Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 1 31–50
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-26460

Research article

Phylogenetic trees: Grammar versus vocabulary

Vladimir N. POLYAKOV, Elena A. MAKAROVA[1] ✉, Valery D. SOLOVYEV[2]

[1]*Institute of Linguistics of Russian Academy of Sciences, Moscow, Russia*
[2]*Kazan Federal University, Kazan, Russia*
✉MakarovaEA@iling-ran.ru

Abstract
Traditionally, genealogical relationships between languages are established on the basis of phonetic and lexical data. The question whether genealogical relationships among languages can be defined based on grammatical data remains unanswered. The objective of this article is to compare two phylogenetic trees: one built using the Automated Similarity Judgment Program (ASJP) project, and one using the World Atlas of Language Structures (WALS). We include data from WALS representing 27 languages from 5 language families of all continents that are deemed to be sufficiently well described. A Hamming distance matrix was calculated for all languages under study, and, based on the matrix, a phylogenetic tree was built. The trees built according to WALS and ASJP data are compared with each other and with a tree built by the classical comparative historical method. Both the ASJP-based tree and the WALS-based tree have their advantages and disadvantages. The ASJP-based tree is a good reflection of the evolutionary divergence of languages. Similarities of languages as calculated based on the typological database of WALS can provide information on the history of languages both in terms of genealogical descent and contact with other languages. The ASJP-based tree reflects genealogical relationship well at a relatively small time depth, while the WALS-based tree reflects genealogical relationship well at large time intervals. We suggest a new variant of a phylogenetic tree that includes information on both the divergence (ASJP project) and the convergence (WALS project) of languages, combining the benefits of both of these trees, although the problem of borrowings remains. The present research reveals prospects for future studies of genealogical relations among languages based on large-scale descriptions of their grammatical structures.

Keywords: *typology, quantitative linguistics, computational linguistics, historical linguistics, phylogenetics*

For citation:
Polyakov, Vladimir N., Elena A. Makarova & Valery D. Solovyev. 2022. Phylogenetic trees: Grammar versus vocabulary. *Russian Journal of Linguistics* 26 (1). 31–50. https://doi.org/10.22363/2687-0088-276460

Научная статья

Филогенетические деревья: грамматика против словаря

В.Н. ПОЛЯКОВ, Е.А. МАКАРОВА[1] ✉, В.Д. СОЛОВЬЕВ[2]

[1]*Институт языкознания РАН, Москва, Россия*
[2]*Казанский федеральный университет, Казань, Россия*
✉MakarovaEA@iling-ran.ru

Аннотация
Традиционно генеалогические отношения между языками устанавливаются на основе фонетических и лексических данных. Вопрос о том, можно ли определить генеалогические отношения между языками на основе грамматических данных, остается без ответа. Цель этой статьи – сравнить два филогенетических дерева: одно построено с использованием проекта Автоматизированной программы оценки сходства (ASJP), а другое – с использованием Всемирного атласа языковых структур (WALS). Мы включаем данные из WALS, представляющие 27 языков из 5 языковых семей всех континентов, которые достаточно хорошо описаны. Для всех исследуемых языков была рассчитана матрица расстояний Хэмминга и на ее основе построено филогенетическое дерево. Деревья, построенные по данным WALS и ASJP сравниваются между собой и с деревом, созданном с опорой на классический сравнительно-исторический метод. И у дерева на основе ASJP, и у дерева но основе WALS есть свои преимущества и недостатки. Дерево на основе ASJP указывает на расхождение языков в процессе эволюции. Сходство языков, рассчитанное на основе типологической базы данных WALS, может предоставить информацию об истории языков как с точки зрения генеалогического происхождения, так и с точки зрения контакта с другими языками. Дерево на основе ASJP хорошо отражает генеалогическое родство на относительно небольшой временной глубине, а дерево на основе WALS – на больших временных интервалах. Мы предлагаем новый вариант филогенетического дерева, который включает информацию как о дивергенции (проект ASJP), так и о конвергенции (проект WALS) языков, объединяя преимущества обоих этих деревьев, хотя при этом остается проблема учета заимствований. В настоящей работе обозначены перспективы будущих исследований генеалогических отношений между языками на основе крупномасштабных описаний их грамматических структур.
Ключевые слова: *типология, квантитативная лингвистика, компьютерная лингвистика, историческая лингвистика, филогенетика*

Для цитирования:
Polyakov V.N., Makarova E.A., Solovyev V.D. Phylogenetic trees: Grammar versus vocabulary. *Russian Journal of Linguistics.* 2022. Vol. 26. № 1. P. 31–50. https://doi.org/10.22363/2687-0088-26460

Introduction

In this paper we are interested in the extent to which structural features of languages are useful for language classification. We are not looking for a replacement of traditional methods of language classification, but we are interested

in learning more about whether structural features can constitute a useful alternative or complement. Claims to this effect exist and deserve to be investigated, and there is a growing body of structural data available which deserve exploration.

As tools for depicting the similarity of languages calculated either through the comparison of lexical or structural features we will be using tree diagrams. Often we use the term 'phylogenetic tree' to refer to these tree diagrams, even if not all aspects of the trees can be interpreted as depicting evolutionary relationships. Strictly speaking, the trees should be referred to as 'phenetic' rather than 'phylogenetic' trees. We urge the reader to keep in mind this broad and relatively vague use of the term 'phylogenetic'.

The problem of defining the genealogical relationship of languages based on grammatical data has a long history. By the 20th century linguists had become aware of structural similarities among languages and were discussing how to interpret them (Trubetzkoy 1939, Benveniste 1954). The dominant point of view in modern linguistics is that genealogical relationships among languages can be established only by the comparative method, through comparisons of lexical and grammatical morphemes, whereas structural features of a language are believed to be more prone to borrowing and, consequently, cannot be a reliable source for defining the family connections between languages. Nevertheless, the discussion about the usefulness of typological features in historical linguistic research is not closed, and it is not much more than a decade ago that sufficiently large databases have become available so as to provide fodder for substantive discussion of the matter. In the following we will briefly review the pertinent literature.

The first attempt to identify deep genealogical relations among languages based on 31 typological features and modern phylogenetic methods was in Dunn & Foley & Levinson & Reesink & Terrill (2005). A discussion ensued between defenders of this approach (Dunn et al. 2007, Dunn et al. 2008) and critics, who argued that structural similarities are prone to come about through contact (Donohue & Wichmann & Albu 2008, Gray & Bryant & Greenhill 2010, Donohue et al. 2011, Wichmann & Holman 2010).

Among those doubting that typological features might finally give us the holy grail of historical linguistics, which is to reach further down in time than the traditional comparative method allows, there have been some who nevertheless paid attention to the utility of typological features in historical linguistic research, showing that more stable features can at least lead to better classification results than some less stable features (Wichmann & Saunders 2007), and it has been shown that many typological features have a rate of change comparable to what has been inferred by lexicostatistics (Swadesh 1955) for the basic lexicon (Wichmann & Holman 2009).

