

Using WALS and Jazyki Mira

Vladimir N. Polyakov

Institute of Linguistics of the Russian Academy of Sciences

Valery D. Solovyev

Kazan State University

Søren Wichmann

Max Planck Institute for Evolutionary Anthropology & Leiden University

Oleg Belyaev

Moscow State University

Abstract

The paper's primary concern is to address the usage of WALS through comparing it with another typological database of similar scope, *Jazyki Mira*. Such a comparison is carried out based on a set of criteria. In Section 2, the scope of the databases is compared, as well as their differences and similarities in structure, the number of errors, and in the existing user interfaces. In Section 3, calculations of typological similarity and temporal stability of language features based on the data provided by both databases are compared. Finally, conclusions are drawn as to the relative efficiency and usefulness of these databases for different aims of research or educational goals.

Keywords

WALS, *Jazyki Mira*, typology, quantitative linguistics, computational linguistics, historical linguistics, classification

0. Introduction

This paper addresses the issue of using WALS through a comparison with another typological database, *Jazyki Mira* ('Languages of the World', henceforth JM). Through this comparison with another typological database advantages and deficiencies of WALS are brought to light. More space will be devoted to describing the structure and contents of JM than to describing WALS, since the latter is much more widely known. After this description we go on to compare the usefulness of the two. It will become clear that the two databases are in several ways complementary, and also that they have a relatively low degree of compatibility, although there certainly are many features in which both databases intersect. In any case, having two databases, even if they are different, is useful in its own way, since it allows us to better verify results obtained by applying various methods to them.

The respective databases have been developed independently of one another over several years and represent the two most ambitious attempts to date to make typological information available for systematic analysis. The existence of these two databases represents an excellent opportunity to consider different choices that have been made or could be made in future approaches to structuring large amounts of typological information.

1. History of JM

In the 1980s a large-scale project was initiated by scholars in Moscow to produce a series of typological sketches of individual languages which were to follow a consistent outline and may be seen as an independent Russian counterpart to the *Lingua Descriptive Studies* (cf. Comrie and Smith 1977). These sketches were eventually published in the series called *Jazyki Mira*¹, but it took many years from the inception of the project until the first volume appeared in 1993². Long before these volumes became finalized, however, efforts towards an encoding of the information in a database format were undertaken. This work was initiated in the 1980s by member-correspondent of the Russian Academy of Sciences Victoria Yartseva and was carried out in the sector of Applied Linguistics of the Academy under the leadership of A. I. Novikov. The concept behind this first version of the database, whose structure is carried over to later versions, is described in Novikov and Yaroslavtseva (1985).

In 2002 the computational aspect of the project took a new turn with the creation of a Windows version under the direction of Vladimir N. Polyakov (programmed by V. Logunov), and subsequently (2007) a web-version became available (programmed by P. D. Khanukaev). This may be accessed through the website <http://dblang2008.narod.ru>³, which also contains some information related to the database.

2. Comparing contents and structure of JM with WALS

¹ Cf. *Jazyki Mira* (1993-2004), Moldovan et al. (2005), Toporov et al. (2006).

² Early ideas on the encyclopædic description of languages which influenced the later work can be found in the following monographs: Yartseva and Besserebrennikov (1976), Yartseva (1980, 1982-1, 1982-2), Zhurinskaya et al. (1986).

³ To access the database, type in the username “demo” and password “demo”.

2.1. Languages represented

JM contains data for 315 languages, mostly from Eurasia, and WALS has data from 2,560 languages from all over the world (including some sign languages). The JM languages belong to the following families (giving numbers of languages for each in parenthesis): Afro-Asiatic (9), Altaic (74), Austro-Asiatic (2), Chukotko-Kamchatkan (4), Eskimo-Aleut (9), Hurrian-Urartian (2), Indo-European (143), Kartvelian (4), Nakh-Daghestanian (28), Northwest Caucasian (5), Sino-Tibetan (2), Uralic (22), Yeniseian (3). In addition, the following isolates are represented: Ainu, Burushashki, Elamite, Japanese, Nivkh, Sumerian, Yukhagir. Finally, a single Austronesian language (Rotuman) is represented. A unique feature of JM is that it contains not only currently spoken languages, but also 54 extinct ones⁴, which can prove useful when analyzing diachronical processes.

2.2. Representation of data

WALS and JM differ fundamentally in that the former represents all features in a flat structure, whereas the latter is hierarchically organized. The only hierarchical structure in WALS is the grouping of features into 11 categories (including an ‘other’ category): phonology, morphology, nominal categories, etc. But this arrangement only serves the

⁴ The extinct languages included in the database are the following: Avestan, Bactrian, Bojnurdi, Parthian, Old Persian, Middle Persian, Sargulyam (a Pamirean language), Scythian, Sogdian, Khotanese, Vanji (Iranian); Celtiberian, Cornish, Gaulish, Lepontic, Pictish (Celtic); Gothic (Germanic); Latin, Oscan, Umbrian, Faliscan (Italic); Old French (Romance); Polabian, Old Church Slavonic (Slavic); Hittite (Anatolian); Bolgar, Chagatai, Cuman, Hunnic, Khwarezmian, Old Oghuz, X-XI century Oghuz, Middle Oghuz, language of the Orkhon-Yenisey script, Mamluk Kipchak, Pecheneg, Old Anatolian Turkish, Turki (Central Asian Turkic), Old Uyghur, Karakhanid Uyghur (Turkic); Akkadian, Biblical Hebrew (Semitic); Kamassian, Mator (Uralic); Kott, Yugh (Yeniseian); Manchu (Tungusic); Moghol, Classical Mongolian (Mongolic); Old Chinese (Sino-Tibetan); Old Japanese (Japonic); Urartian, Hurrian (Hurro-Urartian); Sumerian, Elamite (isolates). Of these 55 languages, only 53 (excluding Parthian, the description of which is not finished yet, and Pictish, which does not contain enough known data) can be used for typological calculations. It should also be noted that many of these languages do not have a generally accepted nomenclature or are only described in the Russian sources of the JM encyclopædia.

purpose of searches in the electronic version. Essentially all features are treated as independent. Moreover, the features take from two to nine values that are juxtaposed rather than organized into some hierarchy. In contrast, JM is hierarchically organized. At the highest level the data are organized into two major parts: metadata (part 1) and data (part 2). The metadata, which are not (yet) included in the online electronic version, consist of information about the language name, variants of the name, genetic affiliation, location, dialectal variants, sociopolitical status, literacy, status in the educational system, associated writing system, a historical periodization, and types of changes undergone as a result of contact. The data (part 2) are headed by the 20 top-level, juxtaposed sections shown in table 1. Although the numbering system suggests a grouping of these 20 sections into 5 (unnamed) higher categories, these higher categories do not determine the organization of the data in terms of how features are encoded. The numbering and structure of the high-level nodes is entirely determined by the structure used in the original JM encyclopædia.

[INSERT TABLE 1 HERE]

Within the 20 highest-level categories shown all features are expressed as hierarchical, binarily organized features. To illustrate how this works table 2 shows the features within the last section, on complex sentences. It gives both the original Russian short descriptions and provisional English translations thereof. Among other things, this table can serve as a good example of how difficult translating feature descriptions may be, which is in part due to the peculiarities of the Russian typological school. We have run

into numerous problems even in this small excerpt; the contents of other categories can be even more complicated.

[INSERT TABLE 2 HERE]

The dots preceding each description indicates the degree of hierarchical embedding. For instance, feature number 3818 is embedded at the third sublevel, as follows

COMPLEX SENTENCE

1st sublevel: type of connection

2nd sublevel: conjunctions

3rd sublevel: absence of conjunctions as a grammatical category

The information is encoded as a simple table where a ‘true’ (present) vs. ‘false’ (absent) value is given for each number in the database.

2.3. Amount of data

The number of datapoints for spoken languages in WALS is 57,916. The number of values in WALS varies from 2 to 9. If the number of possible values is taken into account when calculating the total number of datapoints—e.g., if a 6-value feature attested for a given language counts as 6 datapoints—then the total number of datapoints in WALS is 274,461. JM includes 3801 binary features and 315 languages, so there is a theoretical maximal number of datapoints of $3801 * 315 = 1,197,315$. For some features, however, data were not available. In the online database the combination of the symbol “.O” and a

checkmark following a section heading (e.g., see 2.1.1. ФОНЕМНЫЙ СОСТАВ ‘phoneme inventory’ for Pictish) indicates that data were not available for the given set of features. The frequencies of such ‘missing sections’ and the number of features they contain add up to 83,164 datapoints. The total number of data points, then is 1,114,151. This is almost exactly four times as many as in WALS. Nevertheless one should also take into account the hierarchical structure. If only terminal features are counted as data points there are obviously fewer. The number of terminal features is 3340, of which only 3198 are actually present in any of the languages (which is due to the fact that the database has been designed with other, non-Eurasian language families in mind), which gives a theoretical maximum of 1,007,370 (1,052,100 when considering all of the features).

While the amount of data for individual languages in WALS is extremely skewed and little more than 100 languages are attested for all or nearly all features JM is much more balanced in this respect. Its data matrix of languages and features only has 6.9% empty cells representing gaps in knowledge, whereas the corresponding percentage for WALS is 83.8%.

