American Journal of Applied Sciences

Original Research Paper

Can Grammar Define Similarity of Human Natural Languages?

[1]Vladimir Nikolaevich Polyakov, [2]Ivan Sergeevich Anisimov and [3]Elena Andreevna Makarova

[1]*NUST, "MISIS", Moscow, Russia*
[2]*"Yandex", LLC, Moscow, Russia*
[3]*Institute of Linguistics of RAS, Moscow, Russia*

Article history
Received: 11-06-2016
Revised: 03-10-2016
Accepted: 05-09-2016

Corresponding Author:
Vladimir Nikolaevich Polyakov
NUST "MISIS", Moscow, Russia
Email: pvn-65@mail.ru

Abstract: The aim of the present study is to show that similarity of human natural languages can be conveyed not only by phonetic data, but also by grammar. The paper regards the largest typological database WALS and its possibilities in the sphere of genealogic relationship of languages. Using the method of two-objective optimization and data mining, which is new for linguistic studies, we show that grammatical (structural) data, as well as phonetic data, can deliver information on the similarity of languages. Language isolates and micro-families do not have genealogic relatives based on phonetic information, but they do have genealogic relatives based on grammar information.

Keywords: WALS, Two-Objective Optimization, Data Mining, Language Isolates, Micro-Families, Similarity, Grammar

Introduction

The main source of information for the present study is World Atlas of Language Structures-WALS (Haspelmath *et al*., 2005). It is the world's biggest database describing structural properties of languages. The first version of the database appeared as a book and a CD with a stand-alone application "Interactive Reference Tool (WALS Program)". WALS Program contains description for 2,560 languages according to 142 features. The online version of the database (Dryer and Haspelmath, 2013) was published in 2011. Now, it describes 2,679 languages according to 144 features (Comrie *et al*., 2013) and some of the features are divided into several parts each including a separate map and set of languages.

The core of WALS is made by chapters dedicated to features that are divided into eight main areas encompassing the major structural domains of grammar: Phonology, morphology, nominal categories, nominal syntax, verbal categories, word order, simple clauses and complex sentences. There are also smaller areas describing lexicon, sign languages, writing systems and paralinguistic usage of clicks.

WALS contains 160 maps showing the geographical distribution of the features. Besides, the program allows combining up to four different features on a single map, which can be useful for studying correlation of features.

Full grammar books, dictionaries, dissertations, articles, field notes, questionnaires and authors' own knowledge became the source of information. WALS, being the result of work of 55 specialists, was the first feature atlas on a world-wide scale, which made it possible to compare unrelated languages all over the world.

The present study uses the data WALS in an attempt to solve the problem that has been pending for many decades, if not a hundred years, -classification of language isolates. A language isolate does not have any demonstrable genealogic relationship with any existing language, i.e., the traditional methods of comparative linguistics based on the comparison of basic concepts (100-word lists) (Swadesh, 1952) were unable to define genealogic relations between a language isolate and any other languages.

Many linguists believe that all languages on our planet had originated from a single mother tongue, the reason why the problem of classifying language isolates has become the object of numerous studies (Tallerman and Gibson, 2012).

In contrast to unclassified languages, which are likely to demonstrate a genealogic relationship with some family once they are studied further, language isolates have already been thoroughly described, but they still do not classify with any known language family.

In 1939 Prince Trubetzkoy defined six grammar features characteristic of all Indo-European languages (Trubetzkoy, 1939). Over a decade later, Emile Benveniste started an indirect dispute with him in (Benveniste, 1954). He claimed that one of the languages

of North-American Indians-Takelma-possessed all six grammar features indicated by Trubetzkoy. But Takelma was a language isolate (i.e., it did not have any genealogic relatives), moreover, it was located on a different continent from speakers of Indo-European languages. Thus, there was no way Takelma could be a representative of Indo-European languages. We shall return to this question in the Discussion section.

In the present paper, we suggest using the methods of two-objective optimization (Ehrgott, 2000; Izraylevich and Tsudikman, 2012) for pair-wise comparison of typological language profiles. This might be able to cast light on the genealogic relationship of some language isolates, a question that classic methods of comparative linguistics failed to answer.

In addition, WALS has undergone data cleaning in style of data mining so that only structural features are compared. It allows finding a distinct border between well described and poorly described pairs of languages. In the future, if more information is added to WALS, the problem of classifying language isolates and microfamilies can be successfully solved.

The present research is compared to a seemingly similar study, conducted by Wichmann and Holman (2010), but, as shown below, the method and the results presented in this study, are quite different (cf. Section 4).

Approaches concerning the scenarios of evolution basing on WALS features were used in works of Gray *et al*. (2011).

Materials and Methods

Two-Objective Optimization

The present study uses the method of two-objective optimization, first described in (Edgeworth, 1881). This method approved itself in computer science (Ehrgott, 2000) and is widely used for solving economic and technical problems (Izraylevich and Tsudikman, 2012).

The following two criteria are considered: The percentage of matching features values from WALS Program and the number of coinciding features. The point is that if only the percentage of matching values of features is considered, there is likely to be a language with the number of described features equaling, for example, 2 and their values will both match the values of the language under study. In order to exclude such unreliable data, the second parameter-the number of coinciding features is introduced. It was experimentally established to be 30, while the percentage of matching values must be at least 65%.

In two-objective optimization, one criterion is chosen as the main one and the other becomes subsidiary. In this study, the main criterion is the percentage of matching values of features and the restricting criterion is the number of coinciding features. The latter is restricted from below (language under study and structurally maximally close languages must have at least 30 coinciding features); and the former tends to reach 100%, but must be at least 65%.

Database "Isolates"

The information for the creation of the database "Isolates" (which is used for search of Structurally Maximally Close languages (SMC-languages) was taken from WALS Program, as it is available for download and presented in form of an electronic database, which simplifies computer analysis of the information. The data necessary for the present study include list of languages, list of features and the descriptions of the languages according to the features.

Description of the Method

The aim of the method is to find a language that has the highest percentage of matching values of features for any language under study and, at the same time, meets the requirements, i.e., the number of coinciding features is at least 30 and the percentage of matching values is at least 65%. The given choice of parameters will be discussed further.

MS Access, being the easiest and the most accessible tool, is chosen to be used in this study. Four queries will be needed: QUERY-1, QUERY-2, QUERY-3 and QUERY-4.

QUERY-1 (Fig. 1) allows getting a complete list of languages, features and values of grammar features. A total of 58,019 entries (according to WALS Program) were received.

QUERY-2 (Fig. 2) allows getting from the database a list of languages and an integrated number of their features whose values match the values of the language under study. The list is sorted by the decrease in the number of matches. QUERY-2 is built by uniting two queries QUERY-1.

QUERY-3 (Fig. 3) allows getting a list of languages and an integrated number of their features from the database whose values do not match the values of the language isolate. The list is sorted by the decrease in the number of mismatches.

QUERY-3 is built by uniting two queries QUERY-2.

QUERY-4 allows getting from the DB "Isolates" a list of languages and an integrated number of their features whose values match the values of the language isolate. The list is sorted by the decrease in the number of matches. The query cuts off the results with the total number of coinciding features below 30 and sorts the search results from the highest percentage of matching values (Fig. 4). QUERY-4 is built by uniting two queries: QUERY-2 and QUERY-3.

