Alexei Kassian (Institute of Linguistics of the Russian Academy of Sciences / RSUH)
a.kassian@gmail.com
Draft, 17 March, 2013

The Lezgian linguistic group within the framework of the Global Lexicostatistical Database

The current version of the GLD Lezgian database (Kassian 2011–2012)[1] consists of 110-item wordlists of 20 Lezgian lects. This is the maximum number of Lezgian lects for which available lexicographic sources permit to compile a Swadesh wordlist.

These 20 synchronic wordlists meet the high standards of the *Global Lexicostatistical Database* project.

1) All relevant sources — dictionaries, grammars, text corpora — are taken into account. This includes not only modern publications, but also data collected and published by Peter von Uslar, Adolf Dirr, Adalbert Starchevsky and other Caucasologists in the late 19th – early 20th century.
2) All the analyzed forms are uniformly transcribed in an IPA-based alphabet; additionally, traditional Cyrillic spellings are quoted in parentheses.
3) Many important phonetic, morphological and semantic details are explicitly discussed in the annotations. This especially concerns occasional (quasi-)synonymy and regular quasi-synonymy. For example, notes on the entries 'hand' and 'foot' also include an obligatory discussion of expressions for 'arm' and 'leg' (similarly 'fog' and 'rain cloud' in the entry for 'cloud'; 'hot' in the entry for 'warm'; 'male' and 'husband' in the entry for 'man' and so on).

Compilation of Swadesh wordlists faces difficulties of different kinds, of which the following problems are the most typical.

1) The needed word is missing from the available dictionary or glossary as a separate entry. This may not be a big problem, however, if the used dictionary is illustrated by textual examples or if a text corpus for this lect is available. For example, the basic Budukh-Russian dictionary (*ca.* 8000 entries) lacks separate Budukh entries for the following terms: 'to eat', 'to fly', 'to hear', 'to kill'. However, all these verbs can actually be extracted from various textual examples quoted in the same dictionary.
2) The available dictionary can propose more than one term for the meaning in question, but does not provide us with any additional information. Full synonymy is a relatively rare phenomenon in human language, so it is expected that several terms, quoted as synonyms in our source, in fact differ either semantically or pragmatically. For instance, if we observe two documented terms for the meaning

[1] http://starling.rinet.ru/cgi-bin/response.cgi?root=new100&morpho=0&basename=new100\ncc\lez&limit=-1

‘to sit’, the first verb can in fact have the active meaning ‘to sit down’, the other — the stative meaning ‘to sit’. Or the first word is archaic and obsolete, while the second one is neutral and common. Or the first word may be bound, being only used in fixed expressions (like ‘to mount a horse’), while the second one does not have any such restrictions (in all these cases, it is the second verb that must be taken as a Swadesh item). In such a case, only browsing through the text corpus can clarify the situation. As a typical example, I may quote the Udi words for ‘earth’. There are two documented Udi terms: *očˤal* and *k:ul*, both of them glossed as ‘earth’ or specified as ‘earth, soil’ in Udi dictionaries. Browsing through various text collections suggests, however, that in the Nidzh dialect of Udi, *k:ul* does indeed mean ‘soil’, but this term is synchronically obsolete, whereas *očˤal* is currently the common generic term with polysemy ‘earth, soil, ground, land’. On the contrary, in Vartashen Udi, *k:ul* is a full-fledged term for ‘soil’, whereas *očˤal* means ‘ground’.

3) The morphological structure of the word quoted in the sources is not transparent. For instance, in Lezgian lexicography, there is no tradition to single out verbal spatial (sometimes desemanticized) prefixes. Thus, the Aghul verb *qark-i-* ‘to lie (down)’ actually consists of the root *ark-* modified with the desemanticized spatial prefix *q=*, as follows from the synonymical Aghul verbs *ut:=ark-i-, ʁ=ark-i-* (all of them glossed as ‘to lie’) and the paronymous verb ʢ=*ark-i-* 'to sleep'.

It is only after the high-quality synchronic wordlists have been compiled that it becomes possible to build phylogenetic trees of the linguistic family in question.

Phonetic similarity

For the first stage of marking cognation, we can use a formal algorithm, based on phonetic similarity. There are two most popular approaches to the automatic establishing of cognate word pairs between the given wordlists: Levenshtein distances and consonant classes. In fact, the method of consonant classes may be considered a crude variation on the measurement of Levenshtein distances. Below I will rely on consonant classes, since I am not aware of any publications which demonstrate that consonant classes yield significantly less reliable results than Levenshtein distances.

The method of consonant classes was proposed by A. Dolgopolsky in 1964 (English version: 1986) and successfully tested by various authors on the data of various languages of Eurasia.

This method implies that the phonetic alphabet used in our studies can be divided into several non-intersecting subsets (classes) so that phonetic mutations between the sounds of one class during natural language development are typologically more normal than mutations between sounds that belong to different classes. Below, I operate with classes that are currently accepted in the *Global Lexicostatistical Database* project (GLD):

P-class (labials): p b ɓ β f v…
T-class (dentals): t d ɗ θ ð…
S-class (front affricates & fricatives): c ʒ č ǯ s z š ž…
Y-class (palatal glides): y…
W-class (labial glides): w ʍ…
M-class (labial nasals): m ɱ…
N-class (non-labial nasals): n ɳ ɲ ŋ…
Q-class (lateral affricates & fricatives): ƛ ɬ…
R-class (liquida): r ɾ l ɭ ł…
K-class (velars & uvulars): k g x ɣ q χ ʁ…
zero-class or H-class: ħ ʕ ʜ ʢ ʡ h ɦ ʔ and any vowels.

