COMPARISON OF VARIOUS QUANTITATIVE MEASURES OF PROXIMITY OF LANGUAGES: NORTH CAUCASIAN LANGUAGES

Galeev Timur Ildarovich
Kazan federal university (KFU), Kazan, Russia
tigaleev@kpfu.ru

Solovyev Valery Dmitrievich
Kazan federal university (KFU), Kazan, Russia

ABSTRACT

A comparison of North Caucasian languages is performed in the article according to various measures of proximity constructed using grammatical, lexical and genetic databases. Statistical methods are applied to the study of correlations among these proximity measures, and also between them and both geographical proximity and genealogical kinship. A full correlation has been found among language kinship, geographic situation and genetic kinship of peoples. Also, a high correlation was found between each of them and lexical similarity. In general these correlations persist at different levels – starting at the whole set of studied languages until the level of the smallest groups of related languages. It is shown that a significant factor in the analysis of geographical situation is the existence of a common boundary between the regions of distribution of languages.

Keywords: *Various Quantitative Measures, geographical situation common boundary, Creativity, languages*

1. INTRODUCTION

The classification of languages by genetic kinship, developed in the last two centuries within the framework of historical linguistics applying the comparative historical method, offers a qualitative characteristic of language proximity by including them into macrofamilies, families, branches, groups, etc. Glottochronology provides a quantitative measure of proximity that, in particular, allows assessing the age of families and other language groups.

Unfortunately, in many cases there is no consensus among experts about languages kinship; it must be said also that lexicostatistical data are controversial. The comparative historical method and glottochronology rely generally on lexical material, and apply to it the laws of phonetic changes. This suggests the idea of expanding the set of data under consideration. In the present article grammatical data are considered together with lexical data.

Loanwords are a common problem for both lexical and grammatical data. For the moment, there is no rigorous theory of loanwords. As a result, it is often difficult to determine which are the reasons for the proximity of languages – either a common origin or the presence of loanwords. Since loanwords may appear as a consequence of contacts between neighbouring peoples, it makes sense to consider also the geographical distance as a quantitative parameter. In this research a factor of the existence of a common boundary between distribution areas of languages will be taken into account. Finally, genetic data are available for some ethnicities, making possible to consider also genetic distance between them.

In earlier publications [Cavalli-Sforza 2000, Knight *et al.* 2003, Lansing *et al.* 2007, Hunley *et al.* 2008, Nasidze *et al.* 2001, Limborskaya *et al.* 2002, Derenko *et al.* 2006] genetic and linguistic data were compared. The results obtained are heterogeneous; the existence of correlations depends on the region, the studied set of haplogroups and other parameters. Furthermore, a systematic comparison of proximity

Submit Date: 12.02.2017, Acceptance Date: 05.03.2017, DOI NO: 10.7456/1070ASE/017

measures of languages (and peoples) has never been performed before simultaneously for all available types of data (genealogical, grammatical, lexical, genetic and geographical).

This article is devoted to the comparison of the data types listed above in the case of North Caucasian languages. The choice of this group of languages can be justified by the following reasons. First of all, genealogical classification of these languages is reliably established and rather detailed. Next, Caucasian ethnic groups lived throughout many centuries in fixed places. In this respect they contrast sharply, for instance, with Turkic peoples, with their frequent offensive campaigns and blending of ethnicities. The comparison of various data types should reveal to what extent genes and language depend on kinship and to what extent on contacts, what is easier borrowed – genes, lexicon or grammar. Ultimately, this will allow a more detailed description of the history of peoples' development.

This article is structured as follows. In section 2 a short review of results on co-evolution of languages and genes in the Caucasian region is offered. In section 3 the data used in our research are described. The statistical methods applied are depicted in section 4 together with the numerical data obtained. Section 5 contains the analysis of numerical data at different language levels; here the main results are stated. A method to assess possible geographical position of peoples in the past is proposed in section 6. Finally, the results obtained are discussed in section 7. Appendices contain all the data used in the paper.