DOI: 10.3844/ajassp.2016.1040.1052

QUERY_1

```
SELECT language.lang_id, language.name_, feature.feat_ID, feature.feat_title, x_feat_lang.index_, x_feat_value.svf
FROM [language] INNER JOIN (feature INNER JOIN x_feat_lang ON feature.feat_ID = x_feat_lang.feat_ID) ON language.lang_id = x_feat_lang.lang_ID) INNER JOIN x_feat_value ON (x_feat_lang.index_ =
x_feat_value.index_) AND (feature.feat_ID = x_feat_value.feat_ID)
ORDER BY language.lang_id, language.name_, feature.feat_ID;
```

lang_id	name_	feat_ID	feat_title	svf	index_
1	"A-Pucikwar"	9	The Velar Nasal	Initial velar nasal	1
2	"Aari"	26	Prefixing vs. Suffixing in Inflecti	Strongly suffixing	2
2	"Aari"	33	Coding of Nominal Plurality	No plural	9
2	"Aari"	37	Definite Articles	Definite affix	3
2	"Aari"	38	Indefinite Articles	No indefinite, but definite arti	4
2	"Aari"	51	Position of Case Affixes	Case suffixes	1
2	"Aari"	57	Position of Pronominal Possessi	No possessive affixes	4
2	"Aari"	69	Position of Tense-Aspect Affixes	Tense-aspect suffixes	2
2	"Aari"	82	Order of Subject and Verb	SV	1
2	"Aari"	83	Order of Object and Verb	OV	1
2	"Aari"	85	Order of Adposition and Noun P	Postpositions	1
2	"Aari"	86	Order of Genitive and Noun	Genitive-Noun	1
2	"Aari"	87	Order of Adjective and Noun	Noun-Adjective	2
2	"Aari"	88	Order of Demonstrative and No	Noun-Demonstrative	2
2	"Aari"	89	Order of Numeral and Noun	Noun-Numeral	2
2	"Aari"	90	Order of Relative Clause and No	Noun-Relative clause	1
2	"Aari"	92	Position of Polar Question Partic	No question particle	6
2	"Aari"	95	Relationship between the Order	OV and Postpositions	1

Fig. 1. QUERY-1

QUERY_2

```
SELECT [QUERY_1].lang_id, [QUERY_1].name_, QUERY_777_1.lang_id, QUERY_777_1.name_, Count([QUERY_1].feat_ID) AS [Count-feat_ID], Count([QUERY_1].feat_title) AS [Count-feat_title],
Count([QUERY_1].index_) AS [Count-index_]
FROM QUERY_1 INNER JOIN QUERY_1 AS QUERY_777_1 ON ([QUERY_1].feat_ID=QUERY_777_1.feat_ID) AND ([QUERY_1].index_=QUERY_777_1.index_)
GROUP BY [QUERY_1].lang_id, [QUERY_1].name_, QUERY_777_1.lang_id, QUERY_777_1.name_
HAVING ((([QUERY_1].lang_id)=703))
ORDER BY Count([QUERY_1].feat_ID) DESC , Count([QUERY_1].feat_title) DESC , Count([QUERY_1].index_) DESC;
```

QUERY_1.lang_id	QUERY_1.na	QUERY_777_	QUERY_777_1.name_	Count-feat_ID	Count-feat_title	Count-index
703	"German"	703	"German"	129	129	129
703	"German"	655	"French"	91	91	91
703	"German"	626	"English"	90	90	90
703	"German"	1952	"Russian"	87	87	87
703	"German"	742	"Greek (Modern)"	77	77	77
703	"German"	2089	"Spanish"	75	75	75
703	"German"	647	"Finnish"	72	72	72
703	"German"	849	"Hungarian"	69	69	69
703	"German"	1251	"Latvian"	64	64	64
703	"German"	1008	"Kannada"	61	61	61
703	"German"	825	"Hindi"	58	58	58
703	"German"	2311	"Turkish"	55	55	55
703	"German"	816	"Hebrew (Modern)"	55	55	55
703	"German"	1155	"Korean"	53	53	53
703	"German"	1836	"Persian"	53	53	53
703	"German"	602	"Dutch"	52	52	52
703	"German"	1270	"Lezgian"	50	50	50
703	"German"	109	"Apurinã"	50	50	50

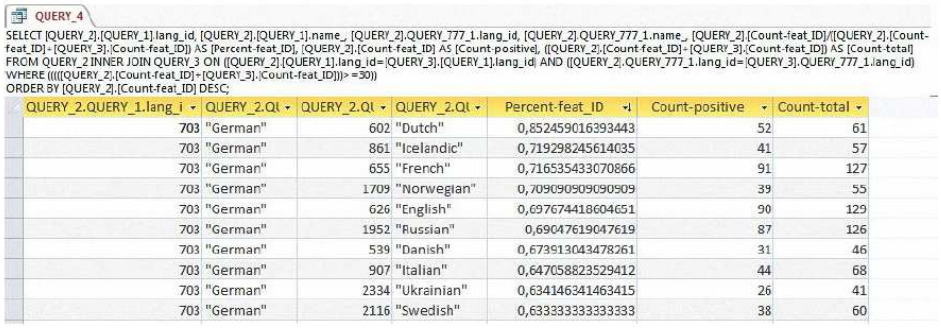
Fig. 2. QUERY-2

QUERY_3

```
SELECT [QUERY_1].lang_id, [QUERY_1].name_, QUERY_777_1.lang_id, QUERY_777_1.name_, Count([QUERY_1].feat_ID) AS [Count-feat_ID], Count([QUERY_1].feat_title) AS [Count-feat_title],
Count([QUERY_1].index_) AS [Count-index_]
FROM QUERY_1 INNER JOIN QUERY_1 AS QUERY_777_1 ON ([QUERY_1].feat_ID=QUERY_777_1.feat_ID) AND ([QUERY_1].index_<>QUERY_777_1.index_)
GROUP BY [QUERY_1].lang_id, [QUERY_1].name_, QUERY_777_1.lang_id, QUERY_777_1.name_
HAVING ((([QUERY_1].lang_id)=703))
ORDER BY Count([QUERY_1].feat_ID) DESC , Count([QUERY_1].feat_title), Count([QUERY_1].index_);
```

QUERY_1.lang_id	QUERY_1.na	QUERY_777_	QUERY_777_1.name_	Count-feat_	Count-feat_title	Count-index
703	"German"	2394	"Wari'"	85	85	85
703	"German"	1244	"Lango"	84	84	84
703	"German"	77	"Amele"	84	84	84
703	"German"	1412	"Maricopa"	83	83	83
703	"German"	2065	"Slave"	82	82	82
703	"German"	827	"Hixkaryana"	81	81	81
703	"German"	1122	"Koasati"	79	79	79
703	"German"	1846	"Pirahã"	79	79	79
703	"German"	36	"Ainu"	79	79	79
703	"German"	1695	"Nivkh"	79	79	79
703	"German"	929	"Jakaltek"	79	79	79
703	"German"	746	"Greenlandic (West)"	78	78	78
703	"German"	1459	"Meithei"	78	78	78
703	"German"	1178	"Krongo"	78	78	78
703	"German"	2556	"Zulu"	77	77	77
703	"German"	2370	"Vietnamese"	77	77	77
703	"German"	878	"Imonda"	76	76	76
703	"German"	1908	"Rapanui"	76	76	76
703	"German"	754	"Guarani"	76	76	76
703	"German"	736	"Goonlyandi"	76	76	76
703	"German"	1437	"Maybrat"	76	76	76
703	"German"	808	"Hausa"	76	76	76