Using this simplified transcription system (*P T S Y W M N Q R K H*), we can encode any real word forms or morphemes included into comparison. Note that elements of the zero-class and such features as coarticulation, prosody, phonation are deleted from the structure. Vocalic onset or vocalic final, however, are coded as *H*. Thus both hypothetical forms *tasa* and *d*^{h}*üʒo* are coded as *TSH*; *alaq* and *ʡärx* = *HRK*; *na* and *ŋoʔ* = *NH*; *pk*^{h}*ot* and *baq'aθ* = *PKT*; *wahat* and *ʍad* = *WT*. Non-initial *Y* and *W* (weak glides) are treated as *H*, thus *ka*, *kay*, *kawa* = *KH*, whereas *kat* and *kayat* = *KT*. As follows from the above, two forms from compared languages possessing identical simplified transcriptions have a better chance to appear to be etymological cognates than forms whose simplified transcriptions differ.

All the Lezgian wordlists have been coded in this way, whereupon cognations have been automatically established in the Starling software: two forms are marked as cognates if the first two consonants in their simplified transcriptions coincide. For instance, the words for 'ashes': Kryts *räq* (*RK*) = Aghul *rüq:*^{y} (*RK*) ≠ Tsakhur *yiq*^{ʕ}- (*YK*) ≠ Archi *diq':*^{ʕ}- (*TK*), even though in reality all the forms originate from one proto-root. On the contrary, Udi *kul* 'hand' (*KL*) = Tsakhur *χɨl*^{y} 'id.' (*KL*), even though these forms actually originate from different proto-roots.

The following tree was produced by the Starling software (neighbor joining method).

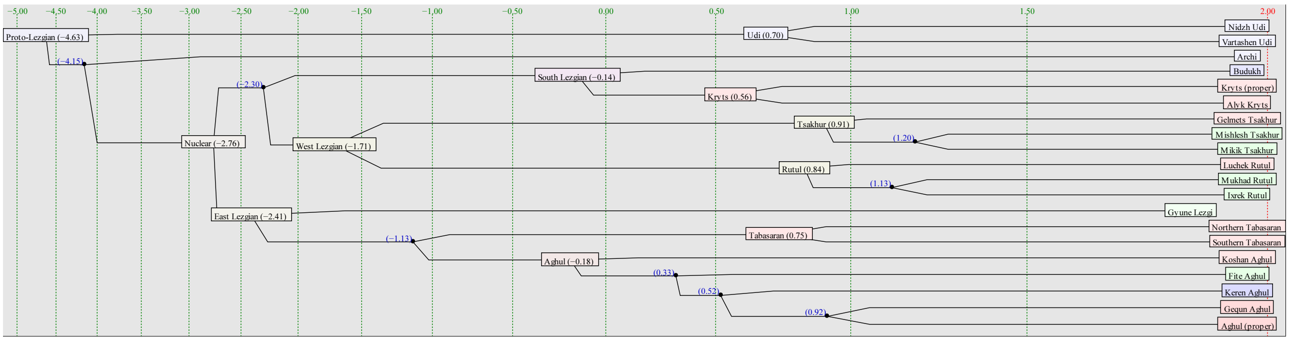


Fig. 1. Phylogenetic tree of the Lezgian group (neighbor joining method).
110-item wordlists, analyzed by phonetic similarity (C_1C_2 matches).

Etymological analysis

The second option for comparison involves setting up cognation indexes for individual forms with the help of the traditional comparative method. I use the Proto-Lezgian reconstruction by the late Sergei Starostin (NCED: 122 ff.; LEDb; Starostin n.d.) with certain corrections and improvements when necessary. S. Starostin's work is the only full-fledged Proto-Lezgian reconstruction, which has been published so far. Recently, Wolfgang Schulze (Schulze 2001; Gippert et al. 2008; etc.) has announced his own version of the Proto-Lezgian reconstruction. The amount of Schulze's etymologies for individual Lezgian roots that has already been published by the author is insufficient for any final decisions, but I must note that a significant number of Schulze's ideas does not look acceptable from my point of view.

The following tree was produced by the Starling software (neighbor joining method).