One of the studies that leaves some potential room for typological features to contribute productively to language classification, even if they might not be suitable for establishing far-flung relationships, is Holman et al. (2008). Here it was shown

that a similarity measure based on both typological and lexical data, weighted such that the typological data accounted for one quarter of the measure and lexical data accounted for three quarters, led to better results than use of lexical data alone. It is still an open question how to explain this result. Perhaps it is the mere addition of more data that caused the improvement. Possibly the sensitivity of the typological features to areal influence could have a positive effect on the classifications inasmuch as geography does influence family trees, genealogical closer languages tending to also be geographically closer than more distantly related languages. It may also be the case that typological features do exhibit a sufficient genealogical signal (in addition to the areal signal) to be directly phylogenetically informative.

In 2009, a study was conducted (Polyakov et al. 2009) one of whose goals was to compare phylogenetic trees built on lexical and phonetic data of ASJP and on structural data from two typological databases: WALS (Dryer & Haspelmath 2013) and the database known as "Languages of the World", compiled at the Institute of Linguistics of the Russian Academy of Sciences (Anisimov et al. 2013). The study demonstrated a considerable advantage of the lexical and phonetic data of ASJP compared to the structural data from WALS and "Languages of the World" with respect to the quality of the language classification based on these different datasets.

Some applications of computational methods in historical linguistics have tried to emulate the framework of the comparative method that developed in the wake of works such as (Bopp 1885) more than a century ago (Ringe, Warnow & Taylor 2002, Nakleh, Ringe & Warnow 2005). Most computational linguistic phylogenies, however, are based on standard lists of concepts, following in the tradition of lexicostatistics (Swadesh 1950, 1952, 1955, Burlak & Starostin 2005) but using modern character- or distance-based algorithms for inferring the trees (Dunn 2015, Wichmann 2017a). Among character-based methods, Bayesian ones in different software realizations such as MrBayes, BayesTraits or BEAST (e.g., Pagel & Meade 2006) are routinely used, and have served to produce phylogenies of several language families including Austronesian (Gray & Jordan 2000), Chapacuran (Birchall, Dunn & Greenhill 2016), and Dravidian (Kolipakam et al. 2018), even if a simulation study indicated that another type of algorithm (Maximum Parsimony) may actually perform better (Barbançon et al. 2013).

Distance-based methods usually involve string similarities (Rama & Borin 2015) and have also been used in the classification of many families, including Austronesian (Wichmann & Rama 2018), just to mention one. Since such methods do not require the assumption that the languages to be classified are related, they have moreover been applied to a large subset of the entire range of the world's languages (Müller et al. 2013). It deserves mentioning that the linguistic distances feeding into distance-based phylogenetic methods also have a potential for investigations of other aspects of language dynamics (Wichmann & Good 2014), such as processes of language divergence (Holman & Wichmann 2017) or the modeling of historical linguistic processes (Wichmann 2017b).

Finally, and closer to the spirit of the present paper, grammatical features have also sometimes been drawn upon for phylogenetic inferencing, e.g., Longobardi et al. (2015). Moreover, such features have been used to find genealogical relationships among languages (Polyakov et al. 2016), for identifying cases of language contact, and for dating language divergence (Solovyev 2009). Using grammatical features for these kinds of historical linguistic purposes is a recent endeavor, so it seems quite reasonable to expect some new results and further developments. For instance, there are types of structural traits that have not yet been used for such purposes, including morphological data.

In the present paper, we wish to continue the discussion about the usefulness of typological features for linguistic phylogenetics. After describing how our data were selected, we build a based on typological data from WALS and contrast it with a tree based on lexical ASJP data. As mentioned, Holman et al. (2008) obtained meaningful results by mixing the two types of data, but we do not carry out a similar exercise here because when mixing the two kinds of data it becomes unclear what the contribution is of each to the results.

1. Materials and methods

1.1. Materials

The present study is based on materials from the two currently largest published linguistic databases (i.e. databases that embrace the biggest number of languages): ASJP (Wichmann, Holman & Brown 2016) and the World Atlas of Language Structures (WALS[1]) (Dryer & Haspelmath 2013).

ASJP was officially launched in 2008 (Brown et al. 2008). It is a project based on the application of computational methods in comparative linguistics using 40-item lists of basic vocabulary of languages. At present, ASJP contains word lists from about 7000 existing languages and dialects, including creole languages, pidgins, mixed languages, and language isolates, and it is constantly being broadened.

One of the original goals of the ASJP project was to create a universal method that would allow linguists to automatically calculate the similarity of words with the same meaning in different languages, and, based on that, to define genealogical relationships among languages, including as yet unclassified ones. In addition, other foci of research has emerged from this project, including the identification of homelands of language families (Wichmann, Müller & Velupillai 2010a), the evaluation of different phylogenetic methods (Wichmann & Holman 2010b), and others.

Pompei, Loreto & Tria (2011) presents an objective analysis of the performance of ASJP. The authors of the study compared phylogenetic trees built

[1] We used the data from the WALS Program (Dryer and Haspelmath 2013), which, largely coincides with the ALS Online version.

on the methods and database of ASJP with trees built manually based on Ethnologue (Lewis 2009). While the performance of ASJP was found to be variable it remains the only existing project allowing linguists to quickly build a phylogenetic tree for any set of languages.

The WALS typological database was developed in 2005 (Haspelmath et al. 2005). The first version of this project, as embodied in the WALS Program, contained data on 2560 languages and 140 features. Later, in 2011, the authors released the online version (Dryer & Haspelmath 2013), which now exhibits 2679 languages and 192 features. The project was originally undertaken by 55 specialists.

WALS Online ensured consistency with the first version in terms of language codes and feature numbers. Since its creation, WALS has been used in a great number of studies, such as (Coloma 2017), to name but one recent paper, and this number keeps growing. Unfortunately, many languages, both in the WALS Program and in WALS Online, are incompletely attested. The fact that it represents a sparse data matrix requires users embarking on any kind of statistical analysis to implement a selection procedure. (Polyakov et al. 2016) contains a detailed description of a procedure for cleaning WALS data and selecting languages based on their degree of attestation, and we follow the same procedure here.

As a first step it is necessary to filter some features and feature values of WALS for the specific purpose at hand. The original goals of WALS would have been broader than that of facilitating a structural comparison of languages. In fact, it was designed by typologists and clearly not intended for phylogenetic uses. Therefore, the database includes lexical features relating to the encoding of lexical categories such as ‘hand’, ‘arm’, ‘finger’, numerals, colors, and ‘tea’. These features evidently have nothing to do with the structure of a language, and for this study, they were all excluded from the list of features that were compared for different languages. Furthermore, WALS also includes two features describing categories of sign languages. These features also were not taken into consideration. Finally, the values of some features, like ‘other’ and ‘not reported,’ which do not necessarily denote the same phenomenon in different languages, will introduce noise into comparisons. These values were also excluded from our data. The present work applies these filters.