2.4. Focus of sampling

The data for WALS were assembled for the purpose of mapping the distributions of different typological features, not for the purpose of documenting languages. JM, however, focuses on the distribution of features both among and within languages. All major differences are consequences of this difference in focus. It explains why WALS has a greater number of languages represented—it has simply not been possible to sample such a large number of JM with the high ambitions for completeness of description. It also explains the strength of JM over WALS that the data can be better for inferring

genealogical relations and the strengths of WALS over JM that the former is better for inferring stabilities of individual features and for discovering worldwide distributional patterns in typological features. There is nothing in the design of JM that limits its superiority in these respects, but the time which would be required to fill in information for all the various languages is a factor which imposes practical limits.

2.5. Data sources

The data contained in JM are mostly drawn from the issues of the *Języki Mira* encyclopædia, continuously published by the Institute of Linguistics of the Russian Academy of Sciences from 1993 up to this day. The structure of JM's feature set also mimics the encyclopædia, which itself follows a more-or-less rigorous and unified specification for its articles. The reliance on published materials by specialists in the languages described is both JM's strength and its disadvantage. Since in most cases it is clear where the information was taken from, it is possible both to correct mistakes and to find out whether they stem from a problem in the original article or in the way it was transcribed into the database. On the other hand, excessive reliance on data that are more than 10 years old data can be a considerable hindrance to JM's further improvement and development. In any case, the fact that it is based on a published encyclopædia is one of JM's unique features, and in many ways determines other design decisions made while drafting its main structure. In spite of this, however, some languages present in the database are noticeably absent in the published descriptions. For instance for the two Sino-Tibetan languages described in the database there is not (yet) any corresponding volume in the encyclopædia. In such cases, the languages are mainly described on the basis of manuscripts for articles that are yet to be published.

The fact that JM is focused exclusively on Eurasia also means that it has a much more detailed selection than WALS of the languages from this region—especially as concerns those spoken in Russia and the former Soviet Union. This specifically applies to the Turkic and Iranian languages. JM contains 54 of the former and 42 of the latter, counting both languages and entities which are usually considered dialects. In contrast, WALS contains 41 Turkic and 26 Iranian languages. WALS, however, seems to cover Finnic (18 in WALS vs. 14 in JM) and Mongolic (13 vs. 11) languages a little bit better. But in general it would seem that JM covers the languages for which data are available in Russian sources only more thoroughly than WALS; this is the case for many Iranian languages (e.g., Pamir languages of Tajikistan and Afghanistan, and some other languages, like Davani and Char-Aimak, for which there are no entries in the Ethnologue⁵), and also for Altaic languages (Turkic, in particular), as well as Caucasian and Palaeoasiatic ones.

JM's feature set is heavily dependent on its data sources. In fact, the feature set, while being based on a template defined by the encyclopædia, is expected to grow as new languages are added, since the data are supposed to be entered by specialists who can use their own terms and concepts. This may in some cases create redundant or duplicate features, but given JM's focus on detailed and exhaustive language descriptions, this seems to be an unavoidable problem. Moreover, new features are not accompanied by detailed descriptions. This is a considerable drawback, since in many cases it is hard to understand what a short name for a feature precisely means. This problem is currently being addressed by analyzing the data contained in the source books and creating feature descriptions based on them.

⁵ These are sometimes, but not always, considered dialects of larger languages.

WALS differs from JM in that the features it contains do not yield a description of a given language as a whole, since the choice of features is rather arbitrary and certainly not aimed at being exhaustive. This difference is but one example of how the initial focus of the databases influences decisions regarding their respective designs.

2.6. Estimates of errors

For users of databases like WALS or JM it is important to know something about their reliability. There is only anecdotal information about the number of errors found in WALS. For users of the database a systematic examination of the information it contains for selected languages made by language experts would be useful.⁶ For JM, an expert examination of select languages from different families has been carried out as part of a project of the Russian Ministry of Education (Scientific and Educational Center of Linguistics, № 02.438.11.7015, director V. Solovyev). For Scandinavian languages, such an examination has been made by D. B. Nikulicheva, who investigated the error rates in the data for Danish, Norwegian, and Swedish. The results are shown in Table 3.

[INSERT TABLE 3 HERE]

The figures in Table 3 show that the number of errors in the printed version of JM seems to be lower than in the database, but these values cannot be compared directly. If we take into account the hierarchical structure of JM, and also the number of features it contains, the percentage of errors in both is probably close to being equal. The number of errors is

⁶ The online version, <http://www.wals.info>, is equipped with an errata page, <http://blog.wals.info/category/errata/>, which allows for users to note mistakes. Cf. also Cysouw (2007) for examples of contradictory statements across WALS chapters.

the number of features where 1 stands in place of 0 and vice versa. As mentioned above, the symbol “.O” followed by the value ‘true’ indicates lack of attestation, but only at the highest hierarchical level of whole sections. For the Scandinavian languages in question there are no such sections, as could be expected. There can be cases when, while a high-level section is not marked as unattested via the “.O” symbol, it still contains some features which are not attested for this language. In these circumstances the structure of the database does not allow for an explicit indication of the fact that the feature is unattested; thus, a zero value can in principle mean both “false” and “unattested”. This is, however, quite unlikely for languages like the ones in the example, since full information for them is available, and languages in JM are filled completely, for all the features. Zero values for features are no less important than the values indicating their presence. If all features are in fact attested the error rates are as indicated in the rightmost column in Table 3. If some are not attested it would mean a slightly higher error rate, perhaps in the order 2%-4%.

2.7. User interface

WALS and JM differ substantially in how their user interfaces are organized. While the data of both databases allow for any kind of application or representation, these interfaces are in many ways determined by the initial objectives of the scholars who designed the databases.

WALS was designed from the start with a specific set of features in mind, and adding new languages can be compared to filling in a questionnaire. It was devised first and foremost as an atlas, as a tool for observing various distributions of typological features. Thus, it is only natural that the user interface of the database works best when

used to getting from features to languages rather than vice versa. The map viewer is one of the most important features of WALS' interface, and it makes it a much better tool than JM for tracking areal groupings.

The idea behind the first versions of JM was to create a reference tool for the languages contained in the books. In other words, it was designed for the user to get information for *languages* and not for *features*, the features only being a way of organizing the descriptions. Other than the maps, however, there is no substantial difference in capabilities between the WALS and JM interfaces: both of them allow for going both from features to languages and vice versa.

To give an impression of how they are organized, screenshots of the JM and WALS user interfaces can be found in Figures 1-5. One notices that the features seem to be translated in the JM screenshots; the quality of this translation is not, however, satisfactory, and it has to be completely redone.

[INSERT FIGURES 1-5 HERE]

The differences which we pointed out, while not that critical, represent limits as to how a researcher can analyze the information contained in both databases. When a thorough description of a language is required JM is unsurpassed, but not because of the superiority of its interface in this regard, rather by the virtue of it having more detailed features. Where it falls short, however, is when one wants to see any kind of information pertaining to the geographical distribution of the features. For instance, if one wanted to discover where typological distributions form areal groupings, one can do so by just browsing the maps of WALS and taking notes. It would even be possible to examine all

the features present, since their number is perfectly processable by a human being. For JM, even if it did contain a mapping tool, analyzing all the features might not be viable. The greater detail with which languages are described is generally a positive trait and has advantages for many different purposes, but it can sometimes be a hindrance for an ordinary human user.

Both WALS and JM contain instruments for combining several features into one, but WALS allows only two to be combined, and JM just three (for WALS this applies only to full features and not values within one feature, where one can combine as many values as is desired). For more complex comparisons the researchers is forced to use additional software. Therefore there is in both cases a need for the further development of user interfaces, and this could follow similar directions.

2.8. Purposes as expressed through designs

WALS and JM also differ considerably in their possible applications.

JM was designed from the ground up with three areas of application in mind: as a reference tool; as an educational instrument; and, finally, as an object of quantitative research. The primary goal of WALS was to create an atlas and a tool for tracking geographical distributions of typological features and determining implicational universals, hence its clear superiority in these regards.

No matter what the databases were initially designed for, they can still be applied for roughly the same purposes, if only with different degrees of success. First of all, the information contained in the databases can be used for research in linguistic typology, both using traditional and quantitative methods. It can help us to find implicational language universals or near-universals (determining which features or feature

combinations are exceptionally frequent can also be an important result) or to analyze areal distributions for features. It can potentially allow us to determine which features are more likely to be borrowed and thus more ‘areal’ or, as a related exercise, to automatically find Sprachbunds or zones with especially high language contact. In these regards, as we have elaborated on in detail in the previous section, WALS is more useful. For finding relations between different features, however, the number of features in WALS is probably too low. As for determining inherent properties of features like stability or diffusability both databases have potentials for measuring these properties; this is confirmed by the fact that stability and diffusability metrics proposed in the literature can and have been applied to both (see section 3.2 below).

Quantitative methods can also be applied to the databases to determine what typology can tell us about possible genealogical relationships between languages. In this regard, JM consistently shows a noticeably better performance using simple phylogenetic algorithms, while the same methods applied to WALS mostly fail to identify known language families (cf. section 3.1 below). Even in JM, however, the interference of areal contact and statistical noise is rather high. Better results could be probably obtained applying stability and diffusability metrics in order to weight character differentially when calculating similarity matrices.

