages-isolates: Ainu, Nivch, Burushaski, Sumeran, Elamite. The unique peculiarity of Data Base "Languages of the World" is a large collection of extinct languages description, that includes 55 essays. There is no analogues of such detailed and systematic description of extinct languages.

The main principles forming the model of language description are binarity, hierarchicity and paradigmaticity. The overall number of binary states in the DB is more than 1.2 million. In order to give an example of the complicity of processing such volume of data we will note that calculating a matrix of measures of similarity between all the languages on a modern personal computer takes more than 10 hours.

3. Review of software connected with the DB "Languages of the World"

Sources of Data for DB JM are:

- Encyclopedic issue "Jaziki Mira" (Languages of the World) – 15 volumes, printed by Institute of Linguistics of Russian Academy of Science from 1993 to 2009.
- Large Encyclopedic Dictionary. Linguistics (Edited by Yarceva V.N.) – includes interpretation of all terms of model of DB.

Main work on language description in DB format was fulfilled by Yelena Yaroslavceva, DSc.

The Kernel version includes full content of database and main functions for adding, editing and searching for data (see table 1).

About the conversion

Text abstracts were chosen as a way of converting the data during a transfer from the DOS-version to the Windows-version of the DB. During the direct data transfer on the level of DBF-files there appeared difficulties connected with the restrictions on the structure of DBF files and the coding of the Cyrillic alphabet. In the end transfer on the level of text abstracts turned out to be the best from the point of view of data safety. It required the support of two formats of text files in the Windows-version, because in the Russian Windows-version the coding Windows-1251 is used, and it is different from the coding of the Cyrillic alphabet CPP-866, which was adopted in the DOS-version.

The second problem that arose during the data transfer was connected with the necessity of replenishment the model of the abstract in the Windows-version. An abstract model is a list of features. In order to exclude possible mistakes each abstract is checked for new features during the launch. If there are new features a specialist in the DB must introduce the necessary changes to the structure of the feature space or correct the mistake in the abstract. Then the abstract is loaded again. And it happens until there are no notices on mistakes. This technology ensures the high level of the control of reliability of the source data.

Table 1.

	Program language or environment / Database Engine and data format	Programmers, year of issue	English interface and content	Main functions	Compatibility
DOS Version	Clipper / Dbase compatible, DBF	Skokan †, 1997 (*)	Yes, but not synchronized with RUS-version on content	Correction of model, add new languages, browse, export, import, save, search , comparison	With Win version via files of essay export/import
Win Version	Pascal Delphi / Borland Database Engine, DBF	Logunov, Polyakov, 2002 (*)	Yes, but not synchronized with RUS-version on content	Correction of model, add new languages, browse, navigation, export, import, save, simple and complex search, comparison, alphabetic and thematic indices	With DOS version via files of essay export/import, with Web version via direct conversion of database files
Web version	C# and .NET / MS SQL Server	Goncharov (1st var.), 2005 Khanukaev (2nd var.), 2006 (**) There is also a Linux-version (at KSU).	The content Is fulfilled (Yaroslavtceva, Makarova). Interface is fulfilled (Khanukaev).	Browse, tree navigation, comparison	Loads data from Win version via direct conversion of database files

* Task formalization was done by Novikov †

** Task formalization was done by Polyakov.

Finally, certain difficulties arose also in connection with the creation of English version of the DB. It was historically established that works on the creation of the Russian and the English versions in the IL RAS were conducted separately. In the end the Russian version was far ahead of the English one in the number of languages. Now these lags are eliminated, but a decision to merge the two versions in one program product was made. It will eliminate the possibility of accidental mistakes and will lead to the unification of the space of languages and features.

It is necessary to note that works on coding the features in the DB require an exceptional mental outlook and patience. A specialist, who makes the coding, must be very good at the very difficult subject area in order to be able to establish binary equivalents of the features in the DB on the base of text descriptions. Moreover, a text of an encyclopedic article can contain hidden information that presupposes further explication or reference to a relative language. Yaroslavtseva E.I. did this work brilliantly. The examination of the DB conducted in the Kazan Center of Linguistics and IL RAS showed a relatively low level of mistakes. In future it is intended to create in the framework of the work group special questionnaires that allow to fill the abstracts with language descriptions in an easier way. An example of such questionnaire for section 2.1.1. PHONEMIC STRUCTURE was made by E. Loginova [Loginova, 2008].