2. CO-EVOLUTION OF LANGUAGES AND GENES

The general mechanism of divergent development of populations with resettlement and isolation leads to similar mechanisms of evolution for languages and genes. This question was considered for the first time by Cavalli-Sforza [Cavalli-Sforza 2000]. He constructed a widely cited scheme (Fig. 1) that clearly demonstrates correlation between language families and large genetically different populations. Afterwards, this work was repeatedly exposed to criticism, in particular, for comparing linguistic groups whose existence has not been confirmed by generally recognized methods and is just hypothetical. Nevertheless, this work was stimulating and inspiring for many other researchers.

During the years passed after that publication a huge amount of data, both genetic and linguistic, has appeared, that can significantly clarify this correlation and consider it at a more detailed level. A considerable attention has been paid in many works to the study of the genetic structure of populations of the Caucasus region.

I. Nasidze and colleagues addressed the following question: what is the best predictor of genetic proximity in the Caucasus – geography or language? The proximity of languages was understood as their belonging to one and the same language group. All main parts of the genome were studied, namely mitochondrial DNA, Y-chromosome, Alu insertion loci of autosomal DNA. The main result obtained is that Azerbaijanians and Armenians are genetically closer to the neighbouring populations of the Caucasus rather than to the peoples of their respective linguistic groups, namely Turkic and Indo-European. No significant correlation between genetic proximity and geographical distance has been detected [Nasidze *et al.* 2001].

The genetic structure of Western Caucasus peoples was studied in [Litvinov 2010]. A statistically significant correlation between Y-chromosome genetic distances and geographical distances, and also linguistic distances was established. At the same time, no correlation was detected between geographical distances and either mitochondrial DNA genetic distances or autosomal Alu insertion genetic distances.

Submit Date: 12.02.2017, Acceptance Date: 05.03.2017, DOI NO: 10.7456/1070ASE/017

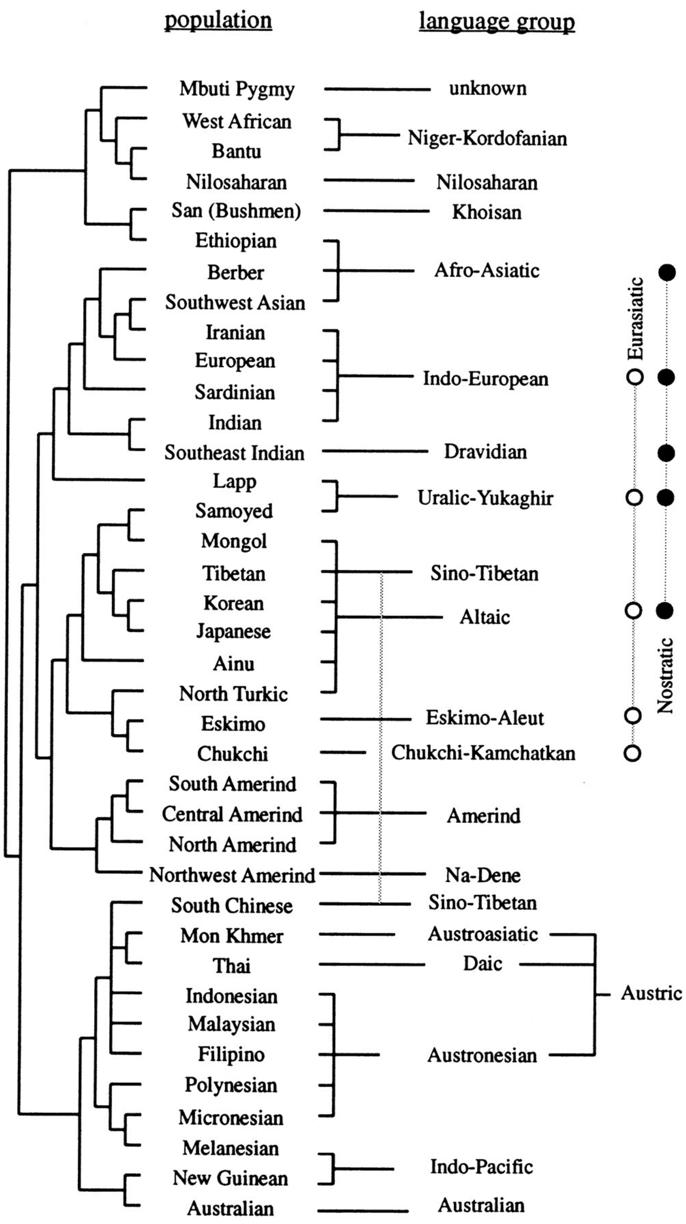


Fig. 1. Correlation between language families and populations.