Next, also following (Polyakov et al. 2016), we select language pairs according to the following criteria, which should all be satisfied: (1) at least 26 features are attested for both languages; (2) at least 65% of the feature values match; (3) the pair should be above a regression line within the space of shared values as a function of overlapping features fitted such that all pairs involving unrelated languages are below the line; (4) one member of the pair is the closest matching language of the other language. Criterion (4) can cause a language to recur in different pairs if that language comes up when searching for language closest to A and then comes up again when searching for the language closest to language B. Swedish is a case where this happens. The results of this selection procedure are shown in Table 1.

Table 1. **Pairs of languages selected. WALS codes are given in brackets**

N	Target language	Structurally maximally close language	Family of both languages	% matching values	Number of overlapping features
1	Shona [shn]	Zulu [zul]	Niger-Congo	96.97	33
2	Russian [rus]	Ukrainian [ukr]	Indo-European	94.59	37
3	Swedish [swe]	Danish [dsh]	Indo-European	95.24	42
4	Hindi [hin]	Panjabi [pan]	Indo-European	92.16	51
5	Kongo [kon]	Nkore-Kiga [nko]	Niger-Congo	85.42	48
6	Iaai [iaa]	Drehu [dre]	Austronesian	85.71	49
7	Dutch [dut]	German [ger]	Indo-European	83.64	55
8	English [eng]	Swedish [swe]	Indo-European	85.94	64
9	Spanish [spa]	Italian [ita]	Indo-European	83.08	65
10	Modern Greek [grk]	Bulgarian [bul]	Indo-European	82.35	68
11	Cantonese [cnt]	Chinese [mnd]	Sino-Tibetan	76.47	68
12	Latvian [lat]	Polish [pol]	Indo-European	74.65	71
13	Malagasy [mal]	Paiwan [pai]	Austronesian	72.15	79
14	Navajo [nav]	Slave [sla]	Na-Dene	70.13	77

1.2. Methods

The main method used for the present research pertains to linguistic phylogenetics (Nichols & Warnow 2008). Modern phylogenetics methods were first developed in biology (Edwards & Cavalli-Sforza 1964, Felsenstein 2003) and were subsequently adopted by linguists as a way of defining the relationship between languages. Some of the first works in comparative linguistics to rely on computational phylogenetics are Gray & Jordan (2000), Ringe et al. (2002), and Gray & Atkinson (2003). Computational phylogenetic methods are by now widely acknowledged as providing valid ways of comparing and classifying of languages. Naturally, just like any other method relying on empirical data, a phylogenetic method is only as good as the data it is given as input. Thus, a computationally produced phylogenetic tree should ideally be based on well-studied and well-described languages. Here we also use a computational method for testing classifications based on the of languages just defined, using data from WALS in comparison with ASJP, and, as the previous discussion has explained, we carefully select the data used.

The list of the selected languages in Table 1 was used to build two trees: a tree based on ASJP data, which is lexical in nature but also contains phonological information inasmuch as words consist of phonemes, and a tree based on the data of the WALS Program, which includes only structural information.

In order to build a tree using the WALS data, for all the pairs of languages, the Hamming distance (Hamming 1950, Wong & Kim 2014), that is, the percentage of unmatching feature values, was calculated, and a distance matrix was built. Finally, using the Neighbor-Joining method as implemented in the MEGA 7 software (Kumar, Stecher & Tamura 2016), we built trees for the selected set of the languages.

In order to build a tree from the 40-item word lists in the ASJP database, the Neighbor-Joining algorithm was again used, this time applied to a matrix of

pairwise edit distances. Specifically, we use the twice-modified Levenshtein distance called LDND (Levenshtein Distance Normalized & Divided), which has been justified Wichmann et al. (2010b) and furthermore tested in Pompei, Loreto & Tria (2011). The LDND is a measure of phonological distance, which is also sensitive to lexical replacement inasmuch as a lexical replacement will incur a large phonological distance. If we were to look at phonological and lexical differences separately, as suggested by a reviewer, we would have to distinguish cognates from non-cognates. This would be beyond our expertise for most of the language pairs involved. Quoting from Wichmann (2013), the LDND "is based on the Levenshtein distance, a distance metric which counts the minimal number of operations (deletions, insertions, and substitutions) required to transform one word into another. The LDN distance between a pair of words is the Levenshtein distance divided by the length of the longer of the two words. Next, the LDND distance between two languages is defined as the average LDN distance between each pair of words with the same meaning, divided by the average LDN distance between each pair of words with a different meaning. The latter division is intended to control for similarity owing simply to similar phonemic inventories of the two languages." The 40-item subset of the Swadesh list given as input to the LDND is described in Holman et al. (2008).

The trees are presented and discussed in the next section. We are aware that a WALS-based tree could alternatively have been produced using a character-based method such as Maximum Parsimony, which has been found to perform better than Neighbor-Joining (Barbançon et al. 2013), but we prefer to compare trees that are produced using one and the same algorithm in order to enhance their comparability.

2. RESULTS AND DISCUSSION

2.1. Results

Figure 1 shows a tree built for the set of 27 languages singled out in Polyakov et al. (2016) and in Table 1 above. It is based on lexical (also implying phonetic) data from ASJP. Figure 2 presents a tree built for the same set of 27 languages using the grammatical data of WALS.

Both trees quite satisfactorily describe the relationship among the languages. Nevertheless, there are certain differences between the two trees that require our attention.

1. According to the ASJP tree, Dutch and Standard German are sisters and the node uniting them is a sister of English; according to the WALS tree, Danish and Swedish are sisters and together form a sister clade of English.

2. According to the ASJP tree, Greek is not closely related to any other Indo-European language. The WALS tree unites Modern Greek with Bulgarian.

3. The ASJP tree suggests the following sequence: after Greek, the first languages to split off from the Proto-Indo-European lineage were Balto-Slavic, then a clade consisting of Romance and Indic, and finally Germanic. According to the WALS tree, the first languages to separate were Indic, with Balto-Slavic—also including Greek—and Germanic constituting clades crystallizing later.