Fig. 3. QUERY-3



QUERY_4

```
SELECT [QUERY_2].[QUERY_1].lang_id, [QUERY_2].[QUERY_1].name_, [QUERY_2].QUERY_777_1.lang_id, [QUERY_2].QUERY_777_1.name_, [QUERY_2].[Count-feat_ID]/[QUERY_2].[Count-
feat_ID]+[QUERY_3].[Count-feat_ID]) AS [Percent-feat_ID], [QUERY_2].[Count-feat_ID] AS [Count-positive], ([QUERY_2].[Count-feat_ID]+[QUERY_3].[Count-feat_ID]) AS [Count-total]
FROM QUERY_2 INNER JOIN QUERY_3 ON ([QUERY_2].[QUERY_1].lang_id=[QUERY_3].[QUERY_1].lang_id) AND ([QUERY_2].QUERY_777_1.lang_id=[QUERY_3].QUERY_777_1.lang_id)
WHERE ((([QUERY_2].[Count-feat_ID]+[QUERY_3].[Count-feat_ID]))> =30))
ORDER BY [QUERY_2].[Count-feat_ID] DESC;
```

QUERY_2.QUERY_1.lang_i	QUERY_2.QU	QUERY_2.QU	QUERY_2.QU	Percent-feat_ID	Count-positive	Count-total
703	"German"	602	"Dutch"	0,852459016393443	52	61
703	"German"	861	"Icelandic"	0,719298245614035	41	57
703	"German"	655	"French"	0,716535433070866	91	127
703	"German"	1709	"Norwegian"	0,709090909090909	39	55
703	"German"	626	"English"	0,697674418604651	90	129
703	"German"	1952	"Russian"	0,69047619047619	87	126
703	"German"	539	"Danish"	0,673913043478261	31	46
703	"German"	907	"Italian"	0,647058823529412	44	68
703	"German"	2334	"Ukrainian"	0,634146341463415	26	41
703	"German"	2116	"Swedish"	0,633333333333333	38	60

Fig. 4. QUERY-4

Table 1. Choice of parameters for the formal method of search for SMC-languages

Family	Language	Range of percentage match of values in features	Range of the number of common features[1]	Number of similar languages[2]
Altaic	Turkish	86.8-87.8%	38-41	2
	Turkmen3	-	-	-
	Tuvan	81.8-90.6%	32-51	7
Uralic	Estonian	76.7-79.2%	30-48	2
	Finnish	79.2-80.5%	41-48	2
	Saami (Northern)	76.7-80.5%	30-41	2
Indo-European	Serbian-Croatian	71.1-96.7%	30-49	7
	Spanish	71.1-83.1%	32-128	6
	Swedish	64.6-95.3%	31-65	10

Notes: 1) The admission was at the level of 30. When the value was lower, the result of the search was inconclusive.
2) Number of similar languages until the first non-relative language appeared.
3) For Turkmen WALS program does not contain enough grammar descriptions that satisfy the requirements of our method.

Choice of Parameters

The parameters of two-objective optimization (at least 30 features and 65% of matching values) have been defined above, but the method has not been specified. The first approach was empirical-a result of numerous experiments. Later, a formal method of defining the boundaries (30 features and 65%) will be suggested (cf. below).

The table contains data on 9 languages from 3 families of languages (Altaic, Uralic, Indo-European). The similarity of these languages is confirmed by phylogenetic tree on the base of lexical and phonetic data (Wichmann *et al*., 2013). Table 1 shows the choice of parameters (% of matching values and number of common features) for the formal method of search for SMC-languages.

Formal Description of the Two-Objective Optimization Method Used for Problem Solution

Problem of search for the Structurally Maximally Close Language (SMCL) *j* for any Language under Study (LUS) *i*

Let:

- $L = \{l_1, l_2 ..., l_i, ...l_n\}$ be a set of languages, be the percentage of matching feature values for languages *i* and *j*
- P_{ij}^{match} be the number of coinciding features for languages *i* and *j*

We have a language *i*. The problem is to find its maximally close language *j* that belongs to set *L* ($l_j \in L$). We will solve the problem of two-objective optimization (Ehrgott, 2000:17) by choosing one criteria as the main one similar to (Izraylevich and Tsudikman, 2012:79):

- We choose P_{ij}^{match} as the main criterion
- N_{ij}^{coinc} Then becomes the restricting criterion

Thus, the solution of the problem comes down to search for language j^{opt} that has the maximal percentage of matches P_{ij}^{match} with the given restriction N^{coinc_lim} for the number of coinciding features N_{ij}^{coinc} for the given language *i* and languages *j* (*j* = 1… *n*):

$$j^{opt} = \operatorname{argmax}\left(P_{ij}^{match}\right), N_{ij}^{match} \geq N^{coinc\_\lim}, j = 1...n$$

Where:

$$P_{ij}^{match} = \frac{N_{ij}^{match} * 100}{N_{ij}^{coinc}}$$

N_{ij}^{match} is the number of matching feature values of the languages.

Criticism of WALS Descriptions and Data Mining

The results received by the method described above seemed to be very encouraging, but still needed additional verification. Moreover, they were criticized by typologists for the fact that using the whole set of features from WALS in the two-objective optimization, we considered some non-structural features. (The information was anonymous, received as critical notes from one of the linguistic journals).

Thus, the verity of the data that can be received using the formal method of search for SMC-languages was checked on a so-called "Test A". Test A included all languages from WALS Programs that begin with the letter "A". Unfortunately, the results were unsatisfactory: SMC-languages belonged to the same family as the language under study only in 15 out of 35 cases, i.e., 42.9%. The data on the classification of a language with a certain language family was taken from WALS Online. The results of Test A are presented in Table 2.

Such results are believed to be accounted for by the fact that WALS does not have enough description for all languages. Besides, it includes such non-grammatical features as "lexicon", for instance, which can easily be borrowed by languages due to areal contacts, besides the features describing the vocabulary of the languages, there is an area "sign language" and values "other" and "not reported", which do not necessarily mean that these values actually match.

Table 2. Results of Search for SMC-languages for test "A"