About the structure

Such question as the data structure deserves a separate discussion. Initially the author of the first (DOS) version of the DB and the structure of the main data table Skokan made a decision to create a flat rectangular table, where rows in this table are features, and columns are languages (see pic. 1). In the field where the column (language) crossed the row (feature) there was the value True, if the given language had this grammar category, and False, if it did not. It is evitable that such approach does not conform to the canons of relational DBMS that are established in the modern ITs. But at that moment – mid 80-s – such data format, taking into consideration the volume of data, available computer powers, was optimal from the point of view of efficiency, speed of search and data retrieval.

	Language 1	Language 2	…	Language m
Feature 1	True	True	…	False
Feature 2	True	False	…	False
…	…	…	…	…
Feature n	False	True	…	True

Pic. 1. Structure of the main table of data

When the data were transferred from the DOS-version to the Windows-version, it was decided to keep the stated data structure. But in the latest Web-version this structure was transformed into a relational format for supporting SQL-queries. The relational format of the DB was also adopted in the work [Omlin, 2010]. Special programs-converters were created in order to transfer the data from the table to the relational structure.

About the content

Initially the content of the DB was limited to the contents of the encyclopedia articles. At the same time some sections of articles (mainly sociolinguistical and ethnographical) were loaded to the MEMO-field and were unavailable for search. All other sections were coded in the binary system of values. During the shift to the educational-inquiry version it was decided to eliminate this shortcoming. Besides, a number of data arrays that present interest for query forming, were added to the DB.

The following data were added to the educational-inquiry version of the DB:

-text of a language description from the encyclopedia in PDF-format (only in Russian);

-glossary;

-genetic index;

-references to literature sources;

-examples demonstrating the meaning of grammar categories;

-status of the language (literature, education, writing);

-changes due to contacts (grammatical and lexical);

-number of speakers;

-frequency of feature spread in macro-families, families, branches, groups, subgroups;

-geographic coordinates [Loginova, 2009];

-the language code according to the standard ISO 639-3 (see www.ethnologue.com);

-for most features there is a link to the corresponding article in the online encyclopedia Wikipedia.

Moreover, every quantitative program product creates its own digital content, which is difficult for transferring to the main DB. These data are usually published by the researchers separately.

About the English version

During the creation of the English version there were found terminological equivalents for languages and grammar categories. Sometimes it was hard to do. For example, in the DB there are dialects of the languages whose names are neither at the web-site www.ethnologue.com nor in the encyclopedia Wikipedia. In this case we had to make a calque from the Russian language. There were similar problems when grammar features were translated.

About the versions of the DB

One of the problems that we cannot manage so far is fixing and operating the archive of changes in the content of the DB. Changes in the DB can be caused by an expertise of existing language descriptions or adding new language descriptions. Due to this some quantitative calculations made in previous versions of the content, can give a wrong link in a newer version during referring to the numbers of features. That is why in publications it is necessary to give the date of calculations and the version of the DB. Besides, it is desirable to specify the full path of a certain grammar category in the tree of features when a reference to this feature is given.

For example:

```
Nominative/accusative<=subjective-objective <=argument case meanings <=2.3.4.CASE MEANINGS (1359)
```

Here the number in brackets is the identifier of the feature in the DB, the tree root is in the right part of the line.

For the past five years of researches connected with the DB, the work group have worked out several program products.

In table 2 different kinds of software products related to DB JM are represented.

The last row of the table (Outer tools applicable to JM data) presents the products of detached developers that are used during the work with the contents of the DB.

Quantitative and other research products connected with DB JM are described in table 3.

The apocryphal structure of the main table turned out to be very convenient during the transfer of the data to the format MS Excel, and the built-in language VBA allowed to create a number of successful quantitative products (Similarity, LangFam). Thus, today the DB is used in three data formats: DBF, XLS, SQL Server.

Also some referential tools are developed (see table 4).

