БАЗА ДАННЫХ «ЯЗЫКИ МИРА» И ЕЕ ПРИМЕНЕНИЯ. СОВРЕМЕННОЕ СОСТОЯНИЕ

Соловьев В. Д. (maki.solovyev@mail.ru)
КФУ, Казань, Россия

Поляков В. Н. (pvn-65@mail.ru)
НИТУ МИСиС, Москва, Россия

Ключевые слова: лингвистические базы данных, типология, квантитативные методы, ареальная лингвистика, языки мира

DATABASE “LANGUAGES OF THE WORLD” AND IT’S APPLICATION. STATE OF THE ART

Solovyev V. D. (maki.solovyev@mail.ru)
Kazan Federal University, Kazan, Russia

Polyakov V. N. (pvn-65@mail.ru)
National University of Science and Technology “MISiS”, Moscow, Russia

The article is dedicated to the largest digital resource in the world that contains a uniform description of language grammars — typological database “Languages of the World” (“Jazyki Mira”). There is information on the contents of the database, the programs for data procession. The database “Languages of the world” has three main areas of application: it can be used for quantitative researches, as a reference linguistic resource and for educational purposes. We give examples of database application in scientific researches in typology and areal linguistics. The examples demonstrate new opportunities of studying such questions as stability of grammatical features, liability to borrowing, typological and areal classification of languages. “Languages of the World” is compared with another famous typological database WALS.

Keywords: linguistic databases, typology, quantitative methods, areal linguistics, languages of the world, Jazyki Mira

1. Introduction

At the turn of the century there appeared various digital linguistic resources aimed at supporting of linguistic researches. An important place among them belongs to typological databases (TDB) that contain the descriptions of formalized grammar features of the languages. The development of this area began with small databases (DB) dedicated to a rather limited number of features, which contained the description of a small number of languages. Examples of such databases and a general review of TDB application can be found in [14, 16].

The new stages of TDB development began with the appearance of The World Atlas of Language Structures (WALS) [6] and database “Languages of the World” («Jazyki Mira»). The latter was created in Institute of Linguistics of Russian Academy of Science (IL RAS) on the base of a series of monographs of the same name (16 volumes). The first publications on this database are [7, 12]. The database is available in the Internet at http://dblang2008.narod.ru/, www.dblang.ru.

WALS and “Languages of the World” can be called big typological databases; each contains over 1 million bits of information. WALS describes over 2,500 languages by 142 features (128 of them are grammatical ones), and each of them has one of a few meanings: from 2 to 9. “Languages of the World” has the descriptions of 315 languages by 3,821 binary features. Both databases embrace all parts of grammar.

Examples of features: free word order, presence of ergative and absolutive cases, presence of exactly 5 monophthongs, etc. The set of features was formed as a result of systematic study of language grammars with the initial development of a formalized model of grammar description, and it was replenished after “Languages of the World” monograph had been written. The aim of the development of the set of features was the most detailed and precise description of grammar. The set of features is open, and it can be broadened when new languages are added.

TDB were initially created as reference books with a user-friendly interface, which helped quickly find the necessary information. But it soon turned out that TDB give us essentially new opportunities to study grammars of the languages by applying mathematical (including statistical) and computational methods. Many phenomena, which were until now regarded only on the qualitative level and on the base of separate examples, can now be studies by quantitative methods and with use of huge arrays of information. An important aspect of such studies is their objective character based on the application of strict mathematical methods. There are several types of the questions, which can be answered with help of TDB.

1. How homogeneous is this or that language areal? Can it be considered a language union? TDB allow applying of quantitative methods in areal linguistics for the estimation of the level of language proximity.
2. How were linguistic features spread during the spreading of the humanity and linguistic evolution? J. Nichols’ [10] conducted her pioneer researches in this direction on a very limited data access. Modern TDB can help define more exactly many aspects of humanity settlement.
3. Linguistic dynamics: what is the speed of grammar changing? What parts of grammar change faster?

4. What grammar features are easier to borrow during linguistic contacts?
5. Typological classification of the languages.

The article contains the description of the DB "Languages of the World" and of the program instruments it uses, it also gives examples of its application.