Language under study	Family of language under study	SMC-language	Family of SMC-language	Number of coinciding features	% of matching values	Same family
Abipón	Guaicuruan	Achumawi	Hokan	32	71.90%	-
Abkhaz	Northwest Caucasian	Kolami	Dravidian	35	71.40%	-
Acehnese	Austronesian	Chrau	Austro-Asiatic	39	64.10%	-
Acehnese	Austronesian	Chrau	Austro-Asiatic	39	64.10%	-
Acoma	Keresan	Luiseño	Uto-Aztecan	34	64.70%	-
Akan	Niger-Congo	Irarutu	Austronesian	32	68.80%	-
Alamblak	Sepik	Kiwai	Kiwaian	38	76.30%	-
Alawa	Mangarrayi-Maran	Nunggubuyu	Gunwinyguan	38	84.20%	-
Albanian	Indo-European	Romanian	Indo-European	56	76.70%	+
Alyawarra	Pama-Nyungan	Kanuri	Saharan	32	78.10%	-
Amahuaca	Panoan	Shipibo-Konibo	Panoan	34	73.50%	+
Amele	Trans-New Guinea	Barai	Trans-New Guinea	33	90.90%	+
Amharic	Afro-Asiatic	Marathi	Indo-European	40	75.00%	-
Anejom	Austronesian	Tigak	Austronesian	34	70.60%	+
Angas	Afro-Asiatic	Yulu	Central Sudanic	33	81.80%	-
Ao	Sino-Tibetan	Usan	Trans-New Guinea	31	77.40%	-
Apalaí	Cariban	Hixkaryana	Cariban	43	76.70%	+
Apinayé	Macro-Ge	Canela-Krahô	Macro-Ge	30	90.00%	+
Apurinã	Arawakan	Seneca	Iroquoian	39	74.30%	-
Arabic (Egyptian)	Afro-Asiatic	Arabic (Moroccan)	Afro-Asiatic	39	84.60%	+
Araona	Tacanan	Tacana	Tacanan	38	71.10%	+
Arapesh	Torricelli	Tiwi	Tiwian	93	68.80%	-
Arawak	Arawakan	Baré	Arawakan	30	66.70%	+
Archi	Nakh-Daghestanian	Avar	Nakh-Daghestanian	42	85.70%	+
Armenian (Eastern)	Indo-European	Brahui	Dravidian	67	68.70%	-
Armenian (Western)	Indo-European	Lezgian	Nakh-Daghestanian	31	67.70%	-
Arosi	Austronesian	Indonesian	Austronesian	30	66.70%	+
Arrernte	Pama-Nyungan	Wahgi	Trans-New Guinea	32	78.10%	-
Asmat	Trans-New Guinea	Shiriana	Yanomam	35	74.30%	-
Atayal	Austronesian	Hawaiian	Austronesian	31	64.50%	+
Au	Torricelli	Bagirmi	Central Sudanic	30	70.00%	-
Avar	Nakh-Daghestanian	Tsova-Tush	Nakh-Daghestanian	33	93.90%	+
Awa Pit	Barbacoan	Mari (Meadow)	Uralic	35	71.40%	-
Awtuw	Sepik	Uradhi	Pama-Nyungan	35	80.00%	-
Aymara (Central)	Aymaran	Jaqaru	Aymaran	45	80.00%	+
Azerbaijani	Altaic	Turkish	Altaic	41	87.80%	+

Thus, it was decided to mark the following features and values as "insignificant" and exclude them from the queries:

- All features from the area "lexicon" (features 129-138)
- All features from the area "sign languages" (features 139-140)
- Values "other" and "not reported" from features 5 ("Voicing and Gaps in Plosive Systems"), 24 ("Locus of Marking in Possessive Noun Phrases"), 74 ("Situational Possibility"), 75 ("Epistemic Possibility"), 95 ("Relationship between the Order of Object and Verb and the Order of Adposition and Noun Phrase"), 96 ("Relationship between the Order of Object and Verb and the Order of Relative Clause and Noun"), 97 ("Relationship between the Order of Object and Verb and the Order of Adjective and Noun"), 142 ("Para-Linguistic Usages of Clicks")

Query-1 was changed so that only values that are not marked as "insignificant" are taken into consideration. Due to the smaller total number of values, it was decided to lower the minimal number of coinciding features to 26.

The total number of languages, for which an SMC language was found, equals 523. As for the other languages from WALS Program, they are either poorly described or the SMC-language had the percentage of matching features below the limit.

This procedure reminds the stage of data cleaning within the framework of the data mining theory, but, at the same time, it has the specificity of linguistic typology. One of the earliest works in the sphere of data mining is presented in (Fayyad *et al*., 1996). Thus, after the cleaning, it is believed that the structural features from WALS were kept for further research.

P-Pairs and G-Pairs of Languages

Comparative linguistics aimed at establishing the historical relatedness of languages compares their phonological and morphological systems, lexicon and syntax (Bynon, 1977), but sometimes it can only consider the lists of basic concepts-Swadesh lists (Swadesh, 1952).

A striking example of such method is Automated Similarity Judgment Program (ASJP) (Wichmann *et al*., 2013)-a project whose major goal is to classify all languages of the world applying computational approaches to comparative linguistics. Other purposes of the program include: Determining the homeland of a protolanguage, evaluating phylogenetic methods, investigating sound symbolism and a few more. Besides the universally recognized language families, the database includes language isolates, creoles, pidgins, mixed languages and constructed languages.

The similarity of languages is calculated automatically by using edit distance (Levenshtein, 1966). Originally, ASJP used 100-word lists, but later the authors came to a conclusion that shorter lists, containing only 40 words, gave the same accurate results. The work on the project is still going on and the number of languages (i.e., lists of words) is constantly being expanded.

The pairs of languages, whose genealogical similarity has been proven by comparative linguistics and ASJP and recognized by the linguistic society, shall be called P-pairs (Phonetic pairs). In order to avoid discrepancy due to different transcription systems and in order to embrace languages that do not have a writing system, Swadesh method, comparative linguistics and ASJP compare phonetic images of the words denoting the basic concepts (40 or 100 word lists); the reason why P-pairs of languages are referred to those pairs of genealogically related languages, whose similarity is stabled by ASJP method.

On the other hand, pairs of languages found by the method of two-objective optimization and data-mining described in this study present grammatically closest languages without considering lexicon at all, thus the term G-pairs (Grammatical pairs) of languages will be applied for them.

Pairs of languages that are recognized to be members of the same family and that present structurally closest languages will be called PG-languages.

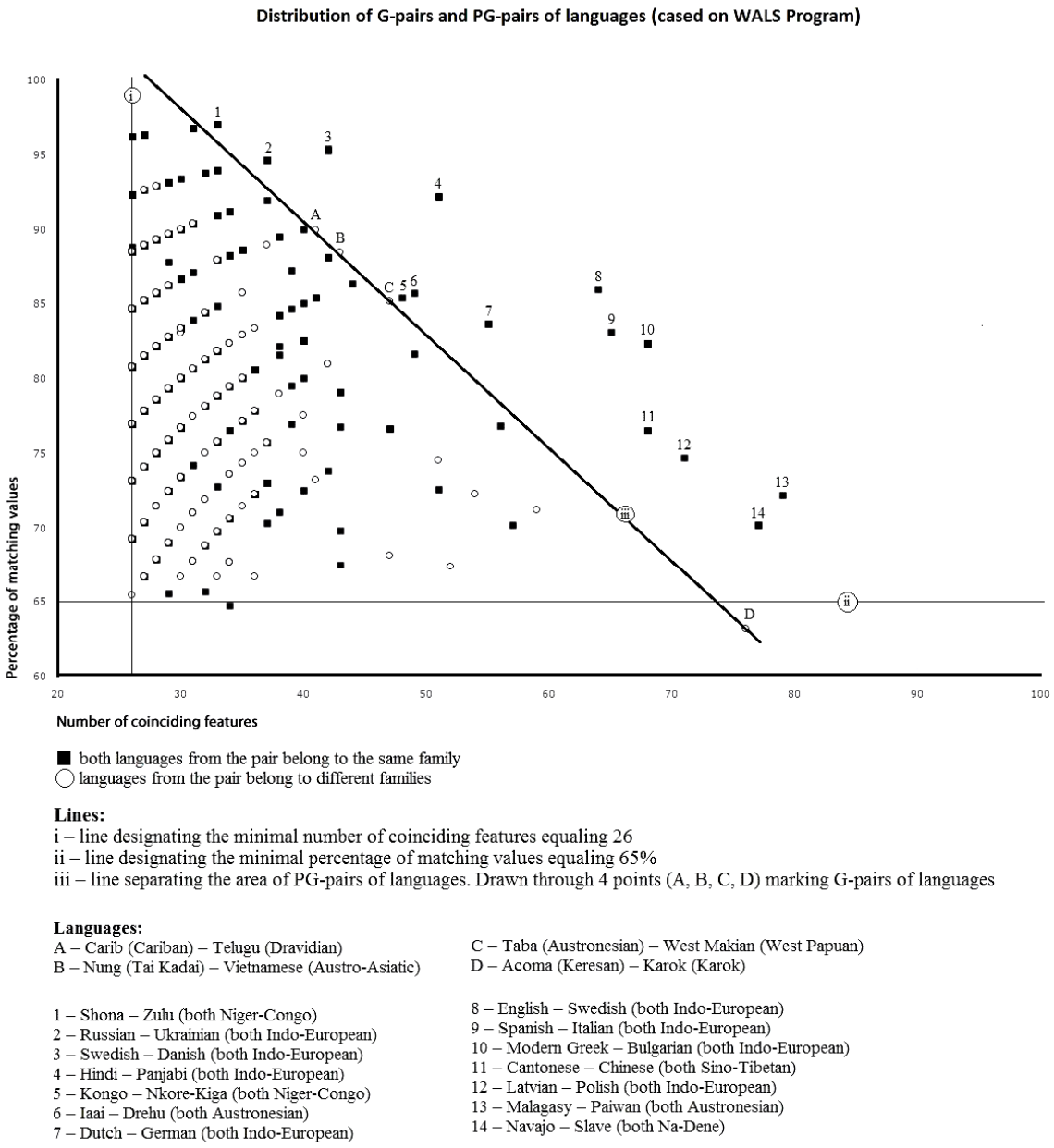
Figure 5 shows a diagram for all 523 languages from WALS Program, for which a SMC-language was found. G-pairs, where the language under study and its SMC language do not belong to the same family, are marked with white circles. Black squares mark PG-languages, i.e., G-pairs where languages belong to the same family. The diagram does not contain language isolates.