It is possible that in future there will appear new specialized products containing a fragment of the DB and supplemented with some new information. There already exists such practice. For example, in the project [Omlin, 2010] the content of the DB “Languages of the World” was united with the content of the project UNESCO (Atlas of the World’s Languages in Danger) [Moseley, Christopher (ed.), 2010]. There are joint works with New Bulgarian University on creating a specialized version of the DB that will be dedicated to the case system of languages.

Table 2.

Versions of Database	DOS Version	Windows Version	Web Version www.dblang.ru
Quantitative And Other Research Products	Includes comparison of two languages as function	Similarity – Software for similarity measure calculations LangFam – Software for language family portraits calculations, genetic markers revealing, deal with rare features filters, investigate typological shift etc. Special software for modeling of evolution Special software for clusterization task Special software for phylogeny with different metrics of feature space BiCoTree –software for easy tree building on DB. Some other research programs, developed for different aims during partial investigations in areal, historical and typological linguistics (Gusareva, Loginova, Fashutdinov, Omlin, Polyakov, Solovyev).	Includes comparison of two languages as function
Reference and Educational Products (under constr.)		Living Diagrams – reference software with possibility of integration source data and quantitative diagrams EduDBLANG – educational version of DB with full spectrum of reference possibilities	The Web-version of "Living diagrams" is prepared.
Outer tools applicable to JM data		R – statistical sotware tools Different phylogeny tools.	

Table 3.

Product	Program language or environment / Database Engine and data format	Programmers, year of issue	Main functions
Similarity	VBA, Excel	Polyakov, 2006	Similarity measure calculations and evaluation
LangFam	VBA, Excel	Polyakov, 2006	Software for language family portraits calculations, genetic markers revealing, deal with rare features filters, investigate typologycal shift etc.
Special software for modeling of evolution	Pascal Delphi	Yuzhikov, 2006 (*)	Modeling of process of appearance, borrowing, extinction of features. Uses different parameters of model, gives different quantitative values.
Special software for clusterization task	Pascal Delphi	Dvoenosova (1st var), 2006 Zheleznovsky (2nd var), 2008 (*)	Clusterization of languages and features by different techniques of classic cluster analysis
Special software for phylogeny wspaceith different metrics of feature	Visual C	Faskhutdinov , 2008 (*)	Use two heuristic ideas of L- and S-metrics for calculation of distance between languages.
BiCoTree –software for easy tree building on DB.	Pascal Delphi	Sarvarov, 2010 (*)	
Some other research programs, developed for different aims during partial investigations in areal, historical and typological linguistics	C, Pascal Delphi	Gusareva, Loginova, Fashutdinov, Omlin, Polyakov, Solovyev	Allow to solve different tasks: - To calculate a core of relevant features for different language families; - To calculate a motherland for different language families using grammar features; - To calculate stability index using different metrics; - Etc.

Table 4.

Product	Program language or environment / Database Engine and data format	Programmers	Main functions
Living Diagrams	C# and .NET MS SQL Server Excel	Khanukaev (*)	Reference software with possibility of integration source data and quantitative diagrams. Allows to draw quantitative pictures or tables and to do queries to source data immediately from picture. Has purpose to improve confidence of linguists to quantitative results.
EduDBLANG	C# and .NET MS SQL Server Excel	Belyaev (*)	Educational version of DB with full spectrum of reference possibilities. Includes genetic and geographic indices, annotation and examples for features, full texts of papers according to the best WALS traditions. New concept of user interface.

4. Lessons, prospects, scientific results

Let us formulate the main lessons that were learnt by the work group from this project:

Long development and creation of different versions of the product during its life cycle (over 20 years), providing its livability against the background of changing of operational systems and paradigms of programming makes us seriously think about a technology of providing for compatibility between different versions of the product, documenting of the code, preserving the key participants of the workgroup.

The structure of the DB is a secondary moment in the relation to the content. In the end, choice of a certain structure of data presentation in a certain realization of the product is a question of comfortable programming. Besides, choice of the structure of the data is in many situations defined by the environment of data storage, dates and budget of the product.