B. Yunusbayev studied the polymorphism of Alu insertions and Y-chromosomes for the following peoples of the Caucasus: Avars, Bagvalals, Chamalals, Andis, Lezgians, Dargins, Tabasarans, Kumyks, Karanogais, Karachays, Abazins, Adyghe, Balkars [Yunusbayev 2006]. The research provides data on specific gaplogroups frequencies for these peoples, and also for a number of other peoples of Europe and Asia, for the purpose of comparative analysis and identification of the place occupied by Dagestan peoples in the genetic map of the world. A statistically significant correlation of genetic and geographical distances for Alu insertion loci was detected. No comparison with linguistic distances was carried out in the aforementioned dissertation.

Submit Date: 12.02.2017, Acceptance Date: 05.03.2017, DOI NO: 10.7456/1070ASE/017

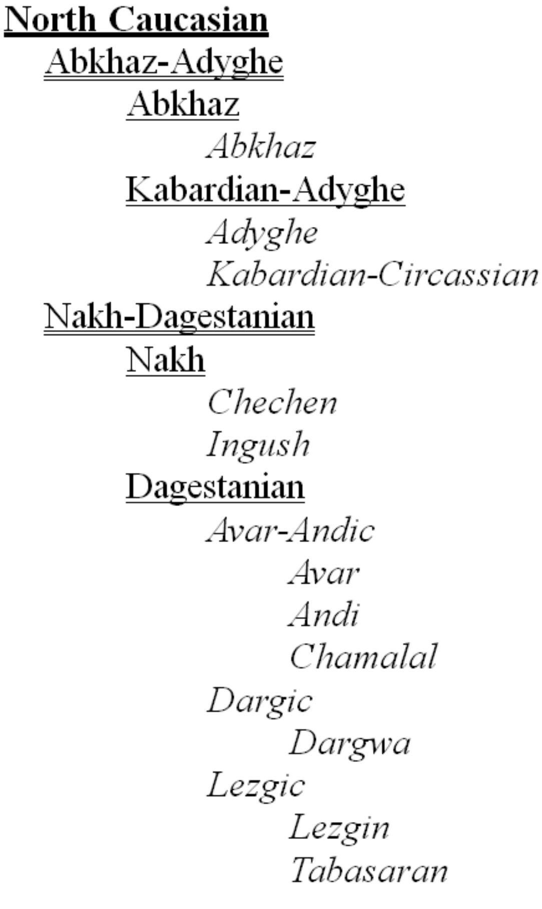
A combined approach using all types of genetic information was applied by I. Kutuev [Kutuev 2010]. In the space of the two main components of the frequency of Y-chromosome haplogroups, the East and West Caucasian clusters are clearly discerned along with the isolated position of Nakh peoples. When considering the Alu insertions of Europe and Asia populations, Caucasus peoples are evidently distinguished in one single cluster; Chamalals stand apart. When examining the mitochondrial DNA, Caucasian peoples fall into the Western Eurasian cluster.

Kutuev detected a correlation between geographical and genetic distances for both Y-chromosome and mitochondrial DNA, but not for autosomal DNA.

Thus, the data provided in different works have controversial character. Let us note that in those works on Caucasus peoples in which correlation of genetics with linguistics is considered, linguistic proximity is deemed only in the sense of belonging to the same genealogical group of languages. There is no differential assessment of grammatical, lexical proximity.

3. DATA SOURCES

The genealogical classification of North Caucasian languages proposed in [Burlak, Starostin 2001] is presented in Fig. 2. It is given in other sources almost in the same form [Grimes 2000].