It can be seen that the more features of two languages coincide and the more values match, the higher the possibility that a G-pair is also a P-pair.

A line can be drawn through the four white circle. These are the four G-pairs of languages with the highest percentage of matching values. All languages above the line are PG-languages. As for the languages below the line, they can be either G-languages or PG-languages. In further studies, the view of this line can be specified due to the expansion of language descriptions.

We believe that the genealogic relatedness of languages proven by lexical data must also be reflected in their grammatical structure. Judging from the diagram, it stands true for the languages above the line.

As WALS is replenished with more information, this conclusion will become more evident.



■ both languages from the pair belong to the same family
◯ languages from the pair belong to different families

Lines:
i – line designating the minimal number of coinciding features equaling 26
ii – line designating the minimal percentage of matching values equaling 65%
iii – line separating the area of PG-pairs of languages. Drawn through 4 points (A, B, C, D) marking G-pairs of languages

Languages:
A – Carib (Cariban) – Telugu (Dravidian)
B – Nung (Tai Kadai) – Vietnamese (Austro-Asiatic)
C – Taba (Austronesian) – West Makian (West Papuan)
D – Acoma (Keresan) – Karok (Karok)

1 – Shona – Zulu (both Niger-Congo)
2 – Russian – Ukrainian (both Indo-European)
3 – Swedish – Danish (both Indo-European)
4 – Hindi – Panjabi (both Indo-European)
5 – Kongo – Nkore-Kiga (both Niger-Congo)
6 – Iaai – Drehu (both Austronesian)
7 – Dutch – German (both Indo-European)
8 – English – Swedish (both Indo-European)
9 – Spanish – Italian (both Indo-European)
10 – Modern Greek – Bulgarian (both Indo-European)
11 – Cantonese – Chinese (both Sino-Tibetan)
12 – Latvian – Polish (both Indo-European)
13 – Malagasy – Paiwan (both Austronesian)
14 – Navajo – Slave (both Na-Dene)

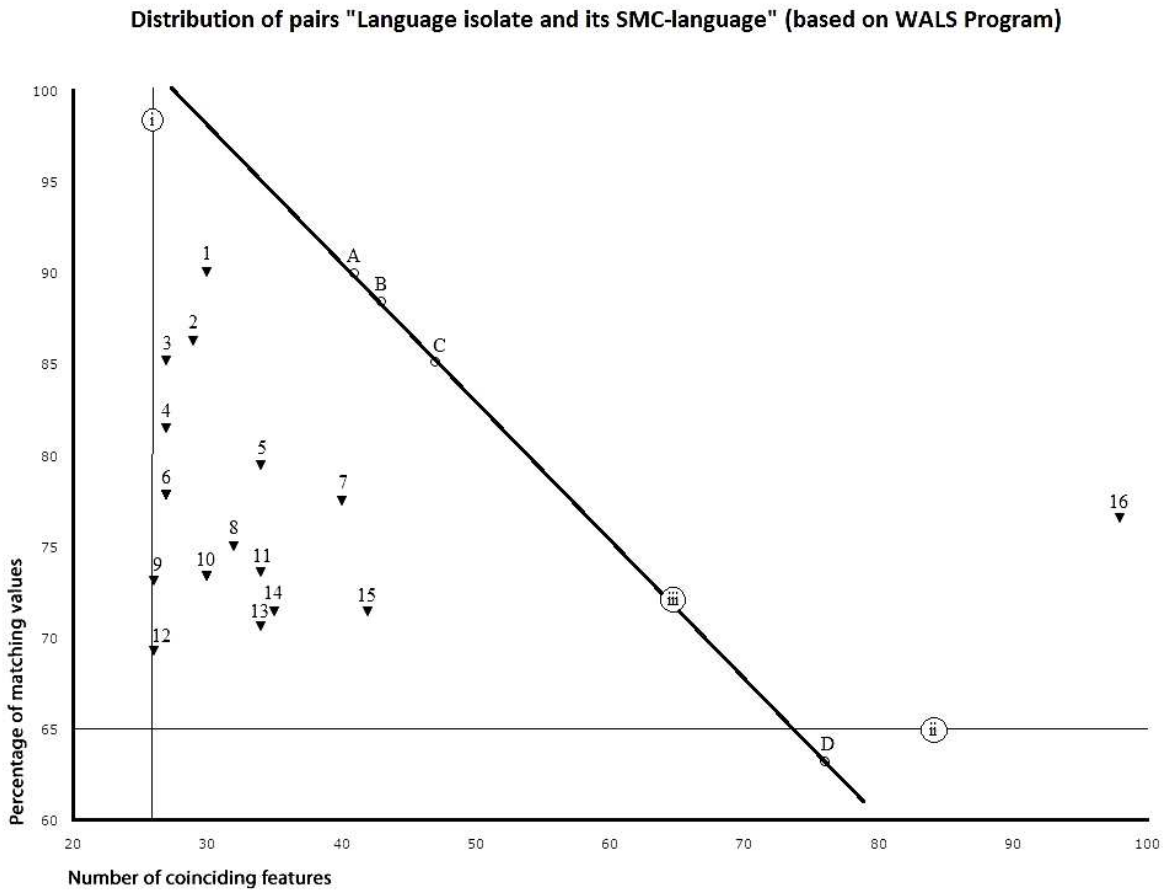
Fig. 5. Distribution of 523 pairs of SMC-languages from WALS PROGRAM

Language Isolates

WALS Program includes descriptions for 58 language isolates. In the present research, only those that allow getting results for search for SMC-languages according to the initial restrictions are considered, i.e., having at least 26 coinciding features and 65% of match. The results for 19 languages isolates are presented in Table 3.

The data from Table 3 were added to the diagram representing all G-pairs of languages form WALS, where only the line separating PG-pairs of languages from the others was left (cf. Fig. 6).

Unfortunately, almost all language isolates from the table above are below the line, which means that the structural description of the languages in question is not complete and the claims on the similarity of the language isolates under study and their structurally maximally close languages can only have a probabilistic nature. Table 3 depicting the table of language isolates is published in order to show the direction of further replenishment of WALS with structural information.