Planning a long life cycle of a linguistic resource for scientific purposes must foresee tools of fixation and archiving the inevitable changes of the contents. Lack of such tools or links to the contents without invariant binding lowers the quality and the value of the received scientific results.

The creation of the bilingual version of the product demanded thorough elaboration of the terminological part of the DB, as well as linkage of the languages to the international system of coding. Along with it, the specificity of Russian scientific linguistic school and a more detailed description of the languages of Eurasia in the DB "Languages of the World" did not allow us to withdraw these contradictions completely.

Let us enumerate the scientific results received for the five past years of using the DB in scientific researches.

- On the base of the data, a new quantitative model of language evolution was worked out and approved by V.D. Solovyev. The use of this model allowed to receive an invariant curve, which reflects the stages of evolution: diagram "Language-Feature" (LF-diagram) [Polyakov, Solovyev, 2006].

- Polyakov V.N. received data on a new diachronic phenomenon, which the author called Typological Shift [Polyakov, Solovyev, 2006], [Polyakov, Yaroslavtseva, 2008]

- Polyakov V.N. worked out and optimized algorithms for calculating measures of language similarity on the base of the similarity of their grammar structure. For the first time a good coincidence of the values of likeness and the data on genetic relationship of languages was shown [Polyakov, Solovyev, 2006] [Polyakov, 2008], and it allowed to bring forward a thesis that the grammar structure carries sufficient information on genetic relationship, which was earlier questioned.

- In the period from 2006 to 2010 a group under the guidance of Solovyev V.D. has made a valuable contribution to adaption of methods of phylogeny to calculation of genetic trees on the base of the grammar structure of languages. New metrics and algorithms of calculating genetic trees were introduced. [Solovyev, 2007], [Solovyev, Fashutdinov, 2008], [Solovyev, Fashutdinov, 2009], [Solovyev, 2009]

- Polyakov V.N. introduced and improved a number of methods for formulating and verifying genetic hypotheses and/or areal contacts, including: method of ranging languages according to the value of similarity measure, method of quantitative maps of language communities, method of quantitative filters, method of filters according to genetic markers [Polyakov, Solovyev, 2006].

- Group under the guidance of Polyakov V.N. revealed genetic markers for the Altai language family, revealed the core of features for the Ural languages (in print).

- Parameters of the age of language families of Eurasia are specified [Solovyev, Fashutdinov, 2009-2], [Solovyev, 2009-2], [Bayrasheva, Solovyev, 2008-2].

- In cooperation with colleagues from Max Planck Institute for Evolutionary Anthropology indices of stability of grammar categories are calculated [Wichmann, Holman, 2009], [Belyaev, 2009], [Solovyev, Fashutdinov, 2009-3];

- A number of new methods and results of quantitative calculations for comparative linguistics, typology and areal linguistics are worked out (in print).

In future the development of the DB will be conducted in the following directions:

- development of new quantitative products and joint DBs for solving the problems in the sphere of linguistic typology, comparative linguistics, historical and areal linguistics;

- generalization of separate quantitative products and technologies in the framework of a united technological environment;

- integration of numerical data of quantitative calculations in the united DB;

- creation of geographic information applications with the use of the contents of the DB.

5. Conclusions

The DB “Languages of the World” is a linguistic resource that has value for the sphere of researches and education. The history of elaboration and development of the DB, software connected with it, researches on its basis, allowed to gain important experience of conducting such large-scale interdisciplinary and interinstitutional projects. At the present time the DB is an infrastructural scientific resource that has a high scientific potential, involved in the international scientific society and providing a few levels of use in the sphere of science and education.

Bibliography

[Bayrasheva, Solovyev, 2008] Solovyev V., Bayrasheva V. Statistic analysis of linguistic databases: the new perspective in the typology and comparative studies. Text processing and cognitive technologies. V.17. 2008, p. 210 – 214.

[Bayrasheva, Solovyev, 2008-2] Bayrasheva V., Solovyev V. Modelling the Evolution of Language Features. Proc. of the intern. conf. “Cognitive and Functional Perspectives on Dynamic Tendencies in Languages”. Tartu: University of Tartu. 2008. p. 198-199.