Several quantitative methods applied to JM are described in Polyakov and Solovyev (2006). The similarity metrics and methods of establishing genetic relationships described in this monograph are efficient and can be successfully applied to the JM data. Among other things, the so-called phenomenon of *typological shift* has been discovered and described. The essential observation here is that there is a growth of the frequency of already frequent features and an elimination of rare ones in the process of the evolution

having taken place within the last 2-3 millennia. Another technique introduced in the monograph is the so-called *language-feature diagram*, which provides the distribution of features according to their frequencies (i. e. for each x , the y value indicates how many features are represented for exactly x languages) . Finally the monograph suggests different data mining methods, but these could be replaced by more efficient methods borrowed from phylogenetics.

Databases can be used not only in academic research, but also for educational or reference purposes. A typological database can be a convenient tool for finding some information about a language or about the distribution of a feature across languages. JM is much more useful in the former regard, while WALS is more useful in the latter; this topic has been elaborated on in detail in the previous section.

The appearance of both WALS and JM potentially has great importance for the scientific community since these databases presents researchers with data in amounts that they could have never been able to operate with before. In this regard it is especially important to develop new tools and methodologies for both of the databases, making full use of their advantages and unique traits. WALS has already been made publicly available with its Interactive Reference Tool; JM has not yet been fully distributed, but is soon expected to become more accessible as well. The lack of an English translation of JM hinders the access of the worldwide scientific community to it. Such a translation would be a large and complicated project, but it has to be done at some point if JM's developers want it to become more widely used. The amount of work required, however, shrinks in comparison with the work already made to create both the encyclopædia and the database, and it can reasonably be expected to be carried out soon, if there is enough interest and demand in the scientific community.

3. Using WALs and JM

3.1. Typological similarity

The typological similarity between languages can be used to infer both areal and genealogical relationships. Neither WALs nor JM were conceived of as data collections serving the purpose of inferring genealogical relations. Nevertheless, using typological data for this purpose has generated some recent interest (Dunn et al. 2005, Wichmann and Saunders 2007), and it is therefore relevant to look into how well each of the two databases serve this purpose. It is impossible to perform a strict, quantified comparison since for this we would need data that are completely comparable, such as a given same number of WALs and JM features for the same set of languages. Instead we limit ourselves to an impressionistic comparison of the results of classifying a set of 38 languages represented both in WALs and JM. This set is selected on the basis of JM-WALS overlap and the languages cover most of the main Eurasian areas. In table 4 these languages are listed, along with the number of WALs features that are attested for each (in the fourth column). We do not give a similar numbers of attestations in JM since, as described above, it is not straightforward to assess precisely how well documented each language is in JM, but on the whole the languages are well documented with hundreds of features having positive values and many more that have negative values.

[INSERT TABLE 4 HERE]

As a yardstick for comparison we use lexicostatistical data. The reason for doing this is that we do not want to restrict ourselves to single families but want to submit the entire set of 38 languages to a classification procedure which is objective and has a principled way of distinguishing between families. The data for the lexicostatistical tree come from the ‘Automated Similarity Judgment Program’ (ASJP). They are constituted by 40-item subsets of the 100-item Swadesh list. The 40 items were chosen because of their higher stability (Holman et al. 2008b). The lists were transcribed in a phonologically simplified system using ascii characters (Brown et al. 2008a: Appendix C). The distances among languages were then measured using the 40-item lists on the basis of Levenshtein distances (Levenshtein 1966), as described in Holman et al. (2008b), cf. also the contribution by Bakker et al. to this issue of *Linguistic Typology*. A Levenshtein Distance (LD) is standardly defined as the minimum number of substitutions, insertions or deletions required to transform one word into another. To take into account the confounding factors of word length and accidental phonological similarities deriving from similar phoneme inventories, we modified each LD by first dividing it by the longest string among the two compared, obtaining LDN. This was then further divided by the average LDN of all pairs of words not having the same meaning and multiplying them by 100, to give LDND. These distances were fed to the Neighbour-Joining algorithm (Saitou and Nei 1987), which is the most standard method for deriving phylogenetic trees from distance data. The WALS and JM data, which are constituted by different values of different features—in other words by different character states—were transformed to distance data and similarly submitted to Neighbour-Joining (as implemented in Splitstree 4, cf. Huson and Bryant 2006). For both WALS and JM we simply used all features available for the languages selected.

[INSERT FIGS. 6-8 HERE]

In comparing the performances of the various trees we shall not go into too much detail, but will mostly just consider whether or not the classifications are in conformity with the 2-level classification of WALS into families and genera, as given in table 4.

The ASJP tree (fig. 6) mostly gets things right. The only places where the classification does not conform to WALS families and genera are the cases of Kabardian and Abkhaz (NW Caucasian), which have been split up by Chechen and Lezgian (Nakh-Daghestanian), and the failure of Hungarian and Khanty (Ugric) to unite under a single node. The placement of isolates (or languages without close relatives in the dataset) is not particularly telling, except that of Breton, which branches off from Romance languages, and appears to have a pulling effect on French, which is the next language to branch off from the cluster. The fact that Georgian is linked with Turkic and Hebrew with Ket is just a consequence of Georgian and Ket being isolates within the set, and they *must* go somewhere in the tree, even if the evidence for a relationship is very weak.

The JM tree (fig. 7) deviates from the commonly accepted classifications in the following ways (working bottom-up). The Romance group has been split up by Germanic intruders. Rather than forming an Indo-Iranian group as in the ASJP tree, Persian and Bengali have been split up. Next comes a group of Turkic languages joined by the Nakh-Daghestanian language Lezgian. This group is coordinate with a group of Uralic languages where Ugric, Finnic, and Samoyedic are not correctly grouped. Then follows a group where two languages—the Kartvelian language Georgian and the Yeniseic language Ket—do not have relatives within the set. The Northern Chukotko-Kamchatkan

language Chukchi appears here, in the vicinity of the Southern Chukotko-Kamchatkan language Itelmen. Between the two, the Northwest Caucasian language Abkhaz intrudes instead of appearing together with its relative Kabardian. In the upper end of the tree the Slavic languages come together (but note their separation from the rest of Indo-European).

When we turn to the WALS tree we encounter a mixture of genealogical information, areal information, and noise, perhaps in roughly equal proportions. Not much could be gained from commenting on the tree in detail. As a test of its performance we can just look at the nodes that unite two languages at the shallowest level of embedding. Going bottom up we find five pairs that make sense in as much as each pair belongs to the same genus: Italian-Catalan (Romance), Danish-Swedish (Germanic), Chuvash-Azerbaijani (Turkic), Tatar-Bashkir (Turkic), Abkhaz-Kabardian (Northwest Caucasian). The Itelmen-Chukchi pair of respectively Southern and Northern Chukotko-Kamchatkan languages also makes sense. Three other pairs are constituted by languages that do not belong to the same family or belong to different genera that are represented by other languages in the set: Ukrainian-Icelandic, Selkup-Uzbek, Bengali-Chechen, and Modern Hebrew-Persian. With the possible exception of the last pair there is no reason to believe that areal contact is responsible for these groupings. Clearly, the languages, as currently attested for the different features in WALS, are not classified with sufficient reliability in either a genealogical or areal sense. We realize, of course, that WALS was not designed with this purpose in mind, but the observation is nevertheless pertinent.

We conclude that while the lexical data yield fairly reliable genealogical classifications, the JM data yield a mixture of genealogical and areal information. Deviations from the standard genealogical scheme may be telling and could provide leads

concerning diffusion areas worthy of further investigation. A logical next step would be to devise a way to separate areal influences from genetic relationships in the JM tree. This could be done, in part, by assessing the stability of features, as explained in 3.2. Finally, WALS data are neither sufficiently reliable for extracting genealogical or areal information. Isolating certain sets of languages can lead both to meaningful genealogical patterns (Wichmann and Saunders 2007) and meaningful areal patterns (Donohue and Wichmann 2008), but one cannot expect much reliable information when picking a random set of languages like the one used here, which was selected using overlap in attestations with JM as the major criterion.

3.2. Measuring stabilities

Temporal stability of a given feature can, in a very general sense, be defined as its susceptibility to change. More precisely we could define stability as the probability that a given feature remains unchanged during an arbitrary period of time. An estimation of such a property could be made from typological distributions of features across languages. One of the first consistent attempts at such an estimate was made in Nichols (1995). A calculation of stabilities based on WALS data, close in spirit to Nichols' metrics, was made in Wichmann and Holman (n. d.). The reader may refer to this paper, to the shorter description in Holman et al. (2007: 414) or to the contribution to the present issue of *LT* by Bakker et al. for an overview of the method used.

In order to compare WALS and JM and to analyze how JM performs when applying such metrics, a calculation based on JM data using the same metric has been carried out. Since the results obtained, albeit preliminary, are important in comparing the two

databases, we will present several examples and highlight the most important conclusions drawn⁷.

In presenting our results we apply the same four-way categorization used in Wichmann and Holman (n. d.), that is:

- very stable: 51.8 – 100.0
- stable: 32.8 – 51.7
- unstable: 19.2 – 32.7
- very unstable: -62.8 – 18.9

for the binary values of the WALS features.