Lines:
i – line designating the minimal number of coinciding features equaling 26
ii – line designating the minimal percentage of matching values equaling 65%
iii – line separating the area of PG-pairs of languages. Drawn through 4 points (A, B, C, D) marking G-pairs of languages

Languages:
1 – Korean – Kalmyk (Altaic)
2 – Burushaski - Karachay-Balkar (Altaic)
3 – Sulka – Luvale (Niger-Congo)
4 – Warao – Fasu (Trans-New Guinea)
5 – Ainu – Kiwai (Kiwaian)
6 – Basque – Digaro (Sino-Tibetan)
Tonkawa – Fulniô (Yatê)
Tunica – Kunimaipa (Trans-New Guinea)
7 – Waorani – Suena (Trans-New Guinea)
8 – Nivkh – Koryak (Chukotko-Kamchatkan)
9 – Kuot – Luo (Eastern Sudanic)
10 – Chitimacha – Tunica (Tunica)
Huave – Luo (Eastern Sudanic)
11 – Siuslaw - Miwok (Southern Sierra) (Penutian)
12 – Yuchi – Manchu (Altaic)
13 – Washo – Daga (Dagan)
14 – Takelma – Daga (Dagan)
15 – Klamath - Nez Perce (Penutian)
16 – Tiwi – Maung (Iwaidjan)

Fig. 6. Diagram of distribution of pairs “Language isolate & its SMC-language” relatively to the area of PG-languages

Nevertheless, one pair of languages lies above the line-Tiwi and Maung. Thus, it can be claimed that these two languages are relative with Tiwi belonging to Iwaidjan family, where Maung is classified with.

Additionally, the pair of languages in question also meets the requirements suggested by Wichmann and Holman (2010). The authors compared profiles of languages presented in WALS pairwise and set the boundary of 45 coinciding features and 81.5% of matching values.

Despite the seeming similarity of the approaches, it should be stressed that in the present study, a descending line as a boundary between the area that includes only PG-languages and the area of both P-pairs and G-pairs of languages is suggested. It means that the more coinciding features a pair of languages has, the lower can the percentage of their matching values for the languages to appear in the area of PG-languages be.

Table 3. Query results for 19 language isolates

Language isolate (Language under study)	SMC-language	Number of matching features	Percentage of matching values in WALS program
Ainu	Kiwai	34	79.41%
Basque	Digaro	27	77.78%
Burushaski	Karachay-Balkar	29	86.21%
Chitimacha	Tunica	30	73.33%
Huave	Luo	30	73.33%
Klamath	Nez Perce	42	71.43%
Korean	Kalmyk	30	90%
Kuot	Luo	26	73.08%
Nivkh	Koryak	32	75%
Siuslaw	Miwok (Southern Sierra)	34	73.53%
Sulka	Luvale	27	85.19%
Takelma	Daga	35	71.43%
Tiwi	Maung	98	76.53%
Tonkawa	Fulniô	27	77.78%
Tunica	Kunimaipa	27	77.78%
Waorani	Suena	40	77.5%
Warao	Fasu	27	81.48%
Washo	Daga	34	70.59%
Yuchi	Manchu	26	69.23%

On the other hand, the whole set of phonetic information (40-item word lists of ASJP, inflexional affixes and endings, prepositions, etc.) does not present any prove of the similarity of these two languages. The question is: How Tiwi and Maung managed to preserve similarity in grammar, but not in phonetics. The possibility of their resemblance due to borrowings is discharged, as they are structurally maximally close languages.

Micro-Families

Micro-families is a term introduced by the authors of the present article. Micro-families are separate families of languages including up to 10 languages. Microfamilies are claimed to be similar to language isolates. The difference is that due to a number of evolutionary and historical reasons, the linguistic diversity of microfamilies is broader than that of language isolates. Another factor is their wider area of distribution and presence of dialects, which indicate more favorable historical conditions. Consequently, there is a reason for searching a relative family for micro-families like for language isolates.

In this study, 35 languages from 25 micro-families (the number of micro-families that have at least one language with enough features described) were studied. As a rule, there were 1-2 such languages in each family. These languages were taken as languages under study and a language from a different family and with the highest percentage of matching values was chosen as their SMC-language. The results are shown in Fig. 7.

Unfortunately, the grammar information in WALS is insufficient for a reliable conclusion on the genealogical relationship of micro-families (all pairs of languages on the diagram are below the line iii).

Results

The method of SMC-language gave hope for establishing the connection between a language isolate and its relative. This idea suggests the following hypotheses.

First, grammatical, or to be more precise, structural data, as well as phonetic data, can deliver information on the genealogic similarity of languages.

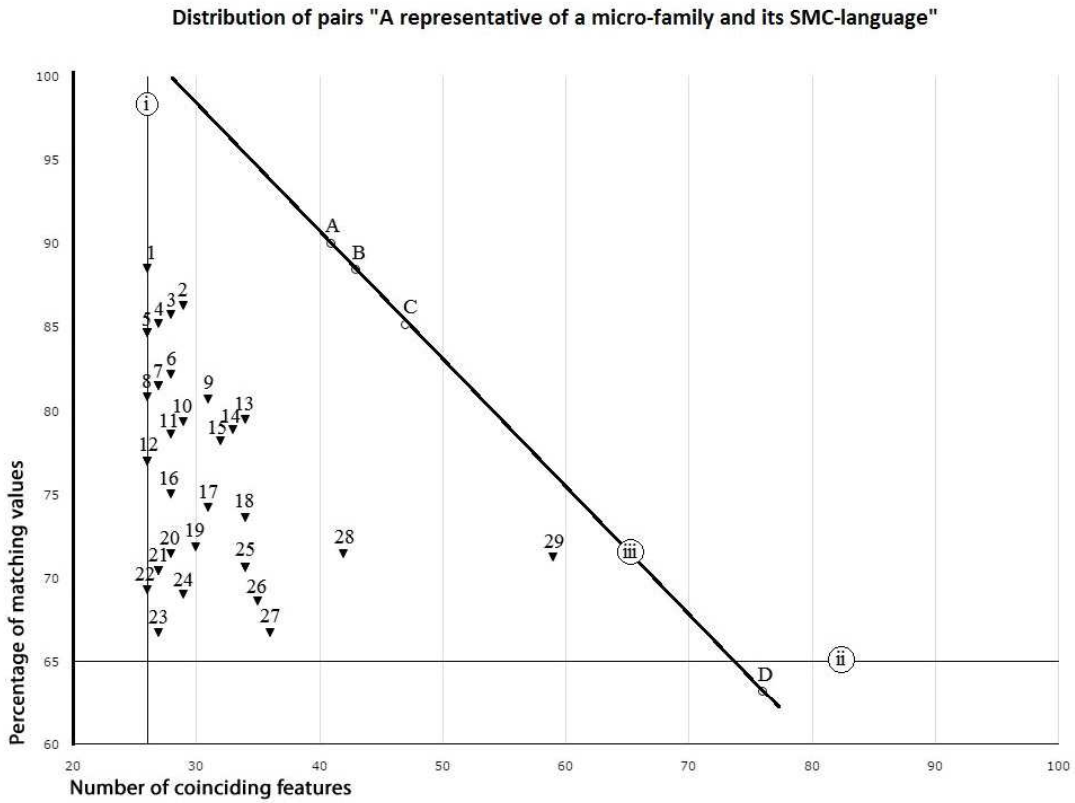
Second, language isolates and micro-families do not have genealogic relatives based on phonetic information (P-pairs), but they do have genealogic relatives based on grammar information (G-pairs).