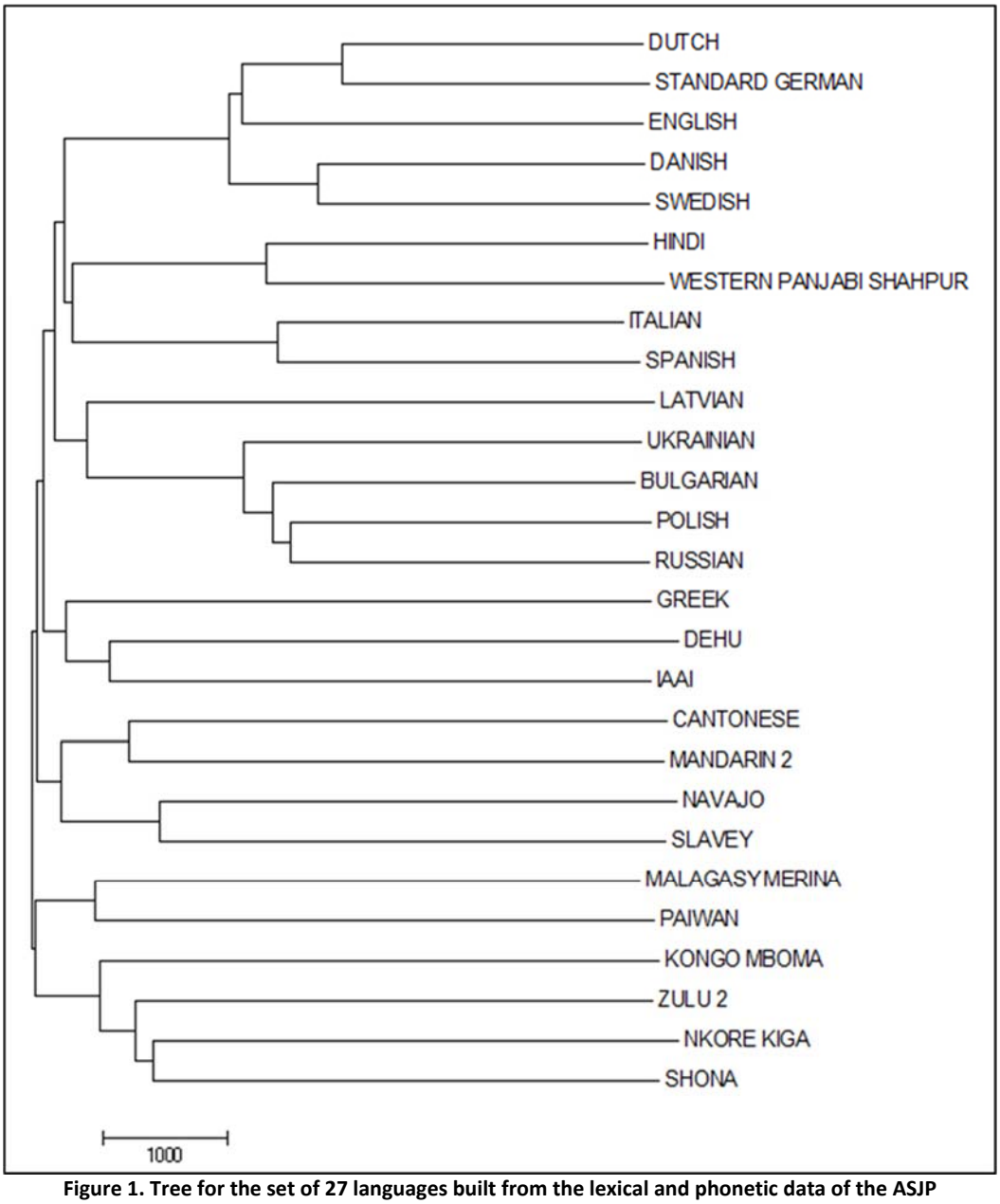

Figure 1. Tree for the set of 27 languages built from the lexical and phonetic data of the ASJP

4. In the ASJP tree, the closest relative of Shona is Nkore Kiga, and in the WALS tree, the closest relative of Shona is Zulu. In the following we compare the features of the trees in Figure 1 and 2 that were just highlighted with trees based on more mainstream methods of comparative linguistics.

Situation 1. According to Seebold (2006), Proto-Germanic divided into three branches: Northern, Eastern, and Western. The Western branch includes Dutch, German, and English. The ASJP tree, unlike the WALS tree, accurately represents this situation.

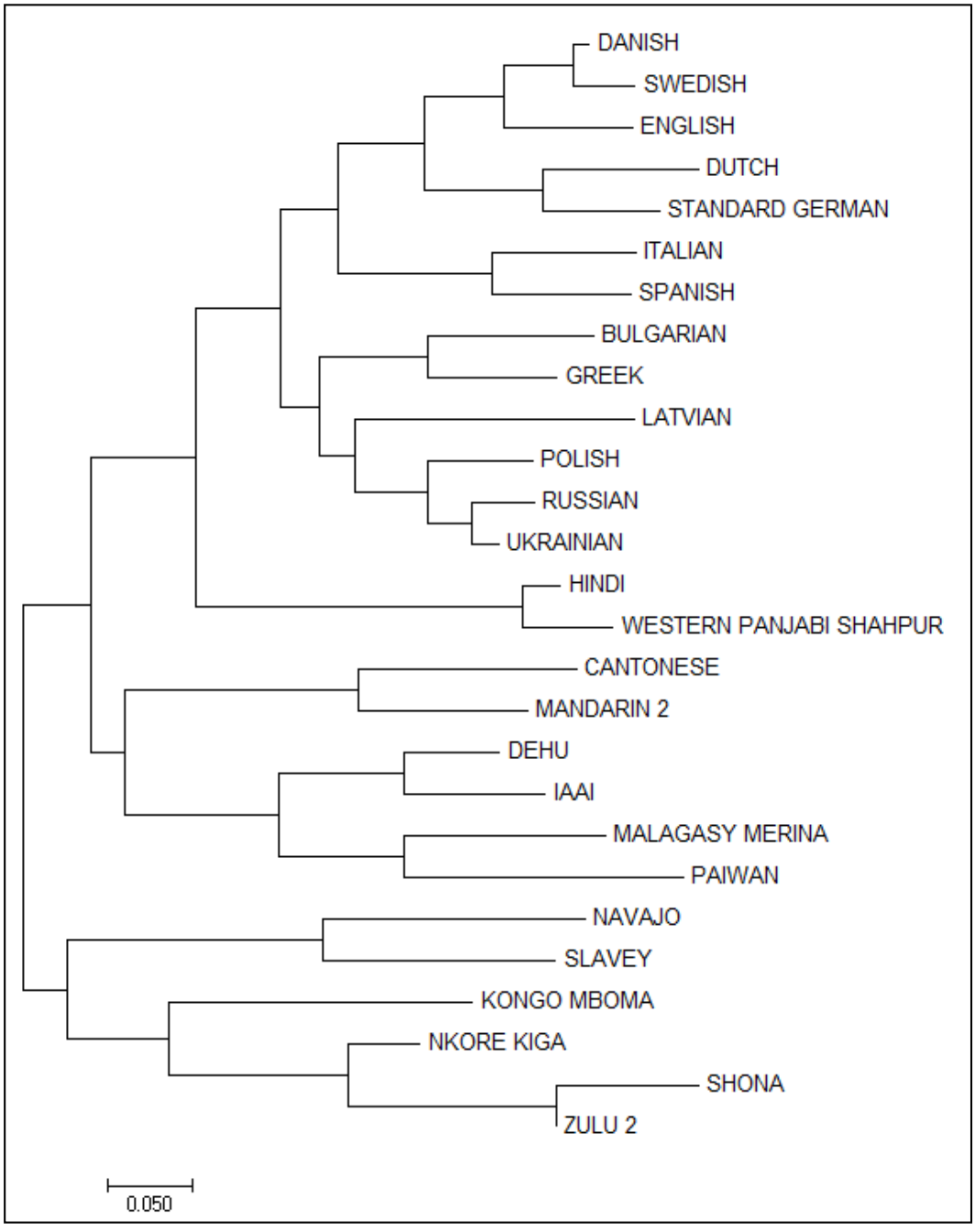

Figure 2. Tree for the set of 27 languages built on grammatical data from WALS

Situation 2. Bulgarian is the descendent of Old Bulgarian which, in turn, is closely related to the South Slavic dialects that formed the base of Old Church Slavonic (Maslov 2005), while Modern Greek has no close relatives. The ASJP tree reflects this situation. According to the WALS tree, Modern Greek and Bulgarian are close relatives.