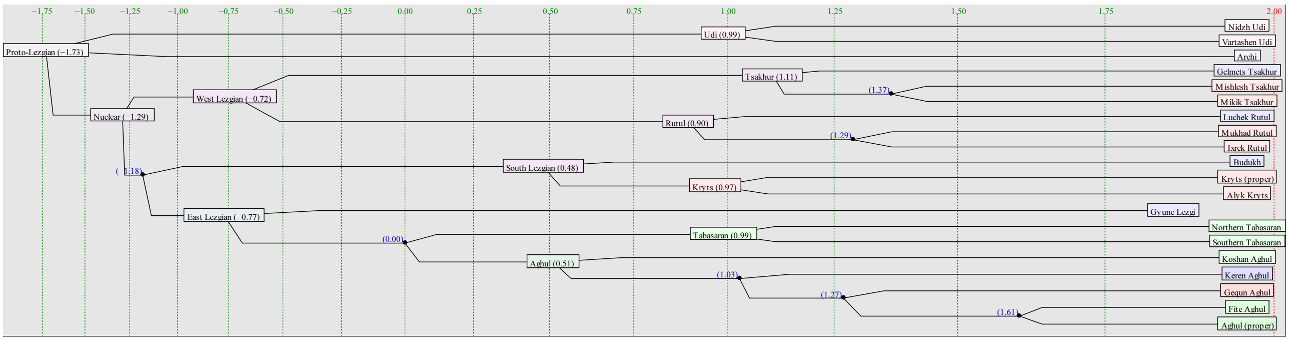


Fig. 2. Phylogenetic tree of the Lezgian group (neighbor joining method).
Etymologically elaborated 110-item wordlists.

As one can see, the initial split was a three-way one: (1) Proto-Udi-Caucasian Albanian, (2) Proto-Archi, and (3) Proto-Nuclear Lezgian. Later, the Proto-Nuclear Lezgian language split into the three following branches: (1) Proto-West Lezgian [Tsakhur, Rutul], (2) Proto-South Lezgian [Kryts, Budukh], and (3) Proto-East Lezgian [Aghul, Tabasaran, Lezgi].

The first interesting thing to note is that the automatic phonetic tree (Fig.1) appears to be rather good. The main discrepancy between Fig.1 and Fig.2 is that the automatic tree (Fig.1) suggests that the split of Proto-Nuclear was rather four-way: West, South, East and Lezgi proper. The second — expected — peculiarity of the automatic tree (Fig.1) is that the datings of the nodes are much deeper, because in many cases, phonetic mutations during natural language development make the relatedness of the forms unrecognizable for the automatic algorithm.

The second important thing is that such a tree with two outliers (Udi & Archi) and a large group of nuclear a.k.a. Samur lects (with three subgroups) conforms with the traditional Lezgian classification (see, e.g., Talibov 1980: 11-16 with further references) as well as with some previous, rougher, lexicostatistical calculations (Alekseev 1984).

On the contrary, the lexicostatistical classification, according to which Archi is the fourth Nuclear Lezgian branch (thus Alekseev 1985: 17-23; Koryakov 2006: 21), is not confirmed and should now be rejected.

Similarly, Schulze's opinion (Schulze 2005; Gippert et al. 2008: II-65-75; despite Schulze-Fürhoff 1994: 450) that Udi-Caucasian Albanian belongs to the East Lezgian subgroup, together with Aghul, Tabasaran and Lezgi, appears to be incorrect. Schulze (Gippert et al. 2008: II-65-75) published his own version of the Caucasian Albanian and Udi Swadesh wordlists and compared these to general Lezgian data. Unfortunately, Schulze did not provide any explanations for his specific version of lexicostatistics, and at the same time the general lexicographic quality of the compiled wordlists is rather low. For these reasons, I have to conclude that Schulze failed to present any formal arguments in support of his tree. On an intuitive basis, Schulze's classification seems just as wrong — I am not aware of any specialists in Lezgian studies that would regard Schulze's tree as acceptable.

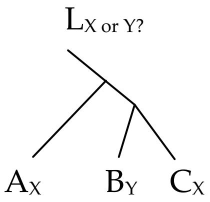
Proto-Lezgian list

The next step is the reconstruction of the Proto-Lezgian 110-item wordlist, since the philosophy of the GLD project implies sequential reconstruction, meaning that the lexicostatistical classification of the next level taxon (North Caucasian family) will be based on the reconstructed wordlists of intermediate groups, not on the synchronic wordlists of individual lects.

The GLD Lezgian database (Kassian 2011–2012) contains the Proto-Lezgian wordlist, which was reconstructed according to the following principles.

1) Topological principle. The configuration of the genealogical tree of a linguistic family is important for semantic reconstruction. In the situation when a proto-word has several different meanings in ancestral languages, one of the strongest criteria

for its semantic reconstruction is the topological one. Let us envision the following genealogical tree where *L* is a proto-language and *A*, *B*, *C* are its daughter languages.



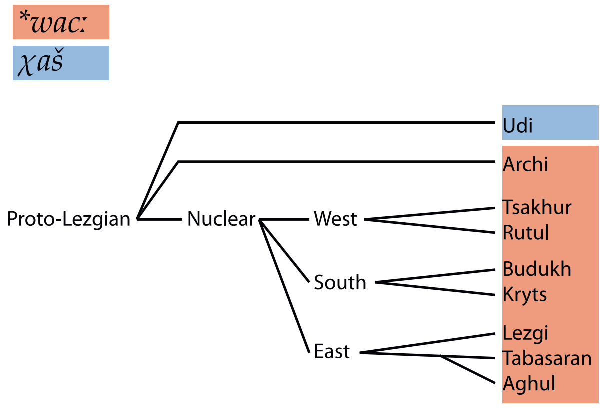
A certain word in *A* means ‘X’, its etymological cognate in *B* has a different meaning ‘Y’, whereas its cognate in *C* also means ‘X’. In the absence of additional evidence, the likeliest solution is that, in *L*, this word denoted ‘X’, not ‘Y’, since, according to general scientific principles, we should prefer the most economic scenario (one semantic shift ‘X’ > ‘Y’ in *B* vs. two independent shifts ‘Y’ > ‘X’ in *A* and *C*).