Third, the results for the language isolate Tiwi and its SMC-language Maung from Iwaidjan micro-family allowed us suggesting a hypothesis that these two languages are genealogically related.

Discussion

It can be concluded that grammar information is an important source of data about the relationship of languages, as phonology, morphology and lexicon.

WALS is the world's biggest typological database, whose first version-WALS Program-was used as the basis for the present study. We used all the information presented in WALS, which is at the present time the most complete collection of typological data on the languages. Realizing that these data are still not enough for reliable theories, we suggest hypotheses hoping that further replenishment of WALS will prove them. Now WALS contains the description of 2,560 languages. But only 523 of these languages (about 20.4%) have structure description sufficient for communication of information about genealogic relationship.



Lines:
i – line designating the minimal number of coinciding features equaling 26
ii – line designating the minimal percentage of matching values equaling 65%
iii – line separating the area of PG-pairs of languages. Drawn through 4 points (A, B, C, D) marking G-pairs of languages

Languages:
A – Carib (Cariban) – Telugu (Dravidian)
B – Nung (Tai Kadai) – Vietnamese (Austro-Asiatic)
C – Taba (Austronesian) – West Makian (West Papuan)
D – Acoma (Keresan) – Karok (Karok)

1 – Cayapa (Barbacoan) – Newari (Sino-Tibetan)
2 – Yukagir (Kolyma) (Yukagir) – Tulu (Dravidian)
3 – Choctaw (Muskogean) – Biloxi (Siouan)
4 – Sanuma (Yanomam) – Kaingang (Macro-Ge)
5 – Jivaro (Jivaroan) - Maidu (Northeast) (Penutian)
Huitoto (Minica) (Huitotoan) – Chuvash (Altaic)
Shiriana (Yanomam) – Daga (Dagan)
6 – Ket (Yeniseian) – Tulu (Dravidian)
7 – Mohawk (Iroquoian) – Wichita (Caddoan)
Qawasqar (Alacalufan) – Wappo (Wappo-Yukian)
8 – Awa Pit (Barbacoan) – Lamani (Indo-European)
9 – Lavukaleve (Solomons East Papuan) – Kunimaipa (Trans-New Guinea)
10 – Wichi (Matacoan) – Cornish (Indo-European)
11 – Kwakw'ala (Wakashan) – Shuswap (Salishan)
12 – Epena Pedee (Choco) – Kaingang (Macro-Ge)
Miskito (Misumalpan) – Dogon (Dogon)
13 – Georgian (Kartvelian) - Tsova-Tush (Nakh-Daghestanian)
14 – Huitoto (Murui) (Huitotoan) – Khalkha (Altaic)
15 – Aymara (Aymaran) – Tamang (Sino-Tibetan)
16 – Pirahã (Mura) – Fasu (Trans-New Guinea)
17 – Koryak (Chukotko-Kamchatkan) – Khanty (Uralic)
18 – Kiowa (Kiowa-Tanoan) – Kiwai (Kiwaian)
19 – Cherokee (Iroquoian) – Amahuaca (Panoan)
20 – Itelmen (Chukotko-Kamchatkan) – Nasioi (East Bougainville)
21 – Koasati (Muskogean) – Kamaiurá (Tupian)
Jebero (Cahuapanan) – Brahui (Dravidian)
22 – Wichita (Caddoan) – Yanesha' (Arawakan)
23 - Wari' (Chapacura-Wanham) – Woleaian (Austronesian)
24 – Gooniyandi (Bunuban) – Waray (in Australia) (Gunwinyguan)
25 – Chukchi (Chukotko-Kamchatkan) – Selepet (Trans-New Guinea)
Seneca (Iroquoian) – Apurinã (Arawakan)
26 – Nuuchahnulth (Wakashan) – Squamish (Salishan)
27 – Selknam (Chon) – Qawasqar (Alacalufan)
28 – Jaqaru (Aymaran) – Quechua (Imbabura) (Quechuan)
29 – Abipón (Guaicuruan) – Wichi (Matacoan)

Fig. 7. Distribution of pairs "A representative of a micro-family & its SMC-language" relatively to the area of PG-languages, Note: If one number designates over one pair, it means that the number of coinciding features and the percentage of matching values were the same for these pairs of languages.

Therefore, it is necessary to replenish WALS with more information about languages, will which make the database even more important and useful tool for further studies.

Optimization methods are widely used in all spheres; they are diverse and acknowledged to be part of computer science (Ehrgott, 2000). In linguistic typology the method of optimization was used for the first time. We decided on the variant of two-objective optimization as it best suits for the nature of the data. As for statistics, we believe that the method of optimization seems more

precise, though we would like to state that we are not criticizing mathematics statistics in any way.

We used a SQL-like languages (language of MS Access), as SQL is de facto the standard in the sphere of informational databases (Codd, 1970).

The article contains all queries for the database "Isolates" in order to make them easier to reproduce in case the queries (Fig. 1-4), the method of two objective optimization and the graphs (Fig. 5-7) need additional verification, as well as the conclusions. Moreover, the queries can be reproduced using any relational DBMS. The database "Isolates" is a slightly changed fragment of WALS Program. All data from WALS Program (Haspelmath *et al*., 2005) and WALS Online (Dryer and Haspelmath, 2013) are available in free access.

One of the goals of the present research was an attempt to classify language isolates using a new formal method, based on the structural information about languages. A similar attempt, though proposing a different aim, was made by Wichmann and Holman (2010) "Pairwise Comparison of Typological Profiles". The authors of the article established the boundaries separating the area of languages belonging to the same families from the area that does not guarantee a necessary genealogic relationship of languages to be at least 45 coinciding features and 81.5% of matching values.

In contrast to their research, we believe that these numbers can be lowered provided that the comparison of typological profiles excludes non-structural features, as well as values that do not necessarily mean a match, e.g., value "other". The minimal number of coinciding features can be 26 and the lowest percentage of matching values is 65%. Moreover, the line indicating the border between the pairs of languages that belong to the same family and the area that includes both G-pairs and P-pairs of languages should be descending. It is conditioned by the fact that the increase of the number of coinciding features decreases the probability of accidental match of a set of values.

The positive result of search of SMC-languages for language isolates included only one pair of languages-Tiwi (a language isolate) and Maung (an Iwaidjan language). The other 18 pairs of languages fell below the boundary of the area of PG-languages. This can be accounted for by the fact that only 20.4% of all languages from WALS Program contain enough information for this study. Thus, once WALS supplements the other language profiles with more data, the application of the formal method for language isolates, as well as micro-families, will give more remarkable results.

We believe that there will be no dramatic change in the percentage of matching values for language isolates and their SMC-languages if WALS is replenished with more descriptions, but the G-pairs of languages will move further to the right of the diagram falling into the area of guaranteed similarity.

For many decades, scientists have been trying to find genealogic relations for language isolates using the whole amount of phonetic information. Hereby, not only limited lists of basic vocabulary, but also the accompanying information (case system (prepositions, endings, postposition, word-forming and inflectional affixes, conjunctions, pronouns, toponyms, onomasticon, names of animals and plants, etc.) is meant. However, there were no results in this sphere.