Situation 3. According to Chang et al. (2015), the Indo-Iranian languages split off from the Indo-European lineage earlier (in the 3rd century BC, according to Oransky 1979) than the Slavic and Germanic languages. This situation is closer to the scenario described by the WALS tree, while the ASJP tree suggests a different sequence.

Situation 4. The classification of the African languages is not yet complete, so it would be premature to make any statements at this point. Thus, the tree built from the ASJP data is closer to the traditional views of comparative linguistics. The tree built from the WALS data brings up problems that require explanation.

For example, why is Danish the closest relative of English according to the WALS tree? It must have borrowed some elements of grammar of a language close to Danish, and more so than of Latin or Celtic grammar, although it is widely known that during the Roman invasion in the 1st to 5th centuries AD the native languages were Celtic (Brittonic) and not Germanic (Schrijver 2013). This may be explained by the fact that in the 9th century, part of England became a territory where laws of the Danes held sway (Hornung 2017). The lexical borrowings from Latin came to English mainly from Medieval French (after the Norman conquest in the 12th century, cf. Lutz 2017).

Another example is Greece and Bulgaria. ASJP classifies Modern Greek in a branch coordinate with the one uniting the other Indo-European languages. In contrast with the ASJP tree, the WALS tree classifies Modern Greek with the Balto-Slavic languages and singles out Bulgarian as its closest relative. Evidently this reflects the fact that Greek and Bulgarian are both participants in the Balkan linguistic area.

While it would take up too much space to discuss each individual WALS feature whose value has contributed to the particular shape of the WALS tree, we may at least consider whether the differential stabilities of WALS features, using the measurements of (Wichmann and Holman 2009, Table 1), have a role to play. In Table 2 we pick the two cases of contact influence, English-Danish and Bulgarian-Greek, and compare the language likely to have been most affected by the contact to its closest relative in the ASJP tree. We supply information on the number of features that match and the stabilities of features with respectively matching and non-matching values.

A hypothesis might be that features whose values are shared between languages that have been in contact will tend to be more unstable than features whose values are shared between languages that are related and not in contact. This hypothesis, however, rests on the assumptions that (1) diffusion of features is a main contributor to their instability and that (2) there is something inherent to different features that make them more or less diffusible. What Table 2 reveals is that the stability of features whose values agree are about the same for English-Dutch vs. English-Danish and Bulgarian-Russian vs. Bulgarian-Greek. The explanation for this–if it is something that needs an explanation–is that the two assumptions just mentioned are probably wrong. As regards (1), it may well be that a much stronger

force than diffusibility accounts for differential stabilities. In the summary of their results, Wichmann & Holman (2009: 51–52) suggest that a single, major driving force behind stability is the importance of a feature in the core morphosyntactic organization of a language. As for assumption (2)–the idea that some features are more diffusible than others–this was tested directly by Wichmann & Holman (2009: 19–20), who found that different features diffuse in different areas in patterns that cannot be predicted. So all in all, we should not be surprised by the numbers we encounter in the column showing mean stability of features with agreeing values in Table 2. Another hypothesis about what Table 2 should be telling us, which is a better one but perhaps also a less interesting one in the present context, is that we would expect features whose values are shared between related languages to more stable than features for which related languages disagree. This is borne out in 3 out of 4 cases in Table 2 (cf. the last two columns).

Table 2. **Statistics on the sharing of feature values and the mean stabilities of those values for 259 language pairs of special interest**

	Matching values	Number of overlapping features	Matching / overlapping	Mean stability of features with matching values	Mean stability of features with non-matching values
English-Dutch	48	60	0.80	37.46	37.93
English-Danish	35	41	0.85	38.41	30.83
Bulgarian-Russian	49	66	0.74	40.05	35.28
Bulgarian-Greek	55	66	0.83	39.42	33.55

Turning now to the third situation—the order in which the languages separated from the Indo-European tree—we note that in the results of Chang et al. (2015) the sequence in which the major subgroups included in our data split from the ancestral lineage is the same as in the WALS tree, namely first Indic and then Slavic and Germanic. In the ASJP tree this order is shuffled such that the split-off of Indic occurs later than that of Slavic.

At this point, we can draw some preliminary conclusions. ASJP allows for building trees based on lexical and phonetic data at relatively shallow time depths with a quality that is not very different from the quality of manually built phylogenetic trees. In contrast, it seems that WALS, which is based on grammatical data, allows for the correct identification of relationships between languages at a further distances, although we provide more discussion of this issue in the following subsection. At the same time, WALS data aids in the detection of diffusion of grammatical structures between languages that are otherwise not close in terms of lexicon or phonology.

2.2. Discussion

When comparing the trees built on the lexical and phonetic data (ASJP) and structural data (WALS) we have seen that the ASJP tree and the WALS tree, in general, represent the relationship of languages quite satisfactorily. Nevertheless,

the trees have some important differences. For example, the ASJP tree represents the relationships of languages in a way more similar to the trees built manually from the data from comparative linguistics.

This may be due to the fact that in both cases lexical and phonetic data play a major role. The WALS tree has two types of differences (A and B) from the ASJP tree:

A) The WALS tree shows very close relationships between some languages that are not close relatives from the point of view of historical linguistics (e.g., Danish & English and Greek & Bulgarian). As discussed in the previous subsection there is reason to believe that these are cases of convergence. Thus, the WALS tree seems to contain additional information on some cases of language convergence.

B) The WALS tree seems to render the distant relationship of languages with higher quality than the ASJP trees.

Concerning difference A), we can consider the following hypotheses: (1) The languages originated from a common proto-language, but the common vocabulary was mostly washed out from them (divergence). (2) Speakers of the languages used to be in close contact, and, as a result, borrowed grammatical structures (convergence). For instance, this linguistic situation occurred when English underwent substratal effects from Norse during the times of the Danelaw. (3) Target and structurally maximally close languages coincidentally evolved in the same way so that they became structurally closest in terms of the features we look at here.

The first hypothesis turns out to be inconsistent for the following reasons. English has many borrowings from a number of languages. For example, it has borrowings from Latin that came from Medieval French (7th century AD), which are administrative or professional terms (Mattila 2006). Danish contact with English took place in the 8th to 10th centuries AD.