The features values to compare were selected on the following basis. Initially, stability was compared for a set of about 57 corresponding features of WALS and JM. As has been stated earlier, JM features, unlike WALS features, lack detailed descriptions, often making it hard to determine the exact meaning of a feature name, although in many cases the advantage of being able to address the original encyclopædia allows to clarify the sense of the feature in question. Thus only those features where the correspondence is more or less certain have been chosen. Besides, some of the correspondences are not one-to-one (when one of the databases is more detailed). Of these features, only those with the *U* value less than 90% have been chosen for both WALS and JM (a value of more than 90% means that the feature value is the same for most of the unrelated languages and the stability indices produced by the metric are dubious in such cases; refer to Wichmann and Holman (n. d.) for details). The following comparison is made for this feature set only.

[INSERT TABLE 5 HERE]

⁷ Cf. Belyaev (2009) for a more thorough comparison.

Table 5 shows that for precise numbers there is hardly any correlation between WALS and JM, except for a few rare cases where they are very close. In the four-way categorization, however, it turns out that most features either lie in the same category or at least in the adjacent ones. Exceptions are few, and they can partly be explained by dubious feature correspondences or errors and/or misrepresentation in JM data.

When JM and WALS features are compared with the statements in the literature presented in Wichmann and Holman (n. d.), the agreement is even better. Of the 13 statements for which correspondences have been analyzed, there is only one case where WALS agrees with the literature and JM does not, and that is verb-initial word order, which is said to be stable; in this case, however, the U value for JM is more than 90%, meaning that the number of languages with VSO word order in it is too small for the results to be trusted.

[INSERT TABLE 6 HERE]

This comparison has even more relaxed conditions than the comparison using the four-way categorization: it does not require a complete correspondence between the features and it is mostly vague with regard to exactly which features are implied. Therefore, as expected, here JM and WALS have an even better correlation.

These results allow us to draw the following conclusion with respect to JM-WALS comparison. The fact that the two databases produce numbers which do not correlate *per se* means that they are significantly different from each other in their structure and composition. That their correspondence increases the more vague and informal the

comparison becomes, however, means that, empirically, they describe roughly the same categories, and that there is, after all, a relatively high degree of overlap. Which of the two database designs is more suited to measuring stability is not entirely clear.

4. Conclusion: evaluations of JM and WALs and prospects for the future

We have made several comparisons of the two databases both from a theoretical standpoint and with respect to several practical applications. One conclusion we draw is that in terms of data structure JM and WALs are quite dissimilar, being based on different philosophies, so they are generally not very compatible. This is not necessarily to be lamented, however, since different structures may serve different needs.

On the other hand, both databases probably have a comparable degree of error and generally have similar data in the cases where they overlap, as has been shown in the stability example. JM, on the other hand, has many more features and describes parts of grammar not touched upon in WALs, something which allows it to perform better in cases where lots of statistics (like in the genealogical relations example) or exhaustive information (when used as a reference tool) is needed. It should be noted, though, that WALs does cover some aspects of grammar better than JM, although the latter is currently in constant development and extension, and these cases are relatively few. In terms of descriptive thoroughness JM by far outweighs WALs. One advantage in this regard is that it contains features for 55 extinct languages, which makes it a unique resource for certain diachronical studies, such as the study of changes in the typological profile of Eurasian languages presented in Polyakov and Solovyev (2006). Where WALs clearly fares much better than JM is in its greater areal coverage. On the basis of JM it would not be possible to even investigate, for instance, whether differential degrees of

stability is something inherent to typological features as opposed to being areally determined since only one (macro-)area is represented. The limitation of the coverage to Eurasia is not a design feature of JM, however, and may be remedied in the future.

The user interface is an area where both databases could be improved. The user interface of WALS is suitable for most basic tasks, but researchers still have to resort to writing their own tools for some more advanced studies, for example to combine more than two features, which is in many cases not enough. JM would benefit a lot from the addition of a mapping tool like the one WALS is equipped with, but, again, it is not imperative that both user interfaces have the same capabilities, since the databases are built for and suited to different kinds of tasks.

Several important studies have drawn upon data from WALS and JM, showing the usefulness of both. As regards JM, the explorations made by Polyakov and Solovyev (2006) have shown the data to be useful for dealing with genealogical relationships and the diachronic evolution of languages. Similarity metrics and other observations and techniques introduced in this monograph (cf. section 2.8) are a significant step forward for research in quantitative typology. As for WALS, we can cite Wichmann and Holman (n. d.) as an important exploration of the problem of stability, which shows that WALS data is quite informative in this regard. The papers in this issue of *Linguistic Typology* attest to many other uses that have been made of WALS.

JM currently contains two important disadvantages over against WALS: it lacks feature descriptions and it is only available in Russian. It is of highest priority to solve these two problems, and in both cases the challenges are great. These disadvantages, however, primarily relate to usability and do not concern the core structure of the database, which scientifically is of main importance. It should be stressed that it is still a

work in progress and should not yet be considered a finished product even after nearly two decades of development. Improvements are regularly being made, and new tools for both ordinary users and researchers are being created to address the problems outlined above.

As a final remark we would like to note that regardless of the advantages or disadvantages of the respective databases the concurrent development of both WALS and JM is of benefit to any future applications of quantitative methods in linguistic typology. As we show in two of our comparisons (typological similarity and temporal stability), having two different databases presents entirely new opportunities for verifying and double-checking methods and techniques used. In the end, both databases are expected to find many uses in the scientific community in the nearest decade, and both will surely serve as sources of inspiration for similar tools to be developed in the future.

Correspondence addresses

(Belyaev; corresponding author) Department of Theoretical and Applied Linguistics, Faculty of Philology, Moscow State University; GSP-2, Leninskie gory, MSU, Uchebniy Korpus 1, Moscow, Russia 119992, e-mail: obelyaev@gmail.com; (Polyakov) Institute of Linguistics of the Russian Academy of Sciences, Bolshoy Kislovsky lane, 1/12, Moscow, Russia 125009, e-mail: pvn-65@mail.ru; (Solovyev) Department of Theoretical Cybernetics, Faculty of Computer Science and Cybernetics, Kazan State University, Kremlevskaya St., 18, Kazan, Russia 420008, e-mail: MAKI.solovyev@mail.ru; (Wichmann) Department of Linguistics, Max-Planck-Institut für evolutionäre Anthropologie, Deutscher Platz 6, 04103 Leipzig, Germany; e-mail: wichmann@eva.mpg.de.

Acknowledgments

For their efforts in gathering and analyzing the empirical data that Figure 6 is based on we would like to acknowledge the members of the Automatic Similarity Judgment Project, in alphabetical order: Dik Bakker, Cecil H. Brown, Pamela Brown, Dmitri Egorov, Anthony Grant, Eric W. Holman, Hagen Jung, Robert Mailhammer, André Müller, Viveka Velupillai, and Kofi Yakpo (Belyaev and Wichmann also participate in this project).

Abbreviations

ASJP – Automatic Similarity Judgement Program, JM – *Jazyki Mira*, LD – Levenshtein Distance, LDN – Levenshtein Distance Normalized, LDND – Levenshtein Distance Normalized Divided, WALs – World Atlas of Language Structures.

References

- 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, scheduled for 2009. Available online at: <http://obelyaev.googlepages.com/BelyaevJMStab.pdf>.
- Brown, Cecil H., Eric W. Holman, Søren Wichmann, and Viveka Velupillai. 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.

- Comrie, Bernard and Norval Smith. 1977. *Lingua Descriptive Studies: Questionnaire* (= *Lingua* 42.1). Amsterdam: North-Holland.
- Croft, William. 1996. *Typology and Universals*. 1st ed. Cambridge: Cambridge University Press.
- Cysouw, Michael. 2007. A social layer for typological databases. In: Andrea Sansò (ed.) *Language Resources and Linguistic Theory*, 59-66. Milano: Francoangeli.
- Donohue, Mark, Søren Wichmann, and Mihai Albu. 2008. Typology, areality and relatedness. *Oceanic Linguistics* 47.1: 223-232.
- Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley, and Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309: 2072-2075.
- Haspelmath, Martin, Matthew Dryer, David Gil, and Bernard Comrie (eds.). 2005. *The World Atlas of Language Structures*. Oxford: Oxford University Press.
- Holman, Eric W., Christian Schulze, Dietrich Stauffer, and Søren Wichmann. 2007. On the relation between structural diversity and geographical distance among languages: observations and computer simulations. *Linguistic Typology* 11.2: 395-423.
- Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008a. Advances in automated language classification. In Arppe, Antti, Kaius Sinnemäki and Urpu Nikanne (eds), *Quantitative Investigations in Theoretical Linguistics*, 40-43. Helsinki: University of Helsinki.
- Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008b. Explorations in automated lexicostatistics. *Folia Linguistica* 42.2: 331-354.