- North Caucasian
  - Abkhaz-Adyghe
    - Abkhaz
      - *Abkhaz*
    - Kabardian-Adyghe
      - *Adyghe*
      - *Kabardian-Circassian*
  - Nakh-Dagestanian
    - Nakh
      - *Chechen*
      - *Ingush*
    - Dagestanian
      - *Avar-Andic*
        - *Avar*
        - *Andi*
        - *Chamalal*
      - *Dargic*
        - *Dargwa*
      - *Lezgic*
        - *Lezgin*
        - *Tabasaran*

Fig. 2. Genealogical hierarchy of languages

Proximity of languages in this tree is presented as a matrix of distances, so that it can be processed by statistical methods. Each vertex of the tree in Fig. 2 has been assigned a rank: 0 to leaves, any of the other vertices has a rank that is a unit greater than the maximum rank of its descendants. Consequently, distances between languages in the genealogical tree are defined as the rank of their nearest common ancestor.

Lexical data were taken from the ASJP [Müller *et al.* 2010]. The data allow to determine the lexical proximity of languages. The distance between two words is defined as the number of transformations that are necessary to transform one word into another. The distance between languages is defined as the normalized sum of the distances between corresponding words. The ASJP approach differs from that of classical lexicostatistics in the following two aspects: in ASJP two words designating one and the same concept are compared independently of whether they are cognates. Thus, part of the information accumulated by historical linguistics is lost here. But on the other side, the comparative historical method does not take into account in any way the degree of similarity of cognates.

Grammar data are taken from the 'Languages of the World' database created on the basis of materials from the homonymous series of monographs (www.dblang2008.narod.ru). It contains descriptions of 315 languages of Eurasia according to 3821 criteria related to the following aspects of language: phonetics, morphology, syntax [Polyakov, Solovyev 2006].

Genetic data were provided by O. Balanovsky on the basis of [Balanovsky 2010].

Geographical coordinates of the distribution areas of languages were taken from the Atlas of the World's Languages in Danger [Moseley 2010] and WALS [Haspelmath *et al.* 2005]. Distances between these coordinates were calculated conforming to Vincenty's formulae, which are the most accurate formulae taking into account the Earth's shape as a geoid.

The following languages were chosen for comparison: Avar, Andi, Chamalal, Dargwa, Tabasaran, Lezgin, Chechen, Ingush, Abkhaz, Adyghe, Kabardian-Circassian.

4. "LOCAL" ANALYSIS

In the genealogical structure of Fig. 1, the following levels can be naturally distinguished: individual languages, branches and families. At the first level we pick out two sets of languages from different families: {Avar, Andi, Chamalal} and {Abkhaz, Adyghe, Kabardian-Circassian}. At the second level – one set of sub-branches: {Avar-Andic languages, Dargi language, Lezgi languages} of the Dagestanian branch. At the third level, two branches of the upper level of the Nakh-Dagestanian family are considered together with the Abkhaz-Adyghe family – Nakh and Dagestanian.

All the data from different tables for each of the considered groups of languages will be merged into one single table. Each cell of the table is divided into four sub-cells containing the distances: lexical in the upper left cell, grammatical in the upper right cell, genetic in the lower left cell, geographical in the lower right cell. In addition to the distance in kilometres, also the existence of a boundary between the corresponding regions is registered.

Some of the most representative examples were selected for analysis from different families, different language levels and reflecting different situations concerning correspondence between genealogical classification and geographical distance.

Submit Date: 12.02.2017, Acceptance Date: 05.03.2017, DOI NO: 10.7456/1070ASE/017

4.1. Language level

4.1.1. Avar, Andi, Chamalal.

Distances between these languages are contained in Tab. 2.

Lexical distances precisely reflect the standard genealogical classification, according to which both Chamalal and Andi languages are included into the Andic subgroup. At the same time, grammatical distances put closer Andi and Avarian languages, while genetic data put closer Avar and Chamalal peoples. This may be consequence of contact phenomena: grammatical borrowings and exchange of genes. Let us pay attention that Andi language is closer to Chamalal geographically, but they do not share a common boundary. The region inhabited by Avars has a common boundary with those populated by Andi and Chamalals.