2) External etymological principle. If there are two or more equal lexical candidates for the status of the proto-term in question, the root that possesses the better external etymology and a better chance to represent the Swadesh proto-term of the next level taxon has the advantage.
3) Internal etymological principle. If there are two lexical candidates for the status of the proto-term in question, and one of them is morphologically primary, whereas the other is a derivative (for example, ‘moon’ ← ‘to shine’, ‘green’ ← ‘grass’ and so on), the first term has an advantage. The same applies to the situation when one of the competing terms possesses some etymological cognates (either internal, that is, within the same language group, or external), whereas the second term is etymologically isolated. In this case, the first term has an advantage, while the second one, on the contrary, is a potential loanword.
4) Semantic principle. If there are two lexical candidates for the status of the proto-term in question, both having various cognates in related lects, the assumed semantic shifts must be examined and weighed up, because many common semantic shifts are indeed bilateral, but in some cases only one specific direction of semantic development is typologically possible. For instance, the shift ‘green’ ↔ ‘grass’ is bilateral (there are a lot of typological instances for both directions), but in the pair ‘moon’ and ‘to shine’ only the development ‘to shine’ → ‘moon’ is possible, not vice versa.
5) Areal principle. If a term has an areal distribution, that is, it is attested in several neighboring lects, and we have strong evidence that these lects are in contact and influence each other, such a word can represent a late introduction, which has spread as an interdialectal loan.

These principles can be illustrated with the help of specific Lezgian examples.

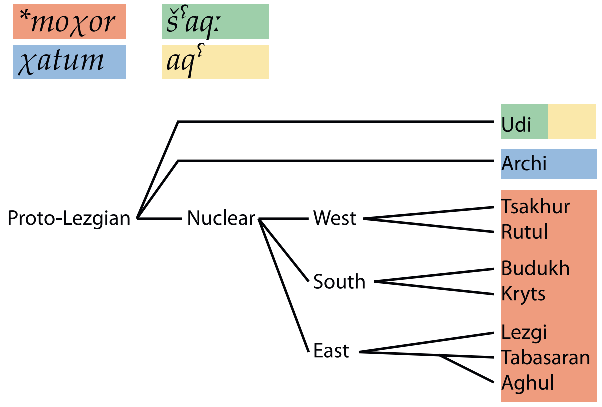
MOON. Two Lezgian terms enter into competition: **wac:* (everywhere except for Udi) and *χaš* (Udi). This is a trivial case, because **wac:* has an advantage according to the topological principle (**wac:* is retained in two of the three branches: Archi and Nuclear), the external etymological one (**wac:* goes back to the obvious Proto-North Caucasian term for 'moon') and the semantic one (Udi *χaš* possesses the polysemy 'moon / shining' that implies the original meaning 'shining').

'moon'

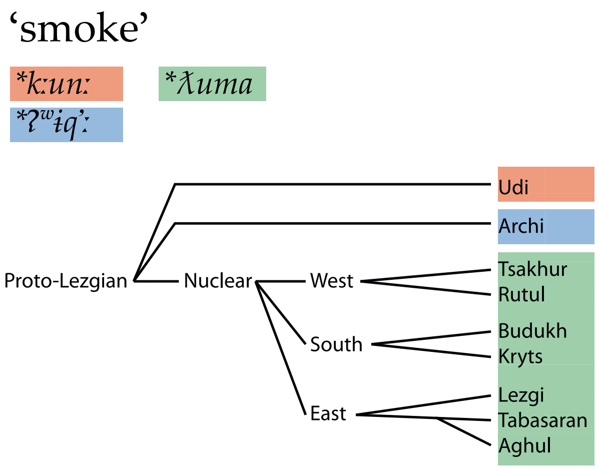


BREAST. This is a more interesting case. Several terms enter into competition: *šˤaq:* and *aqˤ* in the Udi dialects, *χatum* in Archi, and **moχor* in Nuclear. Topologically, all four terms are equally acceptable. The Udi words demonstrate polysemy 'breast / mountain slope', but this does not clarify the situation, because both directions of semantic shift between 'breast' and 'slope' are possible. Fortunately, **moχor* is also retained in Archi, where it means 'brisket'. Since the shift 'breast' > 'brisket' is normal, but not vice versa, **moχor* can be posited as the Proto-Archi term for 'breast'. In sum, we have **moχor* as a word for 'breast' in two of the three branches (Archi and Nuclear) that allows us to reconstruct the Proto-Lezgian item 'breast' as **moχor* according to the topological principle.