- Huson, Daniel H. and David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23: 254-267.
- Jazyki mira: Ural'skie jazyki* (Languages of the World: Uralic languages). 1993. Moscow.
- Jazyki mira: Túrskie jazyki* (Languages of the World: Turkic languages). 1997. Moscow: Indrik.
- Jazyki mira. Paleoaziatskie jazyki* (Languages of the World. Palaeoasiatic languages). 1996. Moscow: Indrik.
- 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: Iranskie jazyki. I. Jugo-zapadnye iranskie jazyki* (Languages of the World: Iranian languages. I. Southwest Iranian languages). 1997. Moscow: Indrik.
- Jazyki mira: Dardskie i nuristanskie jazyki* (Languages of the World: Dardic and Nuristani languages). 1998. Moscow: Indrik.
- Jazyki mira: Germanskie jazyki. Kel'tskie jazyki* (Languages of the World: Germanic languages. Celtic languages). 1999. Moscow: Academia.
- Jazyki mira: Iranskie jazyki. II. Severo-zapadnye iranskie jazyki* (Languages of the World: Iranian languages. II. Northwest Iranian languages). 1999. Moscow: Indrik.
- Jazyki mira: Iranskie jazyki. III. Vostočnoiranskie jazyki* (Languages of the World: Iranian languages. III. East Iranian languages). 1999. Moscow: Indrik.
- Jazyki mira: Kavkazskie jazyki* (Languages of the World: Caucasian languages). 2001. Moscow: Academia.

- Jazyki mira: Romanskije jazyki* (Languages of the World: Romance languages). 2001. Moscow: Academia.
- Jazyki mira: Indoarijskie jazyki drevnego i srednego perioda* (Languages of the World: Old and Middle IndoAryan languages). 2004. Moscow: Academia.
- Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory* 10(8):707-710.
- 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.
- Nichols, Johanna. 1992. *Linguistic Diversity in Space and Time*. Chicago: The University of Chicago Press.
- Nichols, Johanna. 1995. Diachronically stable structural features. *Historical Linguistics 1993. Selected Papers from the 11th International Conference on Historical Linguistics, Los Angeles 16-20 August 1993*, ed. by Henning Andersen, 337-355. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Nichols, Johanna. 2003. Diversity and stability in languages. *The Handbook of Historical Linguistics*, ed. by Brian D. Joseph, and Richard D. Janda, 283-310. Oxford: Oxford University Press. Malden/Oxford/Melbourne/Berlin: Blackwell Publishing.
- Novikov, A. I. and E. I. Yaroslavtseva. 1985. Baza lingvotipologičeskix dannyx I principy eë funkcionirovanija (A Database for Linguistic Typology and Principles of its Operation). *Vesti AN SSSR* 3.
- 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.

- Saitou, Naruya and Masatoshi Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4: 406-425.
- Toporov, V. N., M. V. Zavyalova, A. A. Kibrik et al. (eds.). 2006. *Jazyki mira: Baltijskie jazyki* (Languages of the World: Baltic languages). Moscow: Academia.
- Yartseva, V. N. 1982-1. *Jazyki i dialekty mira* (Languages and Dialects of the World). Moscow: Nauka.
- Yartseva, V. N. (ed.) 1980. *Teoretičeskie osnovy klassifikacii jazykov mira* (Theoretical Foundations of the Classification of the World's Languages). Moscow.
- Yartseva, V. N. (ed.) 1982-2. *Teoretičeskie osnovy klassifikacii jazykov mira. Problemy rodstva* (Theoretical Foundations of the Classification of the World's Languages. Problems of Genetic Relationship). Moscow.
- Yartseva, V. N. and B. A. Besserebrennikov (eds.). 1976. *Principy opisanija jazykov mira* (Principles for Description of the World's Languages). Moscow.
- Wichmann, Søren and Eric W. Holman. N. d. Assessing temporal stability for linguistic typological features. Manuscript under review. Available online at: <http://email.eva.mpg.de/~wichmann/WichmannHolmanIniSubmit.pdf>.
- Wichmann, Søren and Arpiar Saunders. 2007. How to use typological databases in historical linguistic research. *Diachronica* 24.2: 373-404.
- Zhurinskaya, Novikov, Yaroslavtseva. 1986. *Ènciklopedičeskoe opisanie jazykov. Teoretičeskie i prikladnye aspekty* (Encyclopædic description of languages. Theoretical and applied aspects). Moscow: Nauka.

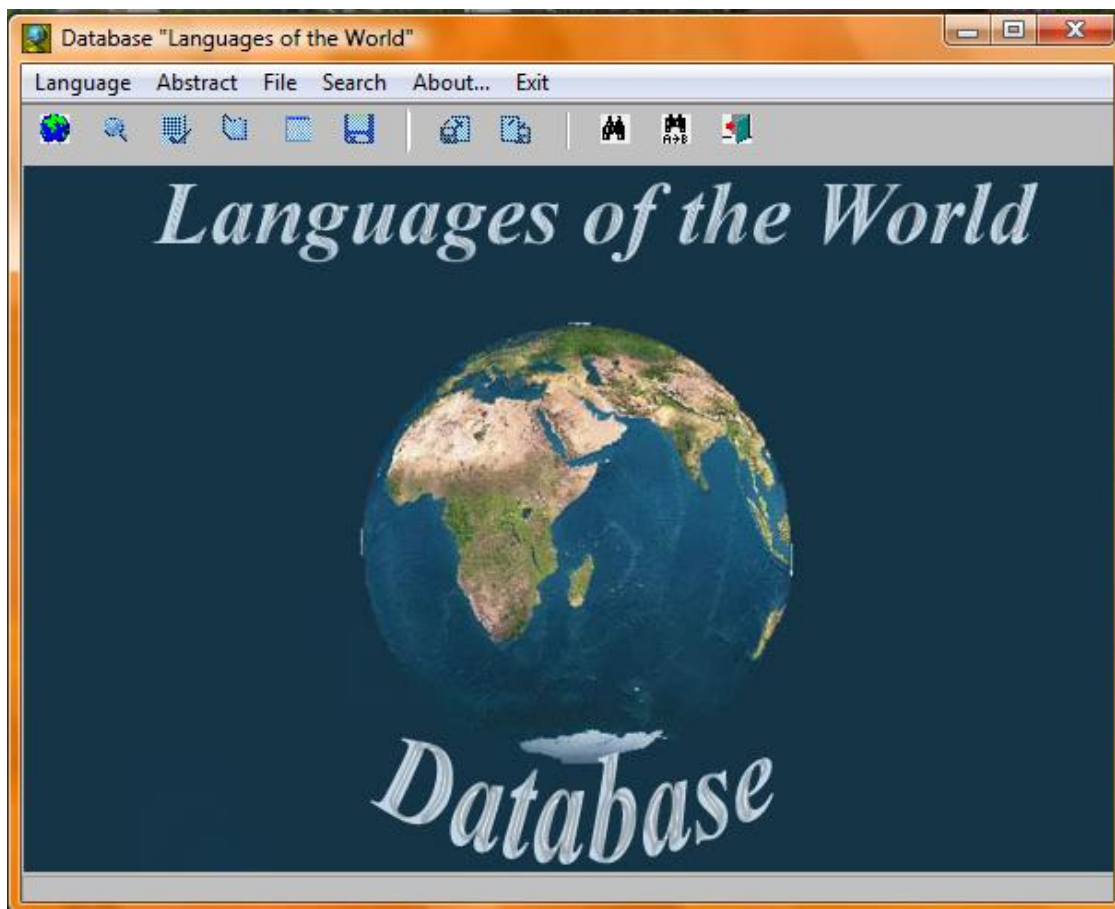


Fig 1. The initial window of the current Jazyki Mira interface.

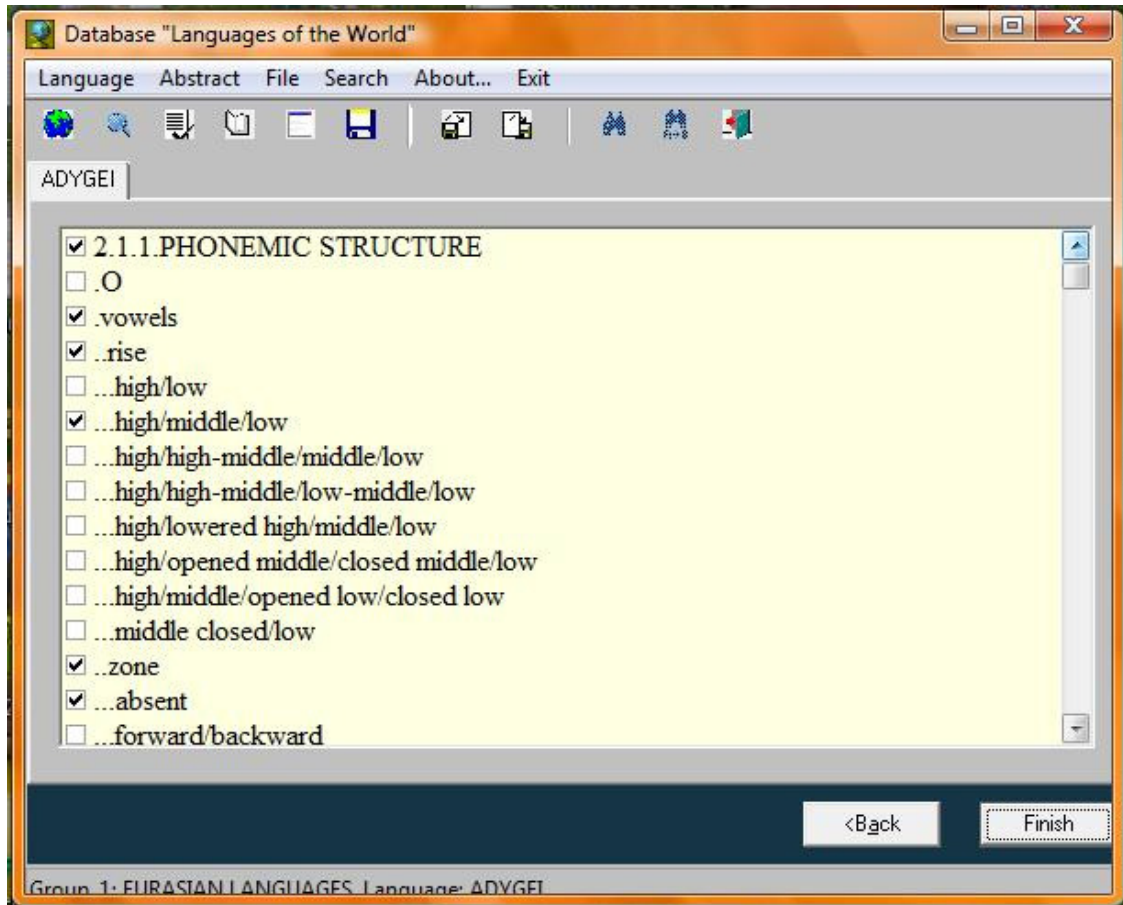


Fig. 2. JM: viewing the Adygh language.