[Belyaev, 2008] Oleg Belyaev. A New Interface Model of the DB “Languages of the World”. In Text Processing and Cognitive Technologies. Paper Collection. (Edited by V. Solovyev, M. Bergelson, V. Polyakov). The X-th International Conference “Cognitive Modeling in Linguistics”. Proceedings. Volume 3. Kazan: KSU, 2008, p. 118-128.

[Belyaev, 2009] Belyaev, Oleg. 2009. Stability of language features: a comparison of the WALS and JM typological databases. Paper presented at Cognitive Modeling in Linguistics–2008, September 6–12, Bechichi, Montenegro. In print. Available online at: http://obelyaev.googlepages.com/BelyaevJMStab.pdf .

[Bultreebank] HPSG-based Syntactic Treebank of Bulgarian. URL: http://www.bultreebank.org/

[DB_JM] Database “Languages of the World”. Institute of Linguistics. Russian Academy of Sciences. URL: http://www.dblang.ru/en/Default.aspx

[Ethnologue] Ethnologue, Languages of the World. An encyclopedic reference work cataloging all of the world’s 6,909 known living languages. URL: http://www.ethnologue.com/

[EXMARaLDA] EXMARaLDA: SFB 538 Corpora. Spoken Language Corpora at the Research Center on Multilingualism. URL: http://www.exmaralda.org/corpora/en_sfbkorpora.html

[Fillmore, Collin, 2010] Fillmore, Charles J. & Baker, Collin F. 2010. A Frame Approach to Semantic Analysis, in Heine, B. & Narrog, H. (eds.) Oxford Handbook of Linguistic Analysis

[FrameNet] FrameNet Project.

URL: http://framenet.icsi.berkeley.edu/index.php?option=com_frontpage&Itemid=1

[Francis & Kucera, 1967] Francis S. and Kucera H., Computational analisys of present-day American English, Providence, RI: Brown University Press, 1967.

[Haspelmath et al., 2005] Haspelmath, Martin & Matthew S. Dryer & David Gil & Bernard Comrie (eds.) The World Atlas of Language Structures. − Oxford: Oxford University Press, 2005. − 695 p.

[Jazyki mira, 1993] Jazyki mira: Ural'skie jazyki (Languages of the World: Uralic languages). 1993. Moscow.

[Jazyki mira, 1997] Jazyki mira: Tûrkskie jazyki (Languages of the World: Turkic languages). 1997. Moscow: Indrik.

[Jazyki mira, 1996] Jazyki mira. Paleoaziatskie jazyki (Languages of the World. Palaeoasiatic languages). 1996. Moscow: Indrik.

[Jazyki mira, 1997] Jazyki mira: Mongol'skie jazyki. Tunguso-Man'čžurskie jazyki, Japonskij jazyk. Korejskij jazyk. (Languages of the World: Tunguso-Manchurian languages. Japanese language. Korean language). 1997. Moscow. Indrik.

[Jazyki mira, 1997] Jazyki mira: Iranskie jazyki. I. Jugo-zapadnye iranskie jazyki (Languages of the World: Iranian languages. I. Southwest Iranian languages). 1997. Moscow: Indrik.

[Jazyki mira, 1998] Jazyki mira: Dardskie i nuristanskie jazyki (Languages of the World: Dardic and Nuristani languages). 1998. Moscow: Indrik.

[Jazyki mira, 1999] Jazyki mira: Germanskie jazyki. Kel'tskie jazyki (Languages of the World: Germanic languages. Celtic languages). 1999. Moscow: Academia.

[Jazyki mira, 1999] Jazyki mira: Iranskie jazyki. II. Severo-zapadnye iranskie jazyki (Languages of the World: Iranian languages. II. Northwest Iranian languages). 1999. Moscow: Indrik.

[Jazyki mira, 1999] Jazyki mira: Iranskie jazyki. III. Vostočnoiranskie jazyki (Languages of the World: Iranian languages. III. East Iranian languages). 1999. Moscow: Indrik.

[Jazyki mira, 2001] Jazyki mira: Kavkazskie jazyki (Languages of the World: Caucasian languages). 2001. Moscow: Academia.