As a result, English has some Scandinavian borrowings as well (Baugh & Cable 2002). Modern English has words both of Danish and of Latin origin, but the latter, especially, do not belong to the basic vocabulary (Baugh & Cable 2002). Emonds and Faarlund (2014) suggested that English can be classified as a Scandinavian languages. This idea goes against widely held opinions and has been strongly criticized (Bech & Walkden 2016).

Consequently, the hypothesis about close contacts resulting in the borrowing of elements of grammatical structure (from Danish into English and vice versa) seems more logical. The nature and the intensity of these contacts are beyond the scope of the present research.

The language pair Greek-Bulgarian was probably in a similar situation. As a result of long-lasting linguistic contact in the Balkans, Bulgarian, perhaps mediated by other languages, borrowed some features of the grammatical structure of (earlier stages of) Greek (or vice versa) without borrowing any basic vocabulary.

Concerning the third suggestion of accidental evolution, the probability of many features independently converging on acquiring the same values is very low—it is something which is more likely to take place when the data is sparse

(Polyakov et al. 2016). Figure 3 shows a tree that combines information on the divergence (ASJP tree) and convergence (WALS tree) of languages. It reproduces the ASJP trees, but dotted lines are added to connect languages that have suffered convergence according to the WALS tree.

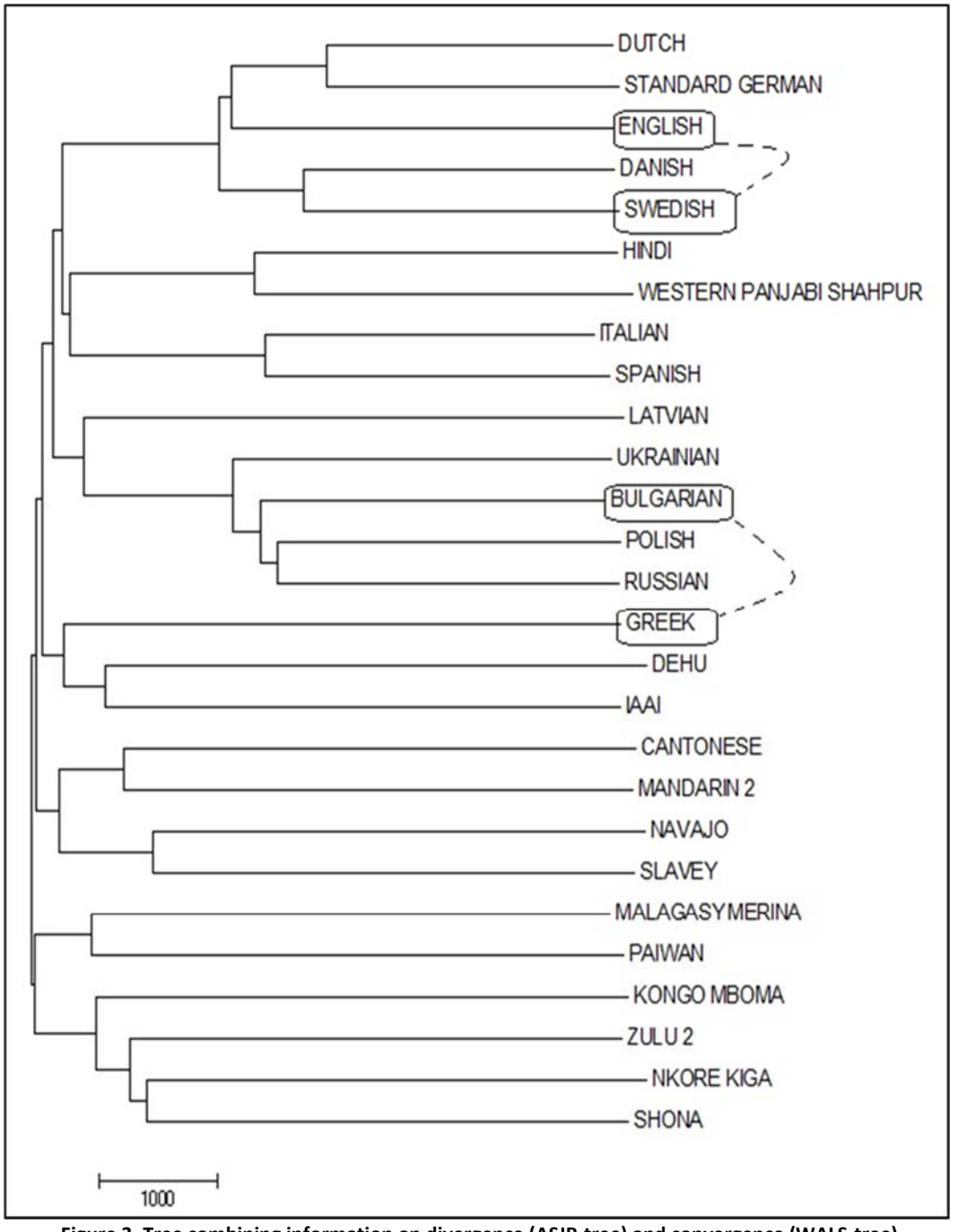

Figure 3. Tree combining information on divergence (ASJP-tree) and convergence (WALS-tree) of languages

As for situation B, it may well be that the areal input coming from typological features has contributed to the clustering of European languages separately from the Indic ones.

Conclusion

In this paper we have presented a comparative analysis of two trees. The first tree was built on lexical and phonetic data from the ASJP database, and the second tree was built on typological data of the WALS Program database, after filtering out non-structural features. Based on a previous study (Polyakov et al. 2016), we selected 27 languages from the WALS data, which formed a set for the present research.

The results show that both trees to some degree satisfactorily describe the relationship among the languages in the sample. Nevertheless, there are four major differences between the two trees. These differences were analyzed with reference to phylogenetic trees built manually using the comparative method. The analysis showed that the ASJP tree, in general, is better at presenting the information about relationships of languages at relatively shallow temporal distances. The WALS tree seems to better presents the information about relationships of languages at further distances, but it is still not clear whether the grouping of European languages as a clade separate from the Indic languages is due to a phylogenetic signal or an areal one. It is, at any rate, clear that certain aspects in which the WALS tree differs from traditional views of comparative linguistics can be explained as representing grammatical diffusion, as in the cases of English-Danish and Bulgarian-Greek. We suggested that different socio-linguistic situations can explain these phenomena. In the case of English and Danish we propose that the reason for large-scale borrowing was the existence of the Danelaw. In the case of Bulgarian and Greek, the reason was suggested to have been participation in the Balkan contact area.

Furthermore, in the present paper, we suggested a new type of tree representation that could unite information on divergence (i.e., information from lexical and phonetic data) and convergence (i.e., information from structural data) of languages.

The use of structural data for building trees for phylogenetic purposes has a number of advantages compared to lexical and phonetic data (ASJP). First, it seems to provide a more reliable view of a more distant historical period, although this matter requires further investigation: a family tree will often reflect geography because of historical migrations, and it may be the input of geographical information coming from typological features which helps to identify nodes that are also geographically defined, such as Indic vs. European languages. Second, despite being prone to borrowing, grammatical data may be preserved in the structure of a language for a longer time, even when all inherited lexical and phonetic elements have been washed out.