2. Structure and software of the database "Languages of the World"

2.1. Composition and structure

DB "Languages of the World" presents the following language families: Austro-Asiatic, Austronesian, Altaic, Afroasiatic, Indo-European, Kartvelian, North Caucasian, Sino-Tibetan, Uralic, Hurro-Urartian, Chukchi-Kamchatkan, Eskimo-Aleut and several isolate languages. The wide range of linguistic families, presented in the DB, justifies the name "Languages of the World". The database is constantly expanded as new monographs of the series are published. This work is conducted in the sector of areal linguistic of IL RAS under the guidance of A. A. Kibrik[1]. There are 10 more volumes planned for publication.

The DB has a genetic reference, which was developed in IL. In general, it corresponds to the classification from [2]. It contains 4 levels: families, branches, groups, subgroups.

The languages are described by a list of features and categories, which was called "Abstract model" in [7], and includes 3,821 features. The description of each language, i.e. a set of meanings of the features, is called its abstract. All languages' abstracts can be found at the web-site of the project: www.dblang.ru. The features are organized in a hierarchy. The top level of the hierarchy: 1.1. Phonemic structure. 1.2. Prosodic phenomena. 1.3. Phonetically motivated processes. 1.4. Syllable. 2.1. Phonological structure. 2.2. Phonological oppositions of morphological categories. 2.3. Phonologically motivated alternations. 3.0. Morphological type of the language. 3.1. Criteria of definition of parts of speech. 3.2. Nominal classifications. 3.3. Number. 3.4. Case meanings. 3.5. Verbal categories. 3.6. Deictic categories. 3.7. Parts of speech. 4.0. Paradigms. 5.1. Word form structure. 5.2. Word formation. 5.3. Simple sentence. 5.4. Composite sentence.

The abstract of a language contains about 300–350 features. 50 languages can be considered poorly described: their abstracts contain less than 200 features. The Russian language is obviously over-described: 536 features.

While using the DB we found mistakes in the data. An expertise was conducted for 30 randomly chosen languages in order to reveal them. On average, less than 3% of feature values were wrong. These mistakes have a different character and are mainly connected to the indistinctness of defining linguistic categories and subjectivism of the researches who described the language. We believe that at the current

[1] http://iling-ran.ru/beta/departments/typol_compar/areal

level of linguistic data formalization it is impossible to eliminate all disagreements in different experts’ interpretations. The database WALS also contains mistakes and contradictions, but they do not influence the results of statistical calculations, as the latter proceed big data arrays, and the mistakes are leveled.

The comparison of WALS and “Languages of the World”, conducted in [13], included building of phylogenetic trees for the same set of languages. It revealed a more serious problem of WALS resource when it is used for statistical calculations: a big number of gaps in the data. On average, languages in WALS are described on less than one third of the features. As a result, due to the lack of data, non-relative languages groundlessly drew closer. “Languages of the World” has a great advantage, as the languages (except for small number of the little-studied ones) are completely described, i.e. by all features.

2.2. Software

The software of the DB “Languages of the World” consists of a nucleus and research tools. The software of the DB “Languages of the World” solves the following tasks:

1) formation and management of the model and abstracts of the database;
2) search for information;
3) binary comparison of abstracts.

The module of binary comparison of abstracts shows lists of common features for the given pair, and also a list of features that are present only in one of the two languages.

The DB “Languages of the World” exists in form of a Web-version, Windows-version and Excel-version. The Windows-version of the DB is a 32-bit application, written in Delphi Pascal (version 7). Borland Database Engine is used as DBMS. The workspace is: Windows 95/98/2000/NT/XP. The volume of installation: 17.4 Mb. The volume of the program and the DB: 18.8 Mb.

The Excel-version gives easy-to-use opportunities for statistical calculations with help of in-circuit tools. Except the nucleus tools, some research tools were created for quantitative investigations. They are:

- Similarity program, for calculation of the level of language proximity;
- LangFam program, for calculation of language portraits of families of languages and revelation of genetic markers.

Standard phylogenetic algorithms, programs for multidimensional scaling and principal component analysis can be applied.

The easiest way to calculate the level of language proximity is Hamming’s metrics (number of unmatched features). Besides Hamming’s metrics, Similarity program provides the calculation of a few other studied measures of language proximity. Moreover, Similarity is an adjustable program. It allows varying different parameters, e.g. choosing groups of features, according to which the calculation of the distance between languages will be implemented. This program helped revealing metrics of calculation of language proximity that describe genetic trees with a high level of precision (up to 80% of match with traditional views) [5].