Tab. 2. Distances between languages of Avar-Andic subgroup

	Avar		Andi		Chamalal	
Avar			77,46	115	82,61	120
			0,21	50/+	0,04	61/+
Andi					54,93	137
					0,26	25/-
Chamalal						

Thus, data on these languages suggest that grammar and genes are easier transferred than lexicon, and a common boundary is an important factor for borrowings.

4.1.2. Abkhaz, Adyghe, Kabardian-Circassian

Distances between languages are provided in Tab. 3.

Table 3. Distances between languages of Abkhaz-Adyghe family

	Abkhaz		Adyghe		Kabardian-Circassian	
Abkhaz			93,49	206	95,96	214
			0,04	217/-	0,08	220/-
Adyghe					32,69	174
					0,17	321/-
Kabardian-Circassian						

Lexical and grammatical distances are in exact correspondence with each other and with the adopted genealogical classification of languages, according to which Adyghe and Kabardian-Circassian belong to the same subgroup. As for geographical distances, in this case they are rather conditional as a result of the extremely low concentration of native speakers of Adyghe and Kabardian-Circassian languages.

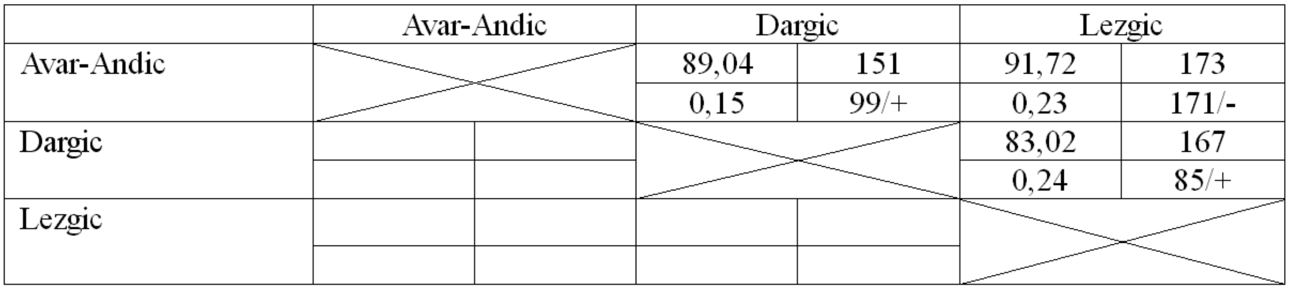
It is interesting that genetic data are completely opposite. Shapsugs act here as native speakers of Adyghe language, while Circassians act correspondingly for Kabardian-Circassian language. The large genetic distance of Shapsugs not only from Circassians, but also from other ethnic groups of the North Caucasus deserves attention. Possibly, if other nationalities speaking dialects of Adyghe language were chosen, then the results involving genetic distances would be different. This question requires additional study.

4.2. Branch level. Avaro-Andic, Dargic, Lezgic languages

Average distances between these groups of languages are provided in Tab. 4.

Dargwa language has many dialects which are considered, for example, in [Koryaks 2006] as independent languages; so it is possible to talk of the Dargic subgroup of languages. Unfortunately, we have no linguistic data on these dialects. However, available genetic data indicate to an undoubted proximity of Dargins, Kaitags and Kubachins. As a result, only literary Dargwa represents this subgroup [Musayev 2001].

Table 4. Distances between branches of Dagestanian group



	Avar-Andic		Dargic		Lezgic	
Avar-Andic			89,04	151	91,72	173
			0,15	99/+	0,23	171/-
Dargic					83,02	167
					0,24	85/+
Lezgic						

The existing genealogical classification does not confirm the existence of a general ancestor for any two of these three groups of languages. Geographically Dargwa language is located strictly between Avar-Andic and Lezgic languages, dividing them and having boundary with languages of these two branches. Therefore, it is necessary to expect that distances between the Avar-Andic and Lezgic branches are greater than the distance from each of them to Dargic. Both lexical and grammatical data confirm this assertion (Tab. 4).