[Jazyki mira, 2001] Jazyki mira: Romanskie jazyki (Languages of the World: Romance languages). 2001. Moscow: Academia.

[Jazyki mira, 2004] Jazyki mira: Indoarijskie jazyki drevnego i srednego perioda (Languages of the World: Old and Middle IndoAryan languages). 2004. Moscow: Academia.

[Jazyki mira, 2005] Moldovan, A. M., S. S. Skorvid, A. A. Kibrik et al. (eds.). 2005. Jazyki mira: Slavânskie jazyki (Languages of the World: Slavic languages). Moscow: Academia.

[Kipper-Schuler, 2005] Karin Kipper-Schuler. 2005. VerbNet: A broad-coverage, comprehensive verb lexicon. Ph.D. thesis, University of Pennsylvania.

[Loginova, 2008] Liza Loginova. The Problems of Representation of Typological Data About Language in the Format of the Database (On the Material of the DB "Languages of the World") In Text Processing and Cognitive Technologies. Paper Collection. (Edited by V. Solovyev, M. Bergelson, V. Polyakov). The X-th International Conference "Cognitive Modeling in Linguistics"(CML-2008) . Proceedings. Volume 3. Kazan: KSU, 2008, p. 188-195.

[Loginova, 2009] Elizaveta Loginova. Technique of Definition of Geographical Coordinates at the Identification of the Area of Distribution of Language (On the Material of DB «Jaziki Mira»). In Text Processing and Cognitive Technologies. Paper Collection. The XI-th International Conference "Cognitive Modeling in Linguistics" (CML-2009). Proceedings. Kazan: KSU, 2009.

[Miller, George A., 1995] George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[Moseley, Christopher (ed.), 2010] Moseley, Christopher (ed.). 2010. Atlas of the World's Languages in Danger, 3rd edn. Paris, UNESCO Publishing. Online version URL: http://www.unesco.org/culture/languages-atlas/

[Omlin, 2010] Samuel Omlin. Study of the Relation of the Linguistic Environment of a Minority Language to Its Vitality. In Text Processing and Cognitive Technologies. Paper Collection. The XII-th International Conference "Cognitive Modeling in Linguistics"(CML-2010). Proceedings. Kazan: KSU, 2010.

[PDT] The Prague Dependency Treebank 1.0. URL: http://ufal.mff.cuni.cz/pdt/

[Penn_treebank] Penn Treebank Project. URL: http://www.cis.upenn.edu/~treebank/

[Polyakov, 2008] Polyakov V. Approaches to improvement of similarity measure based on the structure of language description in the DB "Languages of the World". Text processing and cognitive technologies. v.17. 2008, p. 192 – 209.

[Polyakov, Solovyev, 2006] Polyakov, Vladimir N. and Valery D. Solovyev. 2006. Komp'juternye modeli I metody v tipologii i komparativistike (Computational Models and Methods in Typology and Comparative Linguistics). Kazan: Kazanskiy Gosudarstvennyy Universitet. 208 p.

[Polyakov, Solovyev, Wichmann, Belyaev, 2009] Polyakov V., Solovyev V., Wichmann S., Belyaev O. Using WALS and Jazyki mira. Linguistic Typology. V. 13. 2009. P. 135–165.

[Polyakov, Yaroslavtseva, 2008] Polyakov V.N., Yaroslavtseva E.I. Kvantitativnie zakonomernosti tipologicheskogo sdviga v yazikah Evrazii (The Quantitative Parameters of Typological Shift in Languages of Eurasia). Uchenie zapiski KGU, Vol.150, Book 2, 2008, p. 97-118.

[Roget, 1852] Roget, Peter Mark [1852] (1962), Dutch, Robert A., O.B.E., ed., The Original Roget's Thesaurus of English Words and Phrases (Americanized edition), New York, NY, USA: Longmans, Green & Co./Dell Publishing Co., Inc.

[Roget_Tes] ROGET's Hyperlinked Thesaurus. URL: http://www.roget.org/index.htm

[Solovyev, 2007] Solovyev V.D. Zadachi i metodi lingvisticheskoy filogenomiki (The tasks and the Methods of Linguistic Phylogeny). Conference "Znaniya, ontologii, teorii". Proceedings. Novosibirsk. Siberian Department of RAS, 2007.