We showed that both authors (Benveniste and Trubetzkoy), whom we highly respect, were wrong in their own way. Their misbelief can be explained by the poor elaboration of grammar theory in the 20th century. Trubetzkoy did not have enough data to make conclusions on the similarity of Indo-European languages. Even at the present time genealogic features of a family are unlikely to be reliably defined by the frequency method (Makarova and Polyakov, 2015; Danilova *et al*., 2016). It is only possible to define the closest grammatical relative for a language. The theory and practice of grammar statistics still has to find a convincing answer for this question.

Benveniste was wrong in denying the possibility of classifying languages according to the genealogical principle basing on grammar data. It becomes more obvious that such genealogical classification can exist providing that there is enough structural data on the languages.

Emile Benveniste was right to claim that Takelma does not belong to the Indo-European languages. In the present research (Table 3) we show that the most probable genealogical relative (SMCL) for Takelma (LUS) is Daga from the Dagan family of languages (spoken in Papua New Guinea).

Although the formal method of search of SMC languages can help find genealogic relations between a pair of languages, it does not answer the question on the type of their connection-whether a SMC-language is parent or daughter of the language under study, or they both descended from a common ancestor (not described in WALS or lost historically).

In order to answer the question above we should study the historical context, analyze the possible direction of settlement on continents and archipelagos, compare the areas of language under study and its SMC language by the density of its populating by ethnically close people.

It is noteworthy, that the historical space of the language existence is, as a rule, not connected with the civilization level of development of its speakers. Languages under study have lived in deep geographic isolation for hundreds (or even thousands) of years. So it would be wrong to consider the language of a more civilized people parent and the language of a less civilized people-descendent and vice versa.

Probably, the best way to define the type of relation between a language under study and its SMC-language is to combine formal method suggested in the present paper and methods of phylogeny.

Conclusion

The present research is based on the information from the biggest typological database describing structural properties of languages-WALS. The data were processed using a formal, new for typological studies, method-a method of two-objective optimization.

We introduced the terms: P-pairs of languages (whose genealogic relationship was defined by the methods of comparative linguistics or similar methods based on phonetic data), G-pairs of languages (the relationship was defined by the method of search for SMC-language using structural information from WALS) and PG-pairs (cases when P-pairs and G-pairs match).

A diagram for all languages from WALS that meet the requirements of two-objective optimization was built. It included 523 pairs if languages. We were able to find draw a line through four points separating the area, which includes only genealogically relative languages (PG-pairs), from the area of both P-pairs and G-pairs of languages.

This numerical experiment showed that only 20.4% of all languages presented in WALS have enough grammar description for being able to classify with their relatives basing on structural information. An important conclusion from this experiment is that grammar features carry important information on the genealogical similarity of the languages.

The formal method aimed at finding SMCL for any language was applied to 19 language isolates that contain enough information in WALS Program. The distribution diagram of pairs “Language isolate & its SMCL” (Fig. 6) showed that 18 pairs of languages fell below the boundary separating the area of PG-languages, which mean that no conclusion can be made about these languages. Nevertheless, Tiwi and its SMC-language-Maung-fell into the area of genealogically relative languages. It allows us suggest the hypothesis that Tiwi belongs to the Iwaidjan family along with Maung.

The research was also conducted on 35 languages from 25 micro-families that include up to 10 languages and, in our opinion, are similar to language isolates. A SMCL was found for each of the language from microfamilies. The result presented in a diagram showed that no pairs fell in the area of PG-pairs (Fig. 7). That is the reason why we only suggest a hypothesis that if WALS is replenished with more information pairs of languages from micro-families and their SMC-languages will move to the right on the graph.

Acknowledgment

The paper was checked for the language by Professional editing service (http://www.profediting.com/).

Online Chart Maker (https://live.amcharts.com/) was used for graph visualization.

Funding Information

The research was supported by RFBR grant No. 16-06-00187.

The study was conducted as part of scientific-research work #74 within the framework of the basic part of governmental order in the sphere of scientific activity by order #2014/113.

Author’s Contributions

The idea of the study, discussion and conclusion sections and formal description of the method belong to **Vladimir Polyakov**. The database “Isolates” and the queries were developed by **Vladimir Polyakov and Ivan Anisimov**. The graphs, calculations and tables were made by **Elena Makarova**. The authors have equally contributed to the writing of the article.

Ethics

This article is original and contains unpublished material. The authors confirm that there are no ethical issues involved.

References

Benveniste, E., 1954. La classification des langues. Conferences de l'Institut de Linguistique de l'Universite de Paris, XI, Annees 1952-1953: 33-50.

Bynon, T., 1977. Historical Linguistics. 1st Edn., Cambridge University Press, Cambridge, ISBN-10: 0521291887, pp: 301.

Codd, E.F., 1970. A relational model of data for large shared data banks. Commun. ACM, 13: 377-87. DOI: 10.1145/362384.362685

Comrie, B., M.S. Dryer, D. Gil and M. Haspelmath, 2013. Introduction. In: The World Atlas of Language Structures Online. Dryer, M.S. and M. Haspelmath (Eds.), Max Planck Institute for Evolutionary Anthropology, Leipzig.

Danilova, V., E. Makarova, V. Polyakov and V. Solovyev, 2016. Frequency-based relevant grammar features of the Caucasian languages. Ind. J. Sci. Technol., 9: 1-10. DOI: 10.17485/ijst/2016/v9i11/89415

Dryer, M.S. and M. Haspelmath, 2013. The world atlas of language structures online. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Edgeworth, F.Y., 1881. Mathematical Psychics. 1st Edn., P. Keagan, London.

Ehrgott, M., 2000. Multicriteria Optimization. 1st Edn., Springer, Berlin, ISBN-10: 3540678697, pp: 243.

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth, 1996. From data mining to knowledge discovery in databases. Ai Magazine, 17: 37-54. DOI: 10.1609/aimag.v17i3.1230

Gray, R.D., Q.D. Atkinson and S.J. Greenhill, 2011. Language evolution and human history: What a difference a date makes. Philosophical Trans. Royal Society B: Biol. Sci., 366: 1090-1100. DOI: 10.1098/rstb.2010.0378

Haspelmath, M., M.S. Dryer, D. Gil and B. Comrie, 2005. The World Atlas of Language Structures. 1st Edn., OUP Oxford, Oxford, ISBN-10: 0199255911, pp: 712.

Izraylevich, S. and V. Tsudikman, 2012. Automated Option Trading: Create, Optimize and Test Automated Trading Systems. 1st Edn., FT Press, Upper Saddle River, ISBN-10: 0132491907, pp: 304.

Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions and reversals. Soviet Phys. Doklady, 10: 707-710.

Makarova, E. and V. Polyakov, 2015. The origin of the article in Indo-European languages of Western Europe. Mediterranean J. Social Sci., 6: 61-75. DOI: 10.5901/mjss.2015.v6n5s4p61

Swadesh, M., 1952. Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos. Proc. Am. Philosophical Society, 96: 452-463.

Tallerman, M. and K.R. Gibson, 2012. The Oxford Handbook of Language Evolution. OUP Oxford, Oxford, ISBN-10: 0199541116, pp: 763.

Trubetzkoy, N.S., 1939. Gedenken ihrer das Indogermanen Problem, Acta Linguistica. Copenhague, 1: 81-89.

Wichmann, S. and E.W. Holman, 2010. Pairwise Comparisons of Typological Profiles. In: Rethinking Universals: How Rarities Affect Linguistic Theory, Wohlgemuth, J. and M. Cysouw (Eds.), Walter de Gruyter, Berlin, ISBN-10: 311022092X, pp: 241-254.

Wichmann, S., A. Müller, A. Wett, V. Velupillai and J. Bischoffberger *et al*., 2013. The ASJP Database (version 16).