However, genetic data give a different picture. The distance between Dargian and Lezgian peoples is the greatest. Perhaps, the justification lies in some specific customs of these peoples.

Returning to lexical and grammatical distances, let us consider them in more detail at the level of individual languages of these branches.

Among the languages of the Lezgic group, Dargwa is closer to Tabasaran and has a common boundary with it. Dargwa has no common boundary with Lezgi. Lexical distance from Dargwa to Tabasaran equals 80,72, to Lezgi — 85,31. Grammatical distances: to Tabasaran — 158, to Lezgi — 176. Thus, also in this case there is correlation between geographical distance and the existence of a common boundary, on the one hand, and lexical composition and grammar of languages, on the other hand. Genetically Dargins are closer to Lezgians than to Tabasarans.

Among the languages of the Avar-Andic subgroup, Dargwa is closer to Avar and bounders with it. Lexical distance from Dargwa to Avar is 87,05, to Andi — 89,65, to Chamalal — 90,42. Grammatical distances: to Avar — 148, to Andi — 157, to Chamalal — 148. Ethnicities are arranged by genetic proximity with Dargwa as follows: Chamalal, Avar, Andi.

Submit Date: 12.02.2017, Acceptance Date: 05.03.2017, DOI NO: 10.7456/1070ASE/017

Lexical distances strictly correlate with geographical; grammatical, mostly, also correlates, although the distance between Chamalal and Dargwa proved to be slightly less than expected. Genetic distance between Chamalal and Dargwa also turned out to be less than expected. It is possible that there is some connection between these phenomena.

5. CONCLUSION

In general, all types of data considered in this article are in good compliance with genealogical classification at all levels of hierarchy. This allows to use them as predictors of language kinship.

Statistical analysis confirms a high level of correlation between genetics and geography, as well as a significant level of correlation of the lexicon with all other parameters. Grammar correlates the least with other parameters.

The local analysis performed in section 5 revealed the following regularities.

1. In all examples considered, lexical distances agreed with genealogical proximity and, if the latter was not established, then with geographical. A not so good agreement with the confirmed kinship was showed by grammatical distance. The example from section 5.1.1 reflects geographical position – the existence of a common boundary, but no kinship. An even less precise correspondence is showed by genetic distance.

2. It is shown that not only geographical distance is an important factor, but also the existence of a common boundary.

3. As a rule, when several of the considered metrics deviate from genealogical proximity, they correlate with each other. This may be taken as a signal of existence of common factors influencing the development of languages and peoples. Most often deviations can be explained by geographical proximity, in certain cases additional studies are required for this purpose.

4. The considered metrics can be used to back hypotheses on language kinship in situations when the latter has not been established.

Comparison of diversified quantitative data provides new approaches to the determination of migration paths.

ACKNOWLEDGEMENTS

The author would like to express his gratitude to O. Balanovsky for the genetic data provided and for the valuable discussion of the results, and also to E. Lerner for his assistance in statistical calculations. The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

REFERENCES

Balanovsky et al. 2010 — O. P. Balanovsky, A. S. Pshenichnov, R. S. Sychev, E. V. Balanovskaya. Y-base: frequencies of Y-chromosome haplogroups of peoples of the world. 2010, www.genofond.ru (in Russian)

Burlak, Starostin 2001 — S. A. Burlak, S. A. Starostin. Introduction to comparative linguistics. Moscow: URSS, 2001 (in Russian)

Cavalli-Sforza 2000 — L. Cavalli-Sforza. Genes, Peoples, and Languages. Berkeley: University of California Press. 2000)

Derenko et al. 2006 — M. V. Derenko, B. A. Malyarchuk, M. Voznyak, I. K. Dambuyeva, Ch. M. Dorzhu, F.A. Luzina, H.K. Li, D. Mishchitska-Shlivka, I. A. Zakharov. Diversity of Y-chromosome lines in *indigenous population of Southern Siberia. Reports of the Russian Academy of Sciences. 2006, vol. 411, No. 2, pp. 1–5 (in Russian).*

Grimes 2000 — B. Grimes (ed.) Ethnologue. Dallas: SIL International. 2000 .