Conflict of interest information

There is no conflict of interest.

Acknowledgments

This work was supported by the Russian Foundation for Basic Research, grant 16-06-00187a and by the Kazan Federal University Strategic Academic Leadership Program.

REFERENCES

Anisimov, Ivan, Vladimir N. Polyakov & Valery D. Solovyev. 2013. Database "Languages of the World." New Version. New Research Horizons. In Svetlana Masalóva & Valery Solovyev (eds.), *Proceedings of the First International Forum on Cognitive Modeling 14–21 September, 2013, Italy, Milano Marittima. Part 1. Cognitive modeling in linguistics: Proceedings of the XIV International Conference "Cognitive Modeling in Linguistics. CML-2013,"* 27–34. Rostov-on-Don: Southern Federal University Press.

Barbançon, François et al. 2013. An experimental study comparing linguistic phylogenetic reconstruction methods. *Diachronica* 30(2). 143–170. https://doi.org/10.1075/Dia.30.2.01bar

Bech, Kristin & George Walkden. 2016. English is (still) a West Germanic language. *Nordic Journal of Linguistics* 39(1). 65–100. https://doi.org/10.1017/S0332586515000219

Benveniste, Emile. 1954. La classification des langues. *Conferences de l'Institut de Linguistique de l'Universite de Paris* 11. 33–50.

Birchall, Joshua, Michael Dunn & Simon J. Greenhill. 2016. A combined comparative and phylogenetic analysis of the Chapacuran language family. *International Journal of American Linguistics* 82(3). 255–284.

Bopp, Franz. 1885. *A Comparative Grammar of the Sanskrit, Zend, Greek, Latin, Lithuanian, Gothic, German, and Slavonic Languages*; *Volume 1*. London: Williams.

Brown, Cecil H. et al. 2008. Automated classification of the world's languages: A description of the method and preliminary results. *STUF – Language Typology and Universals* 61(4). 285–308.

Burlak, Svetlana & Sergei Starostin. 2005. *Sravnitel'no-istoricheskoe Yazykoznanie* [Comparative Linguistics]. Moscow: Publishing center "Academia."

Chang, Will et al. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(2). 194–244.

Coloma, Germán. 2017. Complexity trade-offs in the 100-language WALS sample. *Language Sciences* 59. 148–158. https://doi.org/10.1016/j.langsci.201 6.10.006

Donohue, Mark & Simon Musgrave. 2007. Typology and the linguistic macro-history of island Melanesia. *Oceanic Linguistics* 46. 348–387.

Donohue, Mark et al. 2011. Typological feature analysis models linguistic geography. Language 87(2). 369-383.

Donohue, Mark et al. 2008. Typology, areality and diffusion. *Oceanic Linguistics* 47(1). 223–232.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *The World Atlas of Language Structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info (accessed 20 January 2018).

Dunn, Michael. 2015. Language phylogenies. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge handbook of historical linguistics*, 190–211. New York: Routledge.

Dunn, Michael et al. 2007. Statistical reasoning in the evaluation of typological diversity in Island Melanesia. *Oceanic Linguistics* 46. 388–403.

Dunn, Michael et al. 2008. Structural phylogeny in historical linguistics: Methodological explorations applied in Island Melanesia. *Language* 84(4). 710–759. https://doi.org/10.1353/lan.0.0069

Dunn, Michael et al. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309. 2072–2075.

Edwards, Anthony, William Fairbank & Luigi Luca Cavalli-Sforza. 1964. Reconstruction of evolutionary trees. *Phonetic and Phylogenetic Classification. Systematics Association Publication* 6. 67–76. http://www.faculty.biol.ttu.edu/Strauss/Phylogenetics/Readings/EdwardsCavalliSforza1964.pdf (accessed 21 February 2018).

Emonds, Joseph Embley & Jan Terje Faarlund. 2014. *English: The language of the Vikings.* Olomouc modern language monographs 3. Olomouc: Palacký University.

Felsenstein, Joseph. 2003. *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates.

Gray, Russell D. & Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European Origin. *Nature* 426(6965). 435–439. https://doi.org/10.1038/nature02029

Gray, Russell D. & Fiona M. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 405(6790). 1052–1055.

Gray, Russell D., David Bryant & Simon J. Greenhill. 2010. On the shape and fabric of human history. *Philosophical Transactions of the Royal Society* B 365. 3923–3933.

Hamming, Richard W. 1950. Error detecting and error correcting codes. *Bell System Technical Journal* 29(2). 147–160. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x

Haspelmath, Martin et al. 2005. *The World Atlas of language structures. 1 st ed*. Oxford: Oxford University Press.

Holman, Eric W. & Søren Wichmann. 2017. New evidence from linguistic phylogenetics supports phyletic gradualism. *Systematic Biology* 66(4). 604–610. https://doi.org/10.1093/sysbio/syw106

Holman, Eric W. et al. 2008. Explorations in automated language classification. *Folia Linguistica* 42(2). 331–354.

Hornung, Annette. 2017. *English: The Grammar of the Danelaw.* Arizona State University, ProQuest Dissertations Publishing.

Kolipakam, Vishnupriya et al. 2018. A Bayesian phylogenetic study of the Dravidian language family. *Royal Society Open Science* 5(3). https://doi.org/10.1098/rsos.171504

Kumar, Sudhir, Glen Stecher & Koichiro Tamura. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* 33(7). 1870–1874. https://doi.org/10.1093/molbev/msw054

Lewis, Paul M. (ed.). 2009. *Ethnologue: Languages of the World* (Sixteenth edition). Dallas, Texas: SIL International. http://www.ethnologue.com/ (accessed 20 January 2018).

Longobardi, Giuseppe et al. 2015. Toward a syntactic phylogeny of modern Indo-European languages. In Leonid Kulikov & Nikolaos Lavidas (eds.), *Proto-Indo-European syntax and its development,* 125–156. Amsterdam: John Benjamins Publishing Company.

Lutz, Angelika. Norse loans in Middle English and their influence on Late Medieval London English. *Anglia* 135(2). 317–357. https://doi.org/10.1515/ang-2017-0028

Maslov, Yu. 2005. Bolgarsky [Bulgarian]. In *Languages of the World. Slavic languages.* 69–102. Moscow: Academia.

Mattila, Heikki E. S. 2006. Legal language: History. In Keith Brown (ed.), *Encyclopedia of language and linguistics* (2nd ed.), 8–13. London: Elsevier. https://doi.org/10.1016/B0-08-044854-2/04504-1

Müller, André et al. 2013. ASJP World Language Tree of Lexical Similarity: Version 4 (October 2013). http://asjp.clld.org/

Nakleh, Luay, Don Ringe & Tandy Warnow. 2005. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* 81(2). 382–420.

Nichols, Johanna & Tandy Warnow. 2008. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass* 5(2). 760–820.