LangFam program was written in VBA language; it is designed for calculation the frequency of features by all families of languages of all genealogical level that are present in the DB, and by all DB in general. LangFam program helped revealing such phenomenon in the development of the languages as typological shift. It main point is the following: during the linguistic evolution and contacts the feature space is partially "polarized" (most rare features become even more rare, and most widely spread features — even more spread) [11].

The database is constantly replenished with new information and renewed. The version of 2013 is written in C# with use of ASP.NET library and, thus, it requires Microsoft.NET Framework 2.0 and higher. There is a possibility of uploading abstracts from text files. The total volume of installation version is 99 Mb[2]. The program gives a more user-friendly interface for viewing of the main data of the base, it includes annotations of features, examples and references to the source article about the language in the encyclopedia (quantized into pdf). It has more powerful search facilities than the previous version. It also includes "Glossary", which gives a definition of all terms of the language description model; genetic reference; geographic reference, which contains the name of the area where the language is used and geographic coordinates of its center (according to UNESCO's atlas); English translation of features; English names of the language; language code according to ISO 639-2 (Ethnologue, www.ethnologue.com).

3. Examples of application in scientific researches

3.1. Typology

3.1.1. Typological classifications of languages

Evidently, the first serious attempt to classify the languages by their structure is morphological classification, developed in the early 19-th century in Schlegel's, Humboldt's and Schleicher's works. This classification still remains meaningful, but it takes in consideration only one aspect of the linguistic structure: the way morphemes are joined, so, languages, which belong to one class, according to this classification, can radically differ from each other in other aspects.

Other existing typological classifications of languages are also based on one or a few features. It remains unclear, whether holistic classification of languages is possible. It divides all languages into several groups, so that languages within one group are typological homogeneous, and there are sufficient typological differences between the languages from different groups, and they differ in a wide range of features that embraces all main levels of the language.

---

[2] The increase of the volume is due to the big number of graphic materials in the articles of the encyclopedia

With the appearance of big typological databases, like WALS and “Languages of the World”, it becomes possible to build classifications that consider hundreds and thousands of features at the same time.

For the first experiment we chose 27 languages that represent all families and isolate languages from our DB. We apply the well-known phylogenetic algorithm NeighborNet, which was developed in bioinformatics. The results are presents in Pic. 1, where close position of the languages means shorter distance between them, i.e. means bigger typological similarity between them.

In general, the arrangement of the languages in pic. 1 is rather even. Nevertheless, in pic. 1 we can see, though not very clear, 5 main clusters of languages by typological similarity: Indo-European, Uralic-Altaic, Caucasian (probably, with Chukchi and Ket), Far Eastern (several isolate languages) and Afroasiatic. Noteworthy, typological proximity correlates well with linguistic kinship and areal proximity. Thus, Indo-European languages proved to be typologically close, despite the fact that they dispersed from Proto-Indo-European at least 6 thousand years ago and are now spread over a big territory. Nevertheless, during this time they have not acquired such features as vowel harmony (which is characteristic of Turkic languages), incorporation (characteristic of Chukchi-Kamchatkan languages), etc. As a result, typologically, modern Indo-European distinctively differ from Turkic languages, for example. The differences between Proto-Indo-European and Proto-Turkic languages have not been smoothed during this time. Caucasian languages proved to be typologically close, despite their attribution to three different families. This indicates the importance for typological proximity of not only common origin, but also of long-term contacts (several thousand years for Caucasus).

Separate common features can be found in non-relative languages that are located far from each other. Thus, “qualitativeness” (way of action, №2122 in the DB) is found in Aleut and Ethiopian, but it is absent in other languages of the region. Such cases of parallel evolution are rare and they do not influence the general image.