[Solovyev, 2009-2] V.D. Solovyev. Viyavlenie sluchaev parallel'noy evolyucii s pomoshch'yu bazi dannih "Yaziki mira"(Revealing of Cases of Parallel Evolution by Means of a Database "Languages of the World") The 8th Conference on Languages of Far East, Southeast Asia and West Africa (LESEWA-8). Proceedings. Moscow. 2009.

[Solovyev, 2009] Solovyev V.D. Problemi i metodi lingvisticheskoy filogenii (Problems and Methods of Linguistic Phylogeny). Uchenie zapiski KGU. Vol. 151, Book 6, 2009. P. 8-21..

[Solovyev, 2010] Solovyev V.D. Tipologicheskie bazi dannih: perspektivi ispol'zovaniya (Typological databases: prospects of usage). Voprosi yazikoznaniya, 2010, №1. p. 94-110

[Solovyev, Fashutdinov, 2008] V.D. Solovyev, R.F. Fashutdinov. Vibor metriki dlya filogeneticheskih algoritmov (Choice of the Metrics for Phylogenetic Algorithms). Scientific Session of Moscow Engineering Physics Institute. Proceedings, Vol. 10. Moscow: MEPhI, 2008. p. 176.

[Solovyev, Fashutdinov, 2009] V.D. Solovyev, R.F. Fashutdinov. Preobrazovanie metrik, ispol'zuemih v metodah klasterizacii dlya postroeniya filogeneticheskih derev'ev yazikov (Transformation of the Metrics Used in Methods of Clusterization for Construction of Phylogenetic Trees of Languages). Uchenie zapiski KGU. Vol. 151, Book 3. 2009. P. 229–239

[Solovyev, Fashutdinov, 2009-2] V.D. Solovyev, R.F. Fashutdinov. Metodika kolichestvennoy ocenki skorosti evolyucii grammatiki (Technique of a Quantitative Estimation of Speed of Evolution of Grammar). Scientific Session of Moscow Engineering Physics Institute. Proceedings, Vol. 4. Moscow: MEPhI, 2009.

[Solovyev, Fashutdinov, 2009-3] V.D. Solovyev, R.F. Fashutdinov. Metodika ocenki stabil'nosti grammaticheskih svoystv (Technique of an estimation of stability of grammatical properties). Izvestiya RAN. Seriya literaturi i yazika. Vol.68. № 4. 2009.

[Starostin_DB] An Etymological Database Project. URL: http://starling.rinet.ru/main.html

[VerbNet] VerbNet Project. URL: http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

[WALS] The World Atlas of Language Structures Online. URL: http://wals.info/

[Wichmann, Holman, 2009] Wichmann, Søren and Eric W. Holman. 2009. Assessing Temporal Stability for Linguistic Typological Features. München: LINCOM Europa. 82 p.

[Wichmann, Saunders, 2007] Wichmann, Søren and Arpiar Saunders. 2007. How to use typological databases in historical linguistic research. Diachronica 24.2: 373-404.

[WordNet Online] Princeton University "About WordNet." WordNet. Princeton University. 2010. URL: http://wordnet.princeton.edu

[Yaroslavtseva, 2005] Yaroslavtseva E.I. Komp'yuternaya baza dannih "Jaziki Mira" i ee vozmozhnie primeneniya (Computer database « Languages of the World » and its possible applications). Dr.of S. Thes. – IL RAS, 2005.

Authors' Information

***Vladimir Polyakov** – Senior Researcher at the Department of Applied Linguistics of Institute of Linguistics of Russian Academy of Sciences, Assoc. Professor of National Research Technological University (MISIS) and Moscow State Linguistic Universit, PhD.*

Address: Leninsky Avenu, 4. Moscow. Russia; e-mail: pvn-65@mail.ru

Chair of Organizing Committee of International Conference "Cognitive Modeling in Linguistics.

Major Fields of Scientific Research: Cognitive Science and Modeling, Computer Linguistics, Artificial Intelligence.

Projects: www.dblang.ru ; www.cml.msisa.ru ; www.finforecast.ru