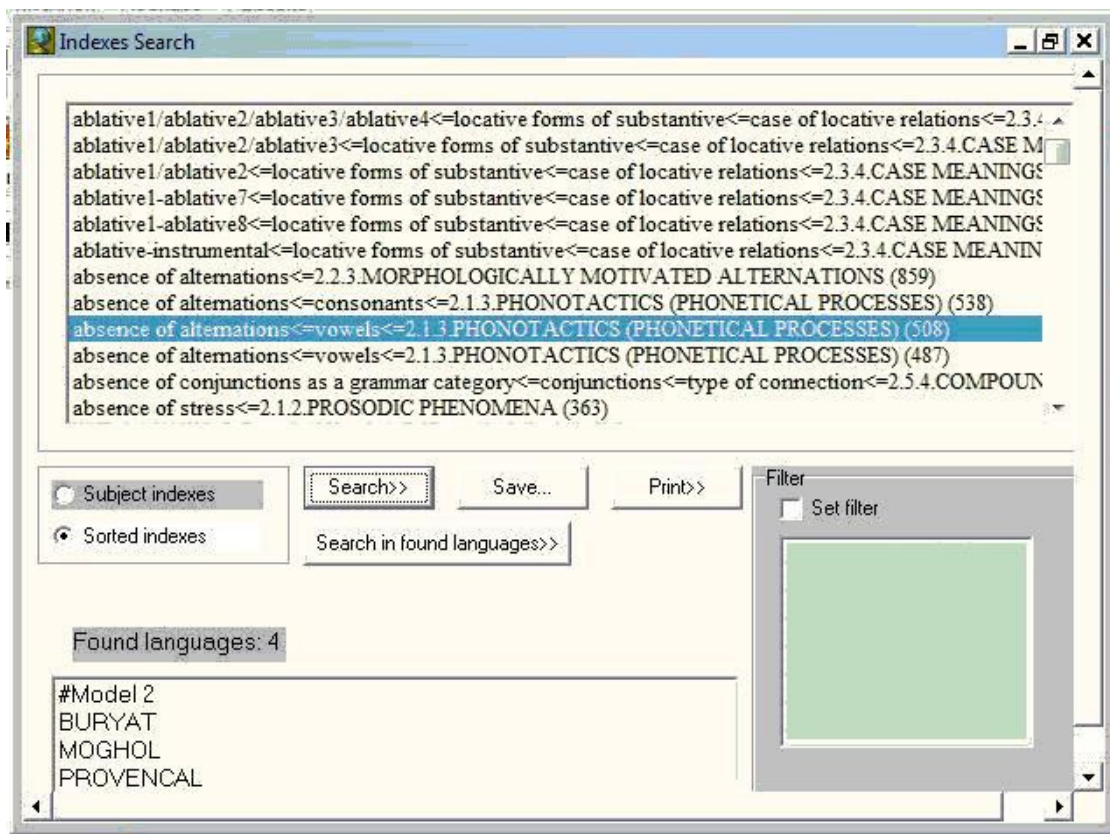


Fig. 3. JM: Search of languages by a feature.

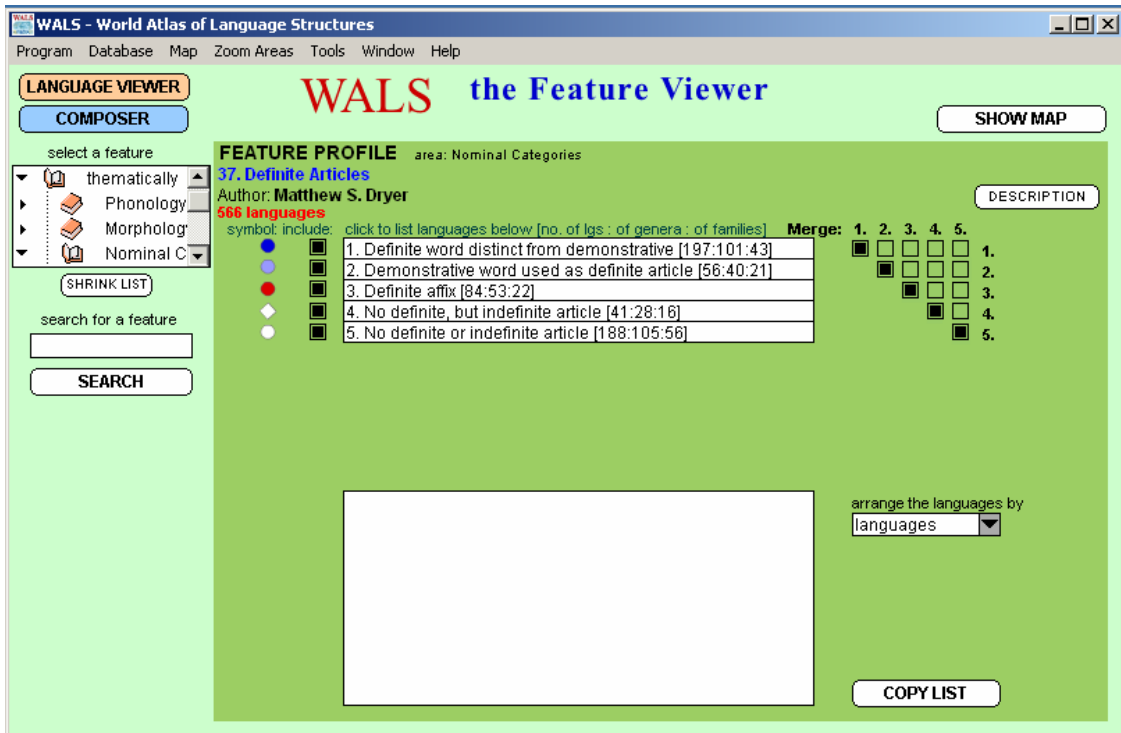


Fig. 4. WALS: the Feature Viewer.

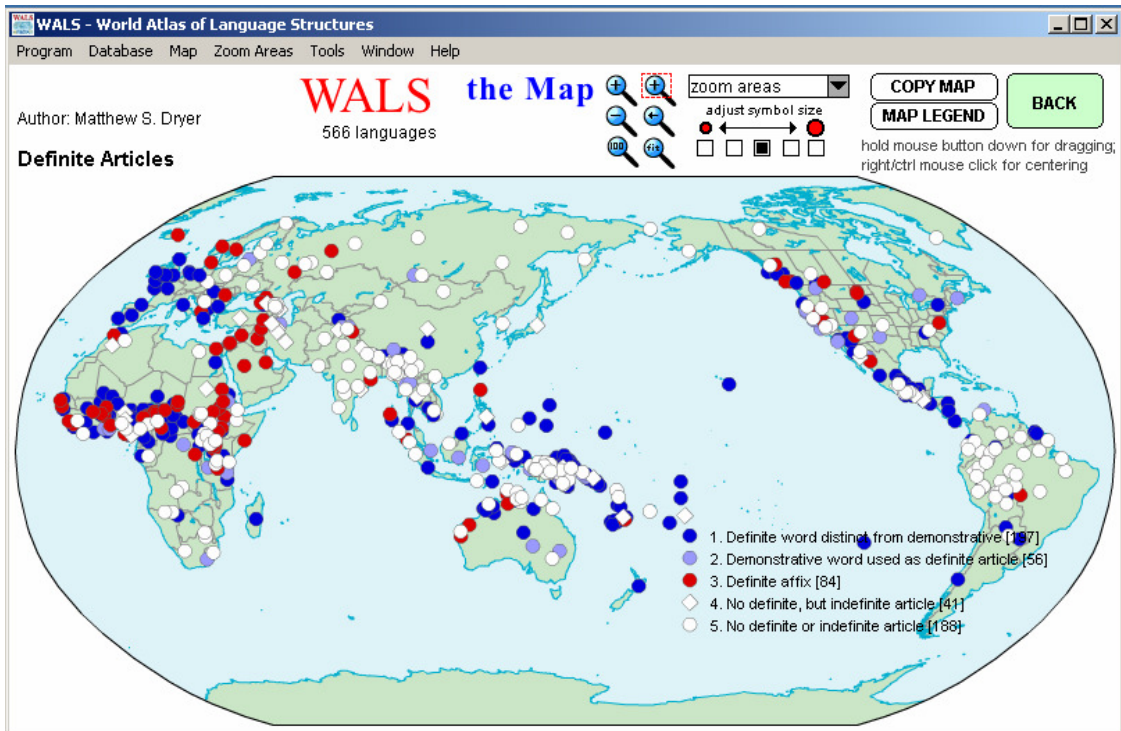


Fig. 5. WALS: an example of the mapping tool.