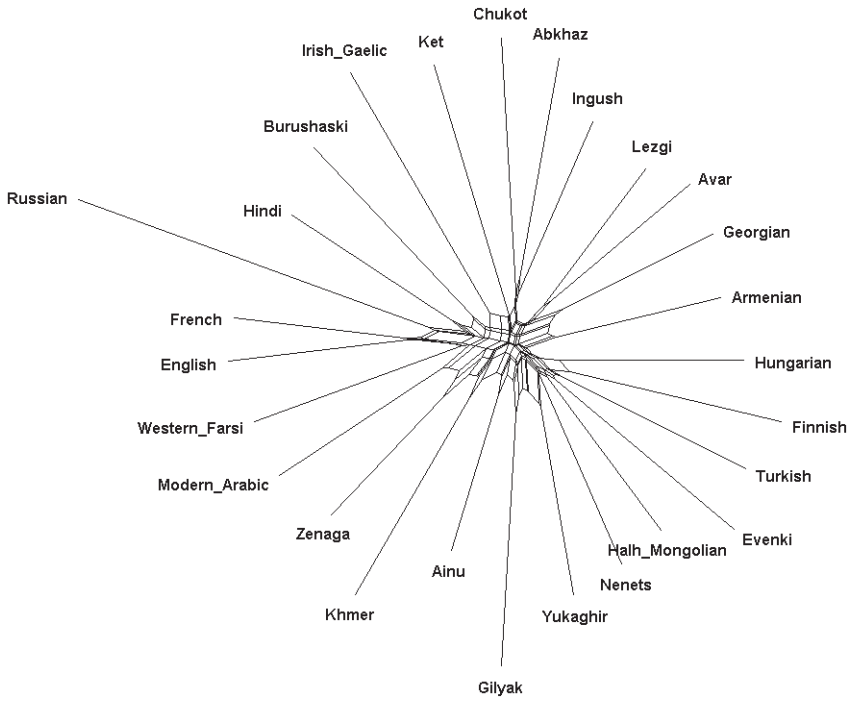

Fig. 1. Languages of Africa and Eurasia according to the data of "Languages of the World"

We shall note that such diagrams do not have an absolute character. Some languages are very strangely positioned in diagrams of this type. For example, the Irish language (Celtic branch of Indo-European family) is placed between Ket and Burushaski. In fact, these languages are not typologically close, they are not related and are geographically very far from each other. The possible reason of such placement if insufficiency of the given method of graphic representation of information for absolutely precise reflection of typological proximity between all pairs of languages. From the mathematical point of view, the DB "Languages of the World" represent languages as points in 3821-dimentional space of features. At the same time, the diagram built by NeighborNet is equivalent to 1-dimentional representation (in a circle). Obviously, when 3821-dimentional space is rolled into 1-dimentional space, there can be distortions.

With help of Similarity program one can find out that Irish is still closer to English (the distance is 285) and Persian (280) that to Burushaski (301). Thus, as well as other tools of computer linguistics (like ancient texts recognition), phylogenetic algorithms require certain post-editing. Nevertheless, these phylogenetic algorithms are more and more widely applied in comparative linguistics, as they allow quickly receiving a rather good result. This method can be compared with Greenberg's method of mass comparison. Articles with use of typological databases and phylogenetic algorithms are published in leading journals, such as Language and Science. A number

of works note that in cases when questions of linguistic kinship have been reliably defined, it turns out that the results of phylogenetic algorithms coincide with the stated ones in 80% of cases.

We should note, even Indo-European languages have not been completely studied from the point of view of their evolution tee reconstruction. For example, [2] enumerates 136 modern Indo-European languages. If their evolution tree was completely studied (i.e. was binary), it would contain 135 tops, which would conform to protolanguages. But the tree presented in [2] contains only 26 tops, i.e. less than a fifth part.

3.1.2. Stability of grammatical features

Let us study the question of the stability of grammatical features. The key idea in the estimation of stability consists in comparison of the prevalence of a feature among related and non-related languages. The biggest part of researches are based on this idea and specify it. The first quantitative researches of stability were conducted by Nichols [10]. She suggested several variants of stability measures. Unfortunately, she did not have a big typological database, which prevented a wide verification (with a big number of languages and features) and spread of her approach.

In [15] there are examples of defining 4 measures of stability of grammatical features. The first one was suggested by Nichols (measure 3 in [10]), the second one was suggested by Wichmann and coauthors [17], the third measure was suggested by Maslova [9], the fourth measure was suggested by one of the authors of the present article, and it is the only measure that realizes the idea of calculation of the number of changes of a feature values during the evolution. Philogenetic algorithms of evolution trees reconstruction are often used for it.

The comparison of these measures on the material of the DB “Languages of the World” showed that there is good correlation between the first and the fourth measures, and also between the second and the third. It is shown that a generalized measure of stability received on the basis of all four measures, in most cases coincide with the qualitative evaluation, previously published in typological literature.