'breast'



SMOKE. Each of the three branches possesses its own term for ‘smoke’: **k:un:* in Udi, **ʔʷɨq’:* in Archi and **ƛuma* in Nuclear. Out of these, **k:un:* is also retained with the meaning ‘dust’ in Archi (but the development ‘smoke’ ↔ ‘dust’ can probably be bilateral), whereas **ʔʷɨq’:* and **ƛuma* lack any cognates in other branches. Thus there are no internal Lezgian hints to choose between the three. According to the external etymological principle, we can posit **k:un:* as the Proto-Lezgian term for ‘smoke’, since it possesses good North Caucasian *comparanda* with the same basic meaning, whereas North Caucasian cognates of **ʔʷɨq’:* and **ƛuma* point to meanings like ‘fumes, stink, wind, air’.



Elaboration of Lezgian data demonstrates that in the majority of cases, the aforementioned principles do not contradict each other (this is a reliable indicator of the high quality of the collected synchronic wordlists). In a couple of cases, however, these principles may come into conflict, see notes on ‘black’ and ‘blood’ (the latter is an especially problematic word in Lezgian).

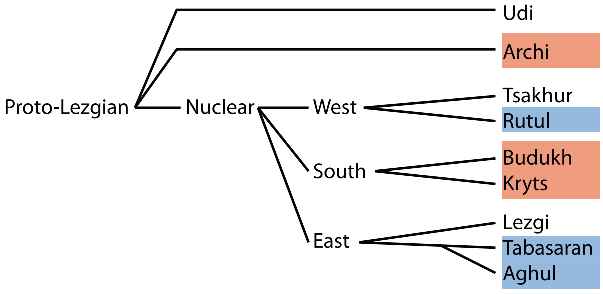
A more common problematic situation is when we have two (or even more) candidates for the status of the proto-term, which are arranged into a so-called «criss-crossed» configuration from the topological point of view. Normally, in such cases, we are dealing with areal, that is, contact-driven isoglosses. Cf. the following example.

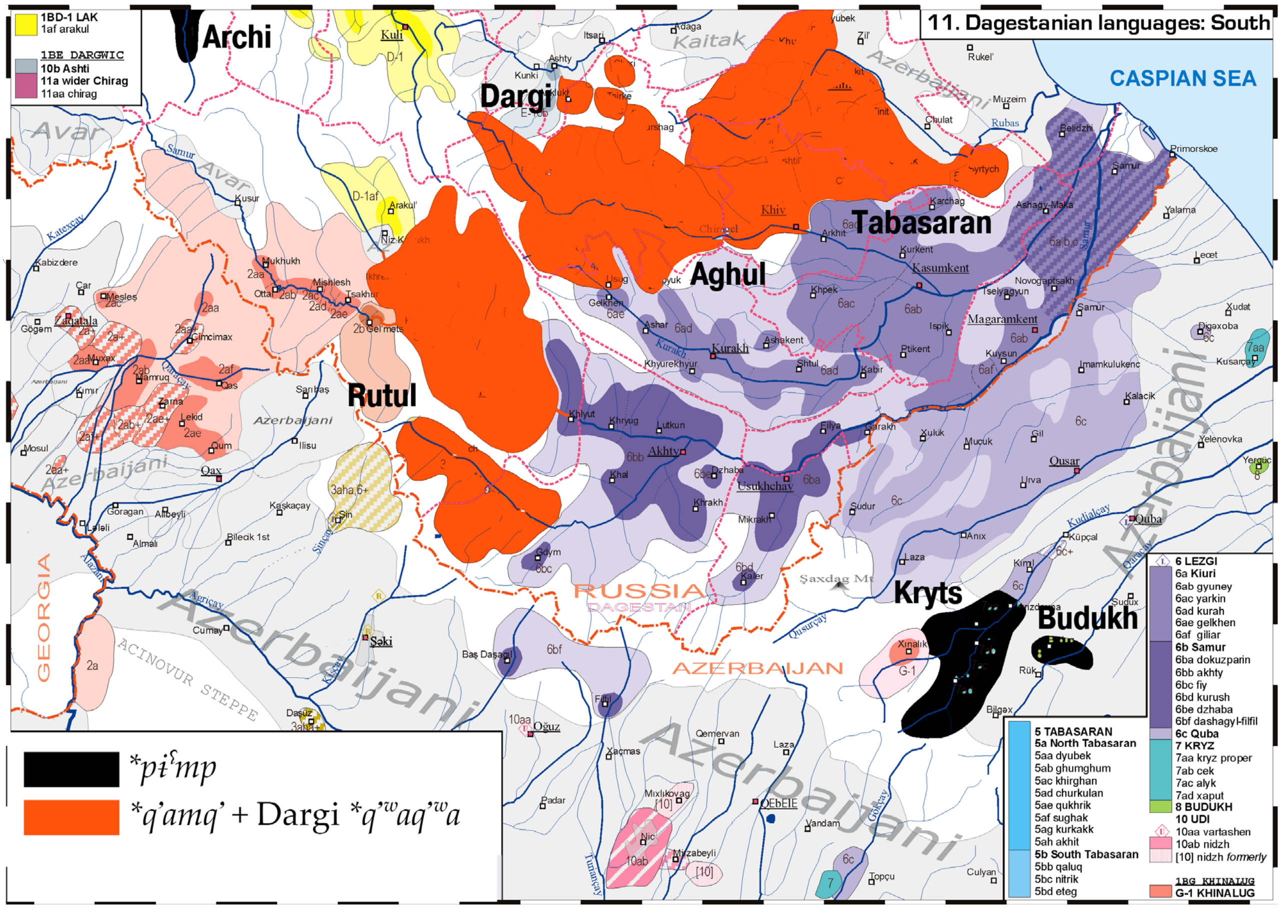
KNEE. Two roots enter into competition in this criss-crossed situation: (1) **pɨˤmp*, which means ‘knee’ in Archi, on the one hand, and in South Lezgian (Kryts, Budukh), on the other; and (2) **q’amq’*, which means ‘knee’ in some West Lezgian (Rutul) and some East Lezgian (Aghul, Tabasaran) lects. The original meaning of the latter Proto-Lezgian root **q’amq’* is unknown, but it is likely that **q’amq’* ‘knee’ together with Proto-Dargi **q’ʷaq’ʷa* ‘knee’ represent a relatively late areal isogloss, which affected several closely related neighboring Lezgian lects as well as adjacent Dargi territory.

