Comparing phonological systems and syllable structure of Botlikh and Zilo Andi: a data–driven analysis

G. Moroz

Linguistic Convergence Laboratory, HSE University, Moscow, Russia

25 February 2020, MPI-SHH, Jena

Presentation is available here: tinyurl.com/rvpqdaa

Phonological description: data-driven analysis

	Traditional analysis	Data-driven analysis
1.	Done by trained linguist	Evaluated by trained linguist
2.	Can be done from scratch	Previous description needed (or at least prior expectations)
3.	Doesn’t care about amount of data	Care more about amount of data
4.	Less reproducible	More reproducible
5.	Can not be automated	Can be automated

Phonological description: data-driven analysis

	Traditional analysis	Data-driven analysis
1.	Done by trained linguist	Evaluated by trained linguist
2.	Can be done from scratch	Previous description needed (or at least prior expectations)
3.	Doesn't care about amount of data	Care more about amount of data
4.	Less reproducible	More reproducible
5.	Can not be automated	Can be automated

Data-driven approach to phonological description and syllable structure analysis:

- was proposed in (Moroz 2018)
- was applied to syllable structure in (Moroz 2019) to Adyghe data
- was applied to syllable structure in (Romanova 2019) to Russian and Macedonian data

Phonological description: data-driven analysis

The main steps:

- start with some language corpus (corpus, dictionary etc.)
- automatically extract phonological units (segments, features, syllable structure etc.)

Phonological description: data-driven analysis

The main steps:

- start with some language corpus (corpus, dictionary etc.)
- automatically extract phonological units (segments, features, syllable structure etc.)

Further steps:

- obtained frequencies and inventories could be compared with the same units from other languages

Phonological description: data-driven analysis

The main steps:

- start with some language corpus (corpus, dictionary etc.)
- automatically extract phonological units (segments, features, syllable structure etc.)

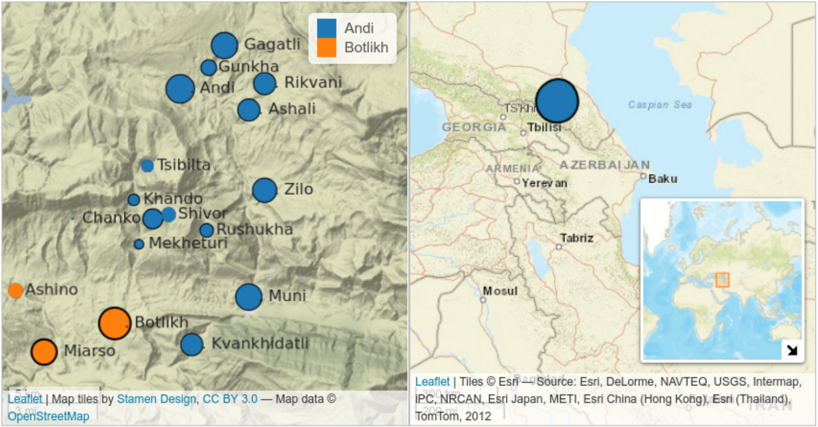
Further steps:

- obtained frequencies and inventories could be compared with the same units from other languages

Advantages:

- more reproducible
- could be updated with new data, see (Moroz 2019) slides from SLE on Bayesian typological research
- answers the question *'How often is X present in language*(*s*)*?'* rather than *'Is X present in language*(*s*)*?'*

Andi and Botlikh villages



Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap

Leaflet | Tiles © Esri — Source: Esri, DeLorme, NAVTEQ, USGS, Intermap, iPC, NRCAN, Esri Japan, METI, Esri China (Hong Kong), Esri (Thailand), TomTom, 2012

- Size of the dot corresponds with number of villages’ inhabitants
- All villages except Botlikh are monoethnic
- Created with lingtypology package (Moroz 2017)

Botlikh < Andic group < EC	Andi < Andic group < EC
Unwritten (can be written with extended Cyrillic script for Avar)	Unwritten (can be written with extended Cyrillic script for Avar)
~5,000–8,000 speakers	~16,500 speakers
Mostly spoken in 3 villages in northwestern Daghestan (Russian Federation): Botlikh, Miarso, Ashino, (Ankho); minor dialectal differences	About 14 villages; There are two main dialect groups: Lower Andi (Muni, Kvankhidatli) and Upper Andi (the rest)
One full reference grammar in Georgian (Gudava 1962)	Several reference grammars (Suleymanov 1957) (Rikvani), (Salimov 1968) (Gagatli), (Tsertsvadze 1965) (Andi)
Two dictionaries: (Saidova and Abusov 2012), (Alekseev and Azaev 2019)	No dictionary except (Kibrik and Kodzasov 1988)

Comparing two Botlikh dictionaries

(Saidova and Abusov 2012)

- Compiled in the 2000s by a native speaker (M. G. Abusov) and an experienced linguist (P. A. Saidova)
- Mostly Botlikh with some notes on Miarso

Comparing two Botlikh dictionaries

(Saidova and Abusov 2012)

- Compiled in the 2000s by a native speaker (M. G. Abusov) and an experienced linguist (P. A. Saidova)
- Mostly Botlikh with some notes on Miarso

(Alekseev and Azaev 2019)

- Compiled in the 1960s / 1970s by a native speaker / philologist (X. G. Azaev) and later (in the 2000s) systematized by an experienced linguist (M. E. Alekseev)
- Subsequently edited by T. A. Maisak and scheduled for posthumous publication last year
- Botlikh only

Comparing two Botlikh dictionaries

(Saidova and Abusov 2012)