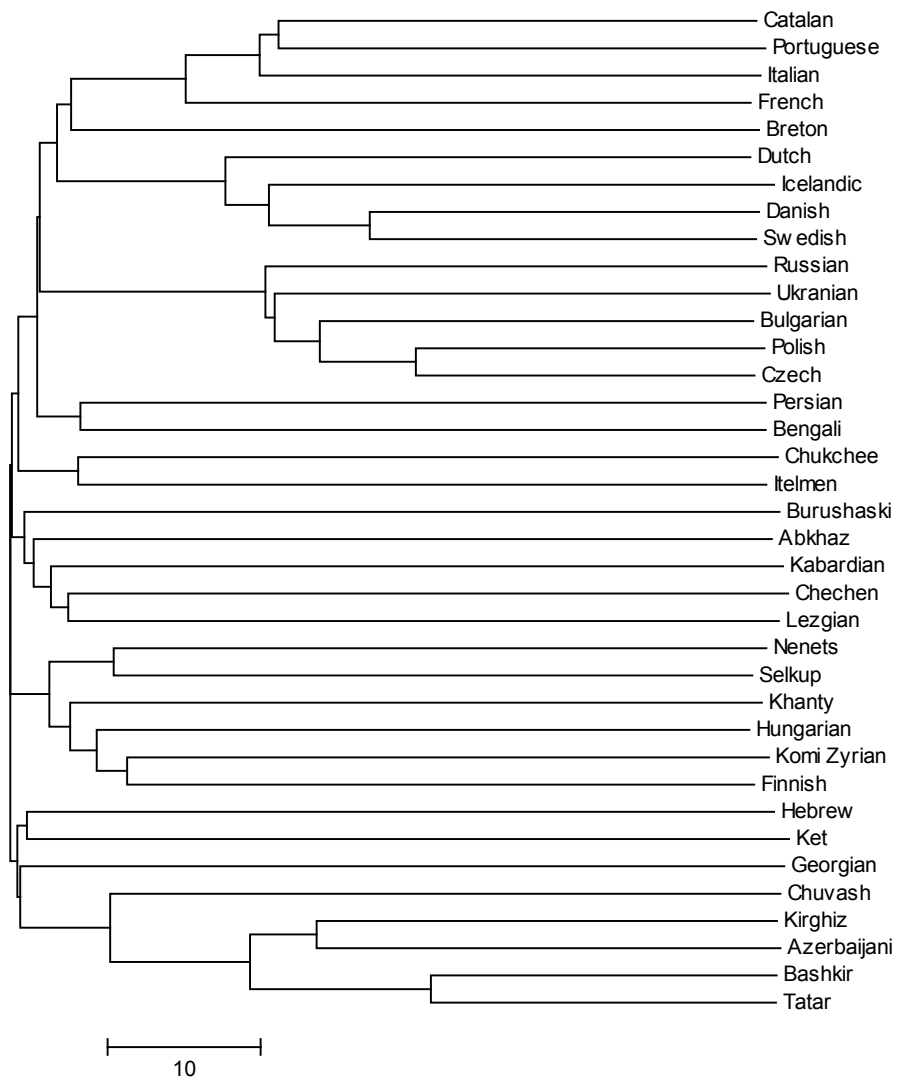


Fig. 6. ASJP tree for the languages selected.

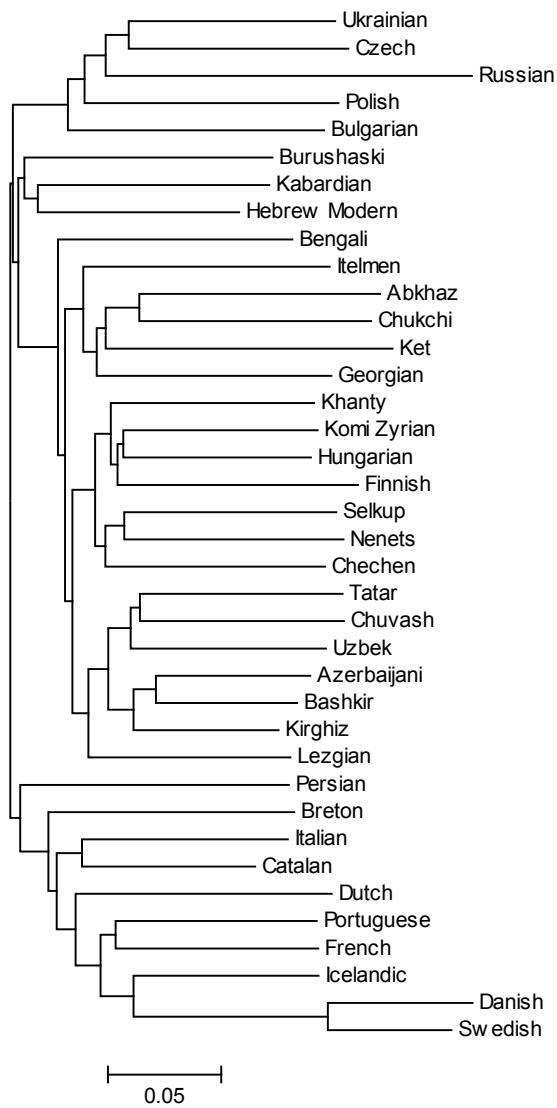


Fig. 7. JM tree for the languages selected.

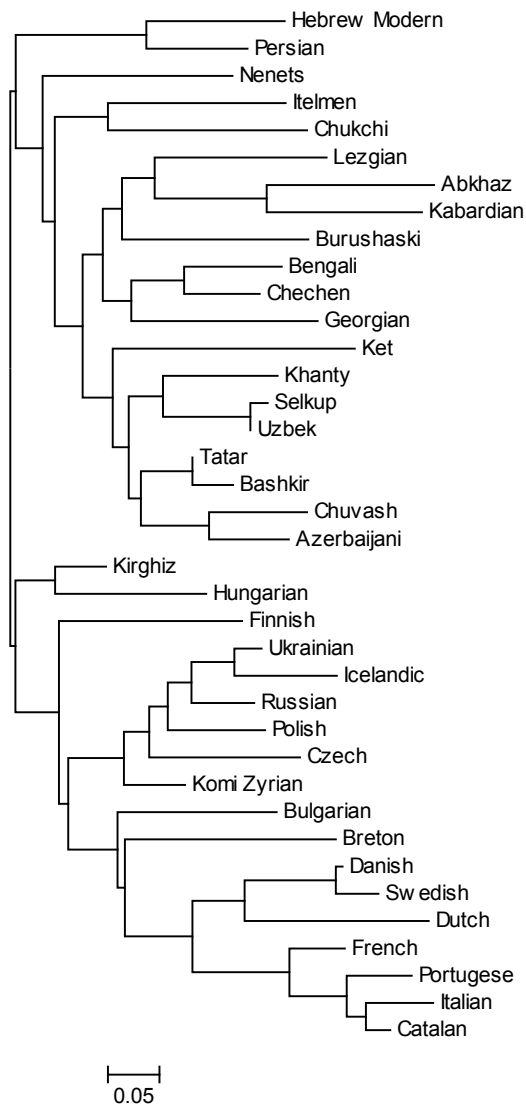


Fig. 8. WALS tree for the languages selected.

2.1.1	Phonological structure
2.1.2	Prosody
2.1.3	Phonetics
2.1.4	The syllable
2.2.1	Phonotactics
2.2.2	Phonological opposition between morphological categories ^a
2.2.3	Morphologically motivated alternations
2.3.0	Morphological type
2.3.1	Criteria for parts of speech assignment
2.3.2	Nouns
2.3.3	Number
2.3.4	Case
2.3.5	Verbal categories
2.3.6	Deictic categories
2.3.7	Parts of speech
2.4.0	Structure of morphological paradigms
2.5.1	Word structure
2.5.2	Word formation
2.5.3	The simple sentence
2.5.4	The complex sentence

Table 1. JM chapter headings

^a The term “morphological categories” refers to categories of phonological units such as words, clitics, and affixes.