‘knee’

**piʕmp*

**q’amq’* + Dargi **q’ʷaq’ʷa*

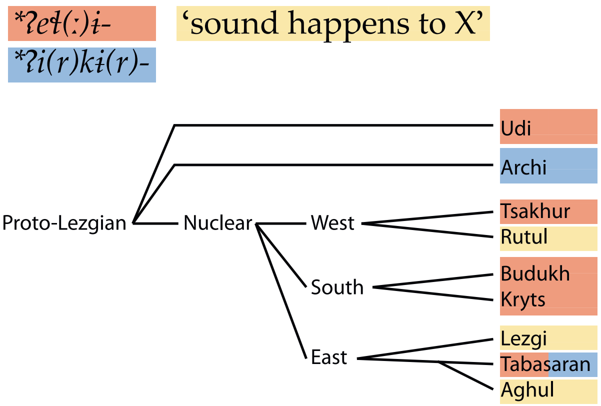


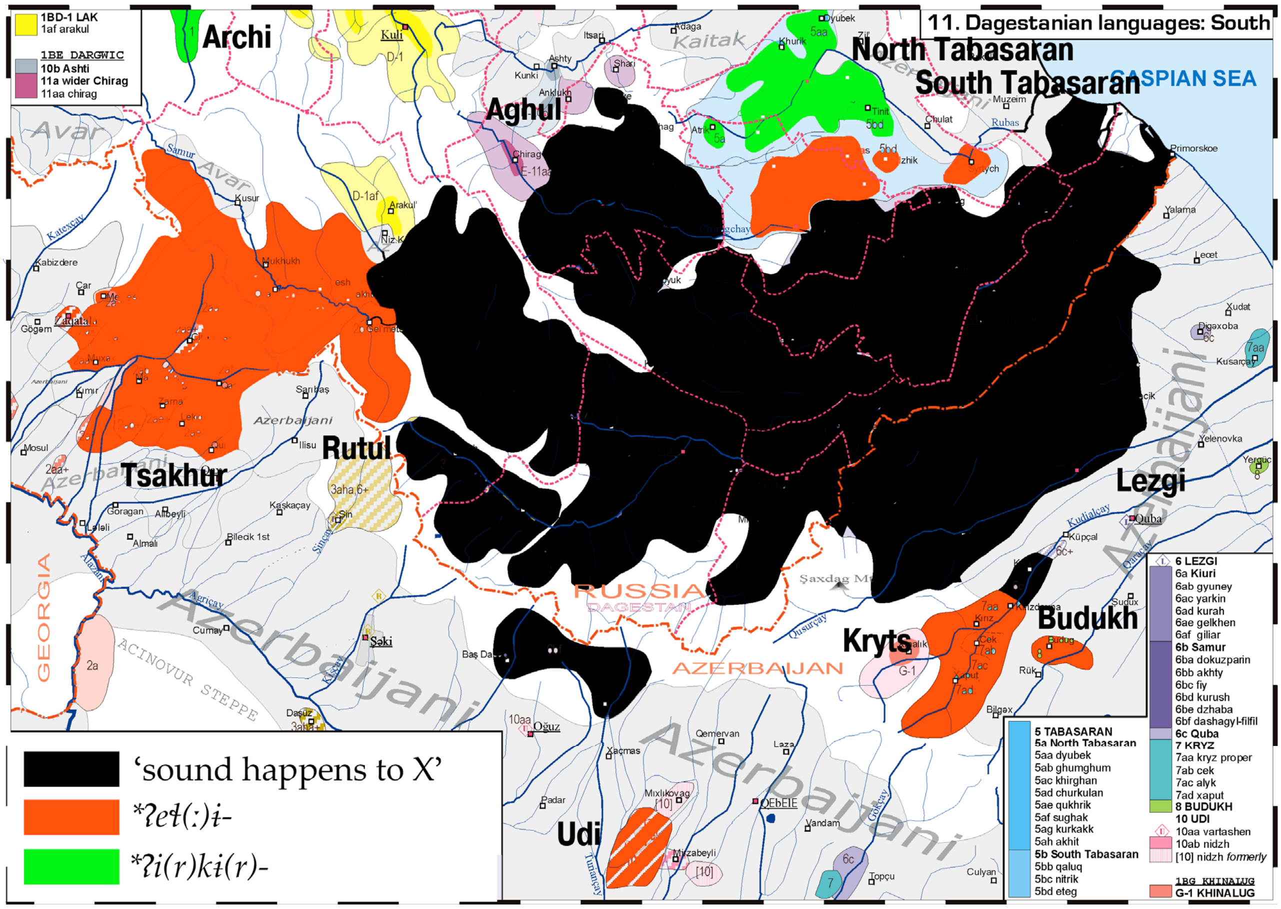


In some cases, however, we are forced to assume an independent semantic development in non-contacting lects.

TO HEAR. Three expressions enter into competition: (1) the verbal root **ʔeɬ(:)ɨ-*; (2) the verbal root **ʔi(r)kɨ(r)-*; (3) the analytic construction ‘sound happens to X’ with different words for 'sound' and different auxiliary verbs. The expression ‘sound happens to X’ is an obvious late areal isogloss that affected the central part of the Lezgian territory: Rutul, Aghul, Lezgi. The real choice consists of two Proto-Lezgian roots (1) **ʔeɬ(:)ɨ-*; (2) **ʔi(r)kɨ(r)-*. The former one, **ʔeɬ(:)ɨ-*, is attested as the basic verb ‘to hear’ in Udi, on the one hand, and in many Nuclear Lezgian lects, on the other: South Lezgian (Kryts, Budukh), Tsakhur and the Southern dialect of Tabasaran. The latter one, *ʔi(r)kɨ(r)-, means ‘to hear’ in Archi and the Northern dialect of Tabasaran. Since *ʔi(r)kɨ(r)- has a quite modest distribution, whereas *ʔeɬ(:)ɨ- displays external North Caucasian *comparanda* with the meaning ‘to hear’, it is likely that Archi and Northern Tabasaran demonstrate an independent semantic development of the verb *ʔi(r)kɨ(r)- ‘to perceive (in some way)’ > ‘to hear’.

‘to hear’





A specific peculiarity of the Lezgian group is a number of cases where the Proto-Lezgian Swadesh term is not reconstructible, since in almost all lects, inherited forms were superseded with loanwords (this applies to such words as ‘all’, ‘person’, ‘to swim’).

Improvement of the tree