The comparison of measure 2 for WALS and “Languages of the World” was conducted in [15, 1]. There were chosen 23 features of WALS (or, to be more precise, values of features) that match or are very similar in WALS and in “Languages of the World”. In most cases data on the stability of features, calculated by both bases, match or are very close. Reasons for the cases when a considerable mismatch takes place require separate study.

3.1.3. Borrowing of features

TDB allow to systematically studying the inclinations of features and groups of features to borrowings. In [3] B. Comrie used WALS to evaluate the number of matching and mismatching features for three languages: Egyptian, Maltese and Spanish, considering that Maltese is related to Egyptian, but it was in contact with Spanish for a long time. We studied groups of three languages with an analogous structure for “Languages of the World”. For the group of Hungarian, Romanian and Khanty with a contact situation Hungarian-Romanian and genetic relationship Hungarian-Khanty we looked at the features common for Hungarian and Romanian and

different for Khanty. One can suppose that they (at least, part of them) were borrowed from the Romanian language to Hungarian. It turned out that among phonological features they made 20.8%, among morphological — 21.6%, and among syntactical — 19.6%. We studied two more groups with the Hungarian language: Hungarian-Slovak-Khanty and Hungarian-German-Khanty. The averaging of the results gave us the following data: among phonological features the percentage of presumable borrowed one in Hungarian made 20.9%, among morphological features — 19.8%, among syntactical features — 17.4%. We received similar results for other regions and language families. For the group of Finnish, Swedish, Komi-Zyrian the corresponding numbers are: 21.7%, 15.6%, 14.6%; for the group of Polish, German, Macedonian: 15.5%, 13.9%, 7.5%, for the group of Tatar, Mari, Turkish: 19.3%, 16.6%, 24.3%.

It is generally accepted that phonology is easier to borrow than morphology and syntax. We showed that, despite it being true, the difference (especially between phonology and morphology) is not great at all, and in some certain situations this regularity may not be followed at all.

3.2. Areal linguistics

During the past years areal linguistics has become one of the most dynamically developing branches of linguistics. A. E. Kibrik [8] noted that studying of areal connection is especially important for the explanatory approach, which will allow linguistics to move closely to understanding of the essence of the language. But one of the central notions of areal linguistics — concept of language union (LU) is being criticized (e.g. L. Campbell in [16]), due to vagueness and indistinctness of the definition of LU. Use of TDB allows making the definition more precise.

We shall introduce a quantitative measure of convergence of languages in a regional community. Let us assume that there are two groups of languages G1 and G2 from two families S1 and S2 in the studied region. We will calculate an average distance d(G1, G2) between languages from groups G1 and G2 and an average distance d(S1, S2) between languages from families S1 and S2. We will call the difference $R = d(S1, S2) - d(G1, G2)$ measure of convergence of languages in the region. When $R = 0$ (and $R < 0$) there are no grounds to postulate any presence of LC in the region. The bigger R is, the stronger the LC is, i.e. the languages of the region drew closer as a result of borrowings. Similar calculations can be made for languages from several families.

For testing we shall apply the suggested method to the classic LC — Balkan. Balkan LC traditionally includes South Slavonic and Balkan-Romanic languages, and also Albanian and Greek. Unfortunately, Albanian and Greek are not presented in “Languages of the World”. So, we apply the approach to the remaining two groups that belong to the Slavonic and Romanic branches of the Indo-European family.

Calculations show that the average distance between Balkan-Romanic and South Slavonic languages equal 237, while the distance between Romanic and Slavonic languages in general equal 243. As it was presumed, the distance between all Romanic and all Slavonic languages is bigger. Although the difference $243 - 237 = 6$ does not seem so big, it still shows higher typological proximity between Balkan-Romanic and

South Slavonic languages, which cannot be explained by their relationship, as their genealogical proximity is the same as between Romanic and Slavonic languages.

Similar results were received for Volga region, which is inhabited by speakers of Turkic and Finnish-Ugric languages. However, further broader studies may require us to specify the suggested simple formula.

4. Conclusion

The aim of this article was not only to present results, which are still very far from final, but to outline a general way one which one can receive new results of quantitative character in typology and areal linguistic with use of typological databases. There are a lot of methodological problems to be solved on this way.

TDB help receive objective numerical estimations of such characteristics of grammatical features as level of stability, inclination to borrowings. These data can be used in researches on the language evolution. Although it is still impossible to give statistically reliable temporal estimations of feature stability, probably, the most stable grammatical features appeared dozens of thousands of years ago, which allows getting deep inside the history of languages.