Oransky, Iosif Mikhailovich. 1979. *Iranian Languages in Historical Perspective*. Moscow: Nauka. http://www.orientalstudies.ru/rus/index.php?option=com_publications&Itemid=75&pub=619 (accessed 20 March 2021).

Pagel, M. & Meade, A. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *American Naturalist* 167. 808–825.

Polyakov, Vladimir N., Ivan S. Anisimov & Elena A. Makarova. 2016. Can grammar define similarity of human natural languages? *American Journal of Applied Sciences* 13(10). 1040–1052. https://doi.org/10.3844/ajassp.2016.1040.1052

Polyakov, Vladimir N. et al. 2009. Using WALS and languages of the world. *Linguistic Typology* 13(1). 137–167. https://doi.org/10.1515/LITY.2009.008

Pompei, Simone, Vittorio Loreto & Francesca Tria. 2011. On the accuracy of language trees. *PLoS ONE* 6(6). e20109. https://doi.org/10.1371/journal.pone.0020109

Rama, Taraka & Lars Borin. 2015. Comparative evaluation of string similarity measures for automatic language classification. Sequences in language and text. http://spraakdata.gu.se/taraka/string-similarities-pdf2doc.pdf (accessed 13 April 2021).

Ringe, Don, Tandy Warnow & Ann Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society* 100(1). 59–129.

Schrijver, Peter. 2014. *Language Contact and the Origins of the Germanic Languages.* New York: Routledge.

Seebold, Elmar S. 2006. Westgermanische Sprachen [West Germanic Languages], *Reallexikon der germanischen Altertumskunde* 33. 530–536.

Solovyev, Valery. 2009. Is grammochronology possible? *Proceedings of the Swadesh Centenary Conference, 17–18 January 2009.* Munich: Official Website of Max Planck Institute for Evolutionary Anthropology. http://www.eva.mpg.de/lingua/conference/09_SwadeshCentenary/pdf/abstracts/Valery_Solovyev.pdf (accessed 15 April 2021).

Swadesh, Morris. 1950. Salish internal relationships. Inte*rnational Journal of American Linguistics* 16. 157–167.

Swadesh, Morris. 1952. Lexicostatistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96. 452–463.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21. 121–137. https://lib.ugent.be/catalog/rug01:002194436 (accessed 5 May 2021).

Trubetzkoy, Nikolai S. 1939. *Gedanken über das Indogermanenproblem* [Commemoration of the Indo-Germanic problem]. *Acta Linguistica. Copenhague* 1(1). 81–89. https://doi.org/10.1080/03740463.1939.10410851

Wichmann, Søren. 2013. A classification of Papuan languages. In: Hammarström, Harald and Wilco van den Heuvel (eds.), *History, contact and classification of Papuan languages* (*Language and Linguistics in Melanesia*, *Special Issue 2012*), 313–386. Port Moresby: Linguistic Society of Papua New Guinea.

Wichmann, Søren. 2017a. Genealogical classification in historical linguistics. In Mark Aronoff (ed.), *Oxford research encyclopedias: Linguistics*. Oxford: Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.78

Wichmann, Søren. 2017b. Modeling language family expansions. *Diachronica* 34(1). 79–101. https://doi.org/10.1075/dia.34.1.03wic

Wichmann, Søren & Eric W. Holman. 2010b. Pairwise comparisons of typological profiles. In Jan Wohlgemuth & Michael Cysouw (eds.), *Rethinking universals: How rarities affect linguistic theory,* 241–254. Berlin/New York: Walter de Gruyter Publishers. https://doi.org/10.1.1.558.3743&rep=rep1&type=pdf

Wichmann, Søren & Jeff Good. 2014. Introduction. In Søren Wichmann & Jeff Good (eds.), *Quantifying language dynamics: On the cutting edge of areal and phylogenetic linguistics,* 1–6. Leiden: Brill.

Wichmann, Søren & Eric W. Holman. 2009. *Temporal Stability of Linguistic Typological Features*. München: LINCOM Europa.

Wichmann, Søren et al. 2010b. Evaluating linguistic distance measures. *Physica A* 389. 3632–3639. https://doi.org/10.1016/j.physa. 2010.05.011

Wichmann, Søren, Eric W. Holman & Cecil H. Brown (eds.). 2016. *The ASJP Database* (version 17). Available at: http://asjp.clld.org/ (Accessed 4 February 2018).

Wichmann, Søren, André Müller & Viveka Velupillai. 2010a. Homelands of the world's language families: A quantitative approach. *Diachronica* 27(2). 247–276.

Wichmann, Søren & Taraka Rama. 2018. Jackknifing the black sheep: ASJP classification performance and Austronesian. In Kikusawa, Ritsuko and Lawrence A. Reid (eds.), *Let's talk about Trees: Genetic relationships of languages and their phylogenic representation*, 39–58. Senri Ethnological Studies 98. Osaka: National Museum of Ethnology, Japan.

Wichmann, Søren & Arpiar Saunders. 2007. How to use typological databases in historical linguistic research. *Diachronica* 24.2. 373–404.

Wong, Kok-Seng & Myung Ho Kim. 2014. On private Hamming distance computation. *The Journal of Supercomputing* 69(3). 1123–1138. https://doi.org/10.1007/s11227-013-1063-z

Article history:
Received: 06 May 2021
Accepted: 14 December 2021

Bionotes:

Valery D. SOLOVYEV is Doctor Habil. of Physics and mathematics, Professor, Chief Researcher of the "Text analytics" Research Laboratory at Kazan (Volga Region) Federal University. His research interests include cognitive sciences, computer linguistics, text complexity.

Contact information:
Kazan (Volga Region) Federal University, 18 Kremlevskaya St., 420008, Kazan, Russia
e-mail: maki.solovyev@mail.ru

Elena A. MAKAROVA is junior researcher, Section of Applied Linguistics, Her research interests embrace typology, typological databases, isolate languages. Research interests: Typology, typological databases, isolate languages.

Contact information:
Institute of Linguistics of the Russian Academy of Sciences
1 bld. 1 Bolshoy Kislovsky Lane, Moscow, 125009, Russia
e-mail: MakarovaEA@iling-ran.ru

Сведения об авторах:
Валерий Дмитриевич СОЛОВЬЕВ – доктор физико-математических наук, профессор, главный научный сотрудник НИЛ «Текстовая аналитика» Казанского (Приволжского) федерального университета. В сферу его научных интересов входят когнитивная наука, компьютерная лингвистика, сложность текстов.
Контактная информация:
Казанский (Приволжский) федеральный университет
Россия, 420008, г. Казань, ул. Кремлевская, 18
e-mail: maki.solovyev@mail.ru

Елена Андреевна МАКАРОВА – младший научный сотрудник сектора прикладного языкознания Института языкознания РАН. В сферу ее научных интересов входят типология, типологические базы данных, языки-изоляты.
Контактная информация:
Институт языкознания РАН
Россия, Москва, 125009, Б. Кисловский пер. д. 1 стр. 1
e-mail: MakarovaEA@iling-ran.ru