After the compiled wordlists have been etymologically and distributionally elaborated, it is possible to improve the tree, rejecting those etymological matches which we consider to be secondary (as described above, these cases represent either contact-driven isoglosses or independent semantic developments).

In a small number of cases, it is possible to establish the direction of lexical influence between two lects. I mark such items as loanwords for the recipient lect. Normally, however, there is no direct evidence for the direction of influence; I mark such items as unrelated forms, even though these originate from one proto-root.

For instance, in the case of 'knee' (treated above), Rutul *q'ʷaq'* 'knee' is marked as unrelated to Aghul *q'ʷaq'ʷ*, Tabasaran *q'amq'* 'knee'.

Similarly, in the case of 'to hear' (treated above), Archi: *=ko-* 'to hear' and Northern Tabasaran *yik-* 'to hear' are marked as unrelated.

As a result of this revision, out of 110 Swadesh entries, internal numeration of cognate forms has been changed in more than 30 entries. Such a solid number is conditioned by the fact of active contacts and interlingual interaction between Nuclear Lezgian lects.

The following revised tree was produced by the Starling software (neighbor joining method).

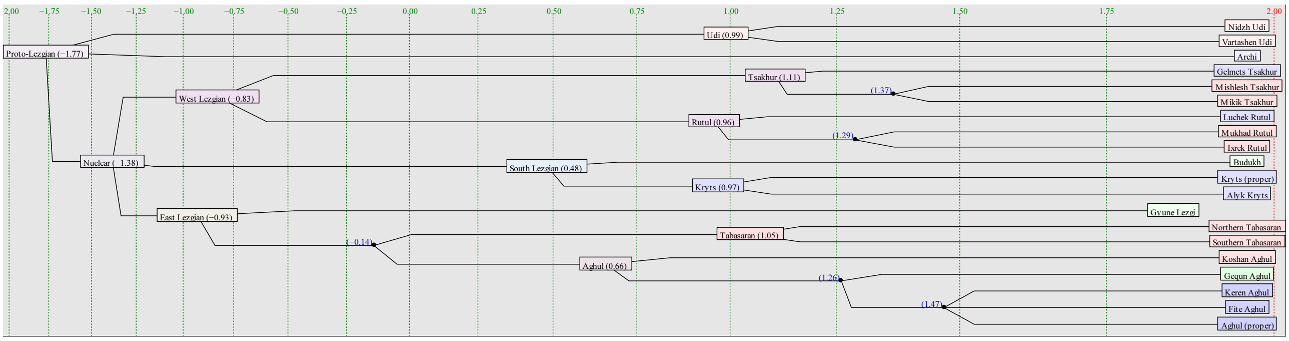


Fig. 3. Phylogenetic tree of the Lezgian group (neighbor joining method). Etymologically elaborated 110-item wordlists. Secondary matches are rejected

The only topological discrepancy between Fig.2 and Fig.3 is the rearrangement of the non-Koshan dialects of Aghul, although the resulting time gaps are too small on the both trees.