Haspelmath et al. 2005 – M. Haspelmath, M. Dryer, D. Gil, B. Comrie (eds.) The World Atlas of Language Structures. Oxford, 2005 .

Hunley et al. 2008 — K. Hunley, M. Dunn, E. Lindström, G. Reesink, A. Terrill, M. E. Healy, G. Koki, F. R. Friedlaender, J. S. Friedlaender. Genetic and Linguistic Coevolution in Northern Island Melanesia. PloS Genetics. V. 4 Issue 10, 2008, e1000239. doi:10.1371/journal.pgen.1000239

Knight et al. 2003 — A. Knight, P. Underhill, H. Mortensen, L. Zhivotovsky, A. Lin, B. Henn, D. Louis, M. Ruhlen, J. Mountain. African Y-chromosome and mtDNA divergence provides insight into the history of click languages. Curr Biol. 2003 Mar 18; 13(6): p. 464–73

Kutuev 2010 — I. A. Kutuev. Genetic structure and molecular phylogeography of the Caucasus people. Avtoref. doct. diss. Ufa: Institute of Biochemistry and Genetics of the Ural Scientific Centre, Russian Academy of Sciences. 2010 (in Russian).

Lansing et al. 2007 — J. Lansing, M. Cox, S. Downey, B. Gabler, B. Hallmark, T. Karafet, P. Norquest, J. Schoenfelder, H. Sudoyo, J. Watkins, M. Hammer. Coevolution of languages and genes on the island of Sumba, eastern Indonesia. Proc Natl Acad Sci USA. 2007 Oct 9; 104(41): p. 16022-6. Epub 2007 Oct 3 .

Limborskaya et al. 2002 — S. A. Limborskaya, E. K. Husnutdinova, E. V. Balanovskaya. Ethnogenomics and genogeography of peoples of Eastern Europe. Moscow: Nauka, 2002 (in Russian).

Litvinov 2010 — S. S. Litvinov. Study of genetic structure of peoples of Western Caucasus according to data on Y-chromosome and mitochondrial DNA polymorphism and Alu-insertions. Avtoref. cand. diss. Ufa: Institute of Biochemistry and Genetics of the Ural Scientific Centre, Russian Academy of Sciences. 2010 (in Russian).

Moseley 2010 — C. Moseley (ed.). Atlas of the World's Languages in Danger. Paris, UNESCO Publishing. 2010. Online version: http://www.unesco.org/culture/en/ endangeredlanguages/atlas. Open access .

Müller et al. 2010 — A. Müller, S. Wichmann, V. Velupillai, C. H. Brown, P. Brown, S. Sauppe, E. W. Holman, D. Bakker, J.-M. List, D. Egorov, O. Belyaev, R. Mailhammer, M. Urban, H. Geyer, A. Grant. ASJP World Language Tree of Lexical Similarity: Version 3. 2010 .

Nasidze et al. 2001 — I. Nasidze, G. Risch, M. Robichaux, S. Sherry, M. Batzer, M. Stoneking. Alu insertion polymorphisms and the genetic structure of human populations from the Caucasus. European Journal of Human Genetics. 2001. 9: p. 267–272 .

Polyakov, Solovyev 2006 — V. N. Polyakov, V. D. Solovyev. Computer models and methods in typology and comparative linguistics. Kazan, 2006 (in Russian).

Solovyev, Faskhutdinov 2009 — V. D. Solovyev, R. F. Faskhutdinov. Technique of evaluation of the stability of grammatical properties. Proc. of Russian Academy of Sciences. Series on literature and language. Vol. 68. No. 4. 2009 (in Russian).

Solovyev, Faskhutdinov 2011 — V. Solovyev, R. Faskhutdinov. Comparative analysis of phylogenic algorithms. FOI ITHEA, 2011 .

Yunusbayev 2006 — B. B. Yunusbayev. Populational and genetic research of the people of Dagestan according to data on Y-chromosome polymorphism and Alu-insertions. Avtoref. cand. diss. Ufa: Institute of Biochemistry and Genetics of the Ural Scientific Centre, Russian Academy of Sciences. 2006 (in Russian).