Number in data-base	Наименование и уровень в иерархии признака	Name and level in the feature hierarchy
3773	2.5.4.СЛОЖНОЕ ПРЕДЛОЖЕНИЕ	2.5.4.COMPLEX SENTENCES
3774	.O	.O (this symbol, when annotated with the value ‘true’ indicates that information for the features within the section is absent)
3775	..линейный порядок компонентов	..linear order of clauses
3776	..главное предшествует придаточному	..the main clause precedes the subordinate clause
3777	..главное предшествует целевому	..the main clause precedes the purpose clause
3778	..не фиксирован	..free order
3779	..придаточное предшествует главному	..the subordinate clause precedes the main clause
3780	..целевое предшествует главному	..the purpose clause precedes the main clause
3781	..особенности подчиненного компонента	..features of the subordinate clause
3782	..особое оформление именных групп	..special marking of noun phrases
3783	..особое оформление подлежащего	..specific marking of a subject
3784	..особый порядок слов	..specific word order
3785	..оформление сказуемого	.. marking of a predicate
3786	...зависимые личные формы	...dependent finite forms
3787	...квотатив	...quotative
3788	...нефинитные формы	...non-finite forms
3789	...оборот с релятивной формой глагола	...phrase with a relativized verb form
3790	...особые глагольные показатели	...specific verbal markers
3791	...особый порядок слов	...specific word order
3792	...таксисные конструкции	...taxis
3793	...финитные формы	...finite verb forms
3794	...абсолютные обороты	... absolute constructions (e.g., Latin Ablativus)

Number in data-base	Наименование и уровень в иерархии признака	Name and level in the feature hierarchy
		Absolutus)
3795	..'тематическое' придаточное	.. topical dependent clause
3796	.особенности построения дискурса	.peculiarities of discourse structure
3797	..обвиатив	..obviation
3798	..переключение референции	.. switch-reference
3799	.структура относительного предложения	.relative clause structure
3800	..главное предшествует придаточному	.. the main clause precedes the subordinate clause
3801	..относительно-местоименные аффиксы в глаголе	..relative pronominal affixes on a verb
3802	..отсутствие союзного слова или его аналога	..absence of a conjunction or its analogue
3803	..придаточное предшествует главному	..the subordinate clause precedes the main clause
3804	..союзное слово в главном	..a conjunction in the main clause
3805	..союзное слово в придаточном	..a conjunction in the subordinate clause
3806	..сказуемое придаточного следует за союзом	..the predicate of the subordinate clause follows the conjunction
3807	.тип построения	.type of construction
3808	..сериализация	..serialization
3809	..сочинение/подчинение	..coordination/subordination
3810	..только подчинение	..only subordination
3811	..только сочинение	..only coordination
3812	.тип связи	.type of connection
3813	..союзная/бессоюзная	..conjunctive/non-conjunctive
3814	..преобладает бессоюзная	..non-conjunctive prevails
3815	..преобладает союзная	..conjunctive prevails

Number in data-base	Наименование и уровень в иерархии признака	Name and level in the feature hierarchy
3816	...союзы	..conjunctions
3817	...знаменательные слова	...full words (as opposed to particles) used as conjunctions
3818	...отсутствие союзов как грамматического разряда	...absence of conjunctions as a grammatical category
3819	...присоединяемые служебные элементы	...attached syntactic (non-lexical) elements
3820	...самостоятельные служебные элементы	...independent syntactic (non-lexical) elements
3821	...союзные формы глагола	...conjunctive verb forms

Table 2. Example of the organization of a section within JM

	No. of errors in database (and in printed version)	No. of features attested with positive values	% errors of all the features (3821) in database
Danish	85 (18)	391	2.2%
Norwegian	36 (12)	415	0.9%
Swedish	50 (18)	378	1.3%

Table 3. Error rates in JM data for three Scandinavian languages

Language	Family	Genus	WALS Features
Modern			
Hebrew	Afro-Asiatic	Semitic	13
Chuvash	Altaic	Turkic	68
Uzbek			52
Bashkir			51
Tatar			43
Azerbaijani			41
Kirghiz			38
Burushaski	Burushaski	Burushaski	121
Chukchi	Chukotko-	Northern Chukotko-Kamchatkan	121
Itelmen	Kamchatkan	Southern Chukotko-Kamchatkan	46
Breton	Indo-European	Celtic	56
Dutch		Germanic	67
Swedish			64
Icelandic			61
Danish			45
Bengali		Indic	40
Persian		Iranian	127
French		Romance	136
Italian			68
Portugese			47
Catalan		47	
Russian		Slavic	134
Polish			76
Bulgarian			70
Czech			45
Ukrainian			41
Georgian		Kartvelian	Kartvelian
Lezgian	Nakh-Daghestanian	Lezgcic	129
Chechen		Nakh	40
Abkhaz	Northwest		128
Kabardian	Caucasian	Northwest Caucasian	46
Finnish	Uralic	Finnic	134
Komi Zyrian			37
Nenets		Samoyedic	95
Selkup			42
Hungarian		Ugric	132
Ket	Yeniseian	Yeniseian	104

Table 4. A comparative selection of languages (classified by families and genera, following WALS)

WALS feature	Stability	JM feature	Stability
26.2. Inflectional Morphology: Predominantly suffixing	66.80% (very stable)	3443 ..преимущественно суффиксальная (predominately suffixing)	22.77% (unstable)
31.1. No gender system	77.80% (very stable)	1150 ..род (gender) ^b	80.44% (very stable)
31.2 Sex-based gender system	81.10% (very stable)	1200 ...пол (sex-based motivation)	61.10% (very stable)
Average of Plural prefix (33.1) and plural suffix (33.2)	55.90% (very stable)	1257 ..аффиксация (plural affix)	37.97% (stable)
33.7. Plural word	25.40% (unstable)	1268 ..служебные слова (functional words)	17.13% (very unstable)
37.2. Demonstrative word used as marker of definiteness	-1.70% (very unstable)	2767 ...указательными местоимениями (demonstrative pronouns) ^c	26.53% (unstable)
37.3. Definite affix on noun	21.00% (unstable)	2745 ..именные аффиксы (nominal affixes)	23.61% (unstable)

^b Since the metric used gives identical results for the presence and absence of a feature when the feature in question is binary (cf. WALS features like “Inflectional optative present” and “Inflectional optative absent”), we sometimes provide a negative counterpart of a feature if no positive counterpart is found.

^c JM features related to definiteness are actually concerned with any definiteness/indefiniteness distinction. This may be the reason for some of the inconsistencies.

38.5. Neither definite nor indefinite article	16.80% (very unstable)	2740 .определенность/неопределенность имени (nominal definiteness/indefiniteness)	37.47% (stable)
41.2. Distance Contrasts in Demonstratives: Two-way contrast	8.50% (very unstable)	2615 ...ближний план/дальний план (close/far plane) and 2616 ...верхний план/нижний план (upper/lower plane)	31.40% (unstable)
41.3. Three-way contrast	20.00% (unstable)	2619 ...ближний план/средний план/дальний план	19.18% (unstable)
54.1. No distributive numerals	37.00% (stable)	1326 ..распределительные (разделительные) (distributives)	38.73% (stable)
65.1. Grammatical marking of perfective/imperfective distinction	36.00% (stable)	1842 ...совершенный/ несовершенный (perfective/imperfective)	26.17% (unstable)
73.1. Inflectional optative present	56.70% (very stable)	2359 ...желательность (optative)	21.85% (unstable)
81.1. Subject-object-verb (SOV)	69.50% (very stable)	3657 ...SOV	50.90% (stable)
81.2. Subject-verb-object (SVO)	59.20% (very stable)	3656 ...SVO	62.24% (very stable)
87.1. Modifying adjective precedes noun (AdjN)	59.20% (very stable)	3664 ...адъективное опред. предшествует определяемому (AdjN)	6.85% (very unstable)

107.1. There is a passive construction	28.30% (unstable)	- 1777 ...пассив (passive)	42.03% (stable)
112.1. Negative affix	36.80% (stable)	2809 ...отрицательные аффиксы	28.21% (unstable)
98.2. Nominative - accusative (standard)	15.80% (very unstable)	3639 ..номинативный (nominative) ^d	62.86% (very stable)
98.4. Ergative - absolutive	65.50% (very stable)	3648 ...эргативный (ergative)	79.25% (very stable)
102.1. No person marking of any argument	12.20% (very unstable)	3712 ..актант определяет форму предиката (lit. "agent determines the form of the predicate")	25.29% (unstable)
102.2. Person marking of only the A argument	38.90% (stable)	3715по лицу (verbal person agrees with agent)	31.87% (unstable)
102.3. Person marking of only the P argument	24.40% (very unstable)	3723по лицу (verbal person agrees with patient)	35.59% (stable)

Table 5. Comparison of WALs and JM stabilities

^d Unlike WALs, JM does not include separate features for case marking of full NPs, pronouns etc, but only a set of general features for the whole language. This may be an explanation for the inconsistencies.

Statement in the literature	WALS feature	WALS stability	WALS agrees	JM feature	JM stability	JM agrees
SVO is possibly stable (Nichols 2003: 286, 305; Croft 1996: 206-7)	Subject-verb-object (SVO)	59.2% (very stable)	<i>Yes</i>	3656 ...SVO	62.24% (very stable)	<i>Yes</i>
Ergativity has a low probability of inheritance (Nichols 2003: 295) ^e	99.4 Alignment of Case Marking of Pronouns: Ergative — absolute	74.9% (very stable)	<i>No</i>	1390 ...эргатив/абсолюти в (ergative/absolute case present)	59% (very stable)	<i>No</i>

Table 6. Two example comparisons with statements in the literature.

^e This statement is at somewhat at variance with or qualifies an earlier assessment of Nichols (1992: 167) that alignment is highly genetically stable.