Conclusions

Due to high quality of the input data, the GLD Lezgian database (Kassian 2011–2012) is a good testing area for various methods of lexicostatistical classification. On all the schemes displayed in the current presentation, the data have been treated by the neighbor joining method, which is the default one in the Starling software. Two other relevant methods, which I plan to test, are the Bayesian MCMC approach (as described in Gray & Atkinson 2003; Kitchen et al. 2009; Bouckaert et al 2012) and the recently proposed method of minimal average errors (Vasilyev 2010; Vasilyev & Kogan forth.).

It is expected that both methods will yield trees that will be topologically compatible with those produced by the neighbor joining method (Fig.2-3).

References

Alekseev 1984 — M. E. Alekseev. K voprosu o klassifikacii lezginskikh yazykov. In: *Voprosy yazykoznaniya*, 1984, No. 5. P. 88–94.

Alekseev 1985 — M. E. Alekseev. *Voprosy sravnitel'no-istoricheskoj grammatiki lezginskikh yazykov. Morfologiya, Sintaksis*. [M. E. Alekseyev. *Issues of comparative-historical grammar of Lezgian languages*]. Moskva, 1985.

Bouckaert et al 2012 — R. Bouckaert, P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard, Q. D. Atkinson. Mapping the Origins and Expansion of the Indo-European Language Family. In: *Science*, 337 (24 August 2012). P. 957–960.

Dolgopolsky 1964 — A. B. Dolgopolsky. Gipoteza drevnejshego rodstva yazykov Severnoj Evrazii s veroyatnostnoj tochki zreniya. In: *Voprosy yazykoznaniya*, 1964, No. 2. P. 53–63.

Dolgopolsky 1986 —A. B. Dolgopolsky. A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia. In: V. V. Shevoroshkin & T. L. Markey (eds.). *Typology, Relationship, and Time: A Collection of Papers on Language Change and Relationship by Soviet Linguists*. Ann Arbor (MI): Karoma, 1986. P. 27–50.

Gippert et al. 2008 — J. Gippert, W. Schulze, Z. Aleksidze, J.-P. Mahé. *The Caucasian Albanian Palimpsests of Mt. Sinai*. 2 vols. Brepols, 2008.

Gray & Atkinson 2003 — R. D. Gray, Q. D. Atkinson. Language-tree divergence times support the Anatolian theory of Indo-European origin. In: *Nature*, 426 (2003). P. 435-9.

Kassian 2011–2012 — *Annotated Swadesh wordlists for the Lezgian group (North Caucasian family)*. Database compiled and annotated by A. Kassian (November 2011 – October 2012). Available at GLD: http://starling.rinet.ru/cgi-bin/response.cgi?root=new100&morpho=0&basename=new100\ncc\lez&limit=-1

Kitchen et al. 2009 — A. Kitchen, C. Ehret, Sh. Assefa & C. J. Mulligan. Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. In: *Proceedings of the Royal Society: Biological Sciences* 276 (2009). P. 2703–2710.

Koryakov 2006 — Yu. B. Koryakov. *Atlas kavkazskikh yazykov. S prilozheniem polnogo reestra yazykov*. [*Atlas of Caucasian languages*]. Moskva, 2006.

LEDb — S. A. Starostin. *Lezgian Etymological Database*. Available at http://starling.rinet.ru/cgi-bin/main.cgi?flags=eygtnnl

NCED — S. A. Starostin, S. L. Nikolayev. *A North Caucasian Etymological Dictionary*. Moscow: Asterisk Publishers, 1994. Reprint in 3 vols.: Ann Arbor: Caravan Books, 2007.

Schulze 2001 — W. Schulze. *The Udi Gospels. Annotated text, etymological index, lemmatized concordance*. München/Newcastle: Lincom, 2001.

Schulze 2005 — W. Schulze. Towards a History of Udi. In: *International Journal of Diachronic Linguistics*, 1 (2005). P. 55–91.

Schulze-Fürhoff 1994 — W. Schulze-Fürhoff. Udi. In: R. Smeets. *The Indigenous Languages of the Caucasus*, vol. 4. Caravan Books, 1994. P. 447–514.

S. Starostin n.d. — S. Starostin. *Lezginskaya rekonstrukciya* [*Lezgian Reconstruction*]. MS, the 1980's.

Talibov 1980 — *Sravnitel'naya fonetika lezginskikh yazykov* [*Historical Phonetics of Lezgian Languages*]. Moskva, 1980.

Vasilyev 2010 — M. E. Vasilyev. Ob ispol'zovanii leksicheskogo kriteriya dlya postroeniya genealogicheskoj klassifikacii. In: *Byulleten' Obschestva vostokovedov RAN*, 17 (2010). P. 530–572.

Vasilyev & Kogan forth. — M. E. Vasilyev, A. I. Kogan. K voprosu o vostochnodardskoj yazykovoj obschnosti. In: *Journal of language Relationship*, forthcoming issue.