Educational program “Databases for Typological and Comparative Researches” was developed on the material of the DB “Languages of the World”. This program was tough as an optional course at the philological faculty of Moscow State University (Department of Theoretical and Applied Linguistics, Moscow) and at linguistics department of South Ural State University (Chelyabinsk).

References

1. *Belyaev O.* (2009) Stability of language features: a comparison of the WALS and JM typological databases. Proceedings of the Int. Conf. “Cognitive Modeling in Linguistics”, FCCL, available at: http://fccl.ksu.ru/conf_CML_2008/jm-wals-stab-2.doc
2. *Burlak C. A., Starostin C. A.* (2005) Vvedenie v lingvisticheskuju komparativistiku [Introduction to comparativistics]. Moscow, Academija.
3. *Comrie B.* (2009) Maltese and the World Atlas of Language Structures. Introducing Maltese Linguistics. Amsterdam, Philadelphia, Benjamins, pp. 3–11.
4. *Everaert M., Musgrave S., Dimitriadis A.* (Eds.) (2009) The Use of Databases in Cross-Linguistic Studies, Berlin, Mouton de Gruyter.
5. *Gusareva U.* Measures of similarity as basis of quantitative researches in the field of historical linguistics. Proceedings of the Int. Conf. “Cognitive Modeling in Linguistics’XI”, Vol. 2, Kazan’, 2009, pp. 391–409.
6. *Haspelmath M., Dryer M., Gil D., Comrie B.* (Eds.) (2005) The World Atlas of Language Structures, Oxford, Oxford University Press, 2005.
7. *Jaroslavtceva E. I.* (2005) Database “Language of the World” and its applications [Kompjuternaja baza dannyh “Jazyki mira” i ee vozmozhnye prilogenija], Linguistic science doctor thesis, Moscow, IJAz RAN.

8. *Kibrik A. E.* (2003) Konstanty i peremennye jazyka [Constants and variables in language]. Sankt-Peterburg, Aleteja.
9. *Maslova E.* (2004) Dynamics of typological distribution and stability of language types [Dinamica tipologicheskih raspredelenij i stabil'nost' lazykovyh tipov], Voprosy jazykoznanija [Problems in linguistics], no. 5, pp. 3–16.
10. *Nichols J.* (1992) Linguistic Diversity in Space and Time, Chicago, London, The University of Chicago Press.
11. *Polyakov V. N., Yaroslavtseva E. I.* (2008) Quantitative laws of typological shift in the Eurasian languages (based on "Languages of the world" data base) [Kvantitativnye zakony tipologicheskogo sdviga i jazykah Evrazii]. Uchenie zapiski Kazanskogo Gosudartsvennogo Universiteta. Seriya Gumanitarnie Nauki. [Scientific notes of Kazan State University. Humanity Series], Vol. 150, Book 2, pp. 97–118.
12. *Polyakov V. N., Solovyev V. D.* (2006), Komp'uternye modeli i metody v typologii i komparativistike [Computer models and methods in typology and comparativistics]. Kazan, Kazan University.
13. *Polyakov V., Solovyev V., Wichmann S., Belyaev O.* (2009) Using WALS and Jazyki mira, Language Typology, Vol. 13, pp. 135–165.
14. *Solovyev V. D.* (2010) Typological databases: perspectives of using [Tipologicheskie basy dannyh: perspectivy ispol'zovanija], Voprosy jazykoznanija [Problems in linguistics], no. 1, pp. 94–110.
15. *Solovyev V. D., Faskhutdinov R. F.* (2009) Evaluation method for stability of grammar features [Metodica otcenki stabil'nosti grammaticheskih svoystv], Izvestija RAN. Serija jazyka I literatury [RAS News. Language and Literature series], Vol. 68, № 4, pp. 44–57.
16. *Vinogradova V. A., Novikov A. I., Jaroslavtceva E. I.* (2003) Database "Languages of the World" as the tool for linguistic researches [Baza dannyh "Jazyki mira" kak instrument lingvisticheskih issledovanii], Voprosy jazykoznanija [Problems in linguistics], no. 3.
17. *Wichmann S., Holman E.* (2009) Assessing temporal stability for linguistic typological features, München, LINCOM Europa.