- Compiled in the 2000s by a native speaker (M. G. Abusov) and an experienced linguist (P. A. Saidova)
- Mostly Botlikh with some notes on Miarso

(Alekseev and Azaev 2019)

- Compiled in the 1960s / 1970s by a native speaker / philologist (X. G. Azaev) and later (in the 2000s) systematized by an experienced linguist (M. E. Alekseev)
- Subsequently edited by T. A. Maisak and scheduled for posthumous publication last year
- Botlikh only

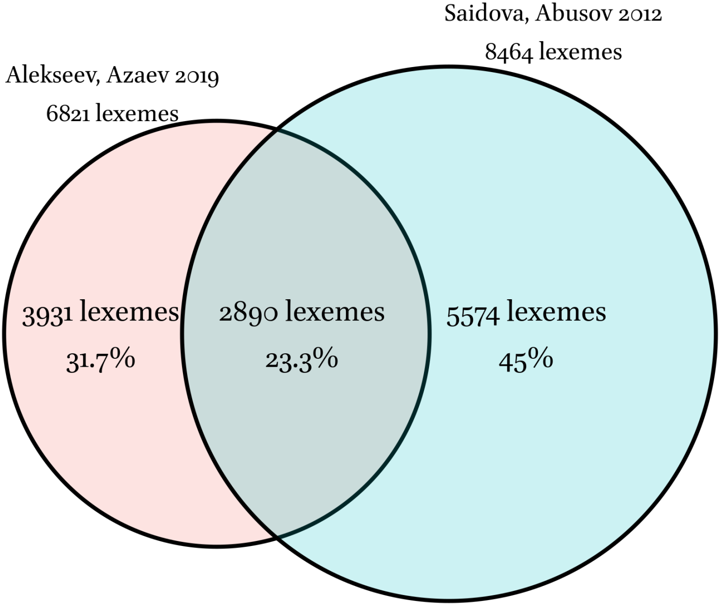
Summary:

- Dictionaries were compiled independently of each other
- with no metadata on the speakers consulted
- data collection was separated with several decades break

Comparing two Botlikh dictionaries: data preparation

- Automatically merge two `.doc` file into one unified `.xls` file
- Manually check for similarities (S. Verhees, C. Naccarato and me)

Comparing two Botlikh dictionaries: data preparation



Comparing two Botlikh dictionaries: results

- There are 1996 lexemes which look phonetically the same, and 909 are different (31%)
- If we remove the stress sign, there are 2449 lexemes which look phonetically the same, and 456 are different (16%)
- ⇒ 15% of lexemes have different stress pattern?..

Comparing two Botlikh dictionaries: results

- There are 1996 lexemes which look phonetically the same, and 909 are different (31%)
- If we remove the stress sign, there are 2449 lexemes which look phonetically the same, and 456 are different (16%)
- ⇒ 15% of lexemes have different stress pattern?.. Yes, but including 265 (9%) cases where stress is present in one dictionary and absent in the other.

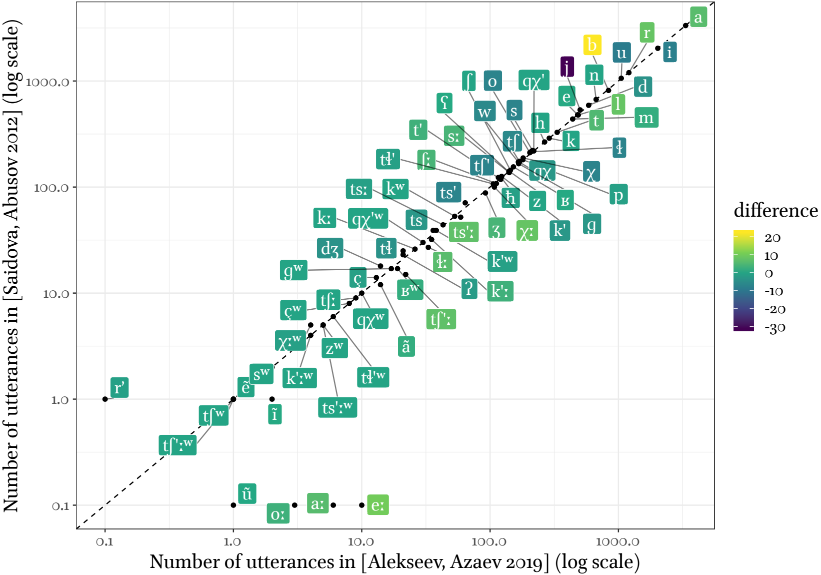
Comparing two Botlikh dictionaries: results

- There are 1996 lexemes which look phonetically the same, and 909 are different (31%)
- If we remove the stress sign, there are 2449 lexemes which look phonetically the same, and 456 are different (16%)
- ⇒ 15% of lexemes have different stress pattern?.. Yes, but including 265 (9%) cases where stress is present in one dictionary and absent in the other.
- What causes the difference between dictionaries?
  - Stress pattern differences in 188 lexemes (about 6%)
  - Multiple cases where there is a small difference that could be explained either as a typo or in terms phonological variation: *čuhí* ‘to run’ [aa] vs. *čũhí* [sa], *kusu* ‘cherry plum’ [aa] vs. *kusːu* [sa]
  - Multiple cases where Russian borrowings were adopted differently: *awtobus* ‘bus’ [aa] vs. *abtabus* [sa], *biton* ‘milk can’ [aa] vs. *bitun* [sa], *apteka* ‘pharmacy’ [aa] vs. *abteka* [sa]
  - Morphological preferences: *dinija=**w*** ‘pious’ [aa] vs. *dinija=**b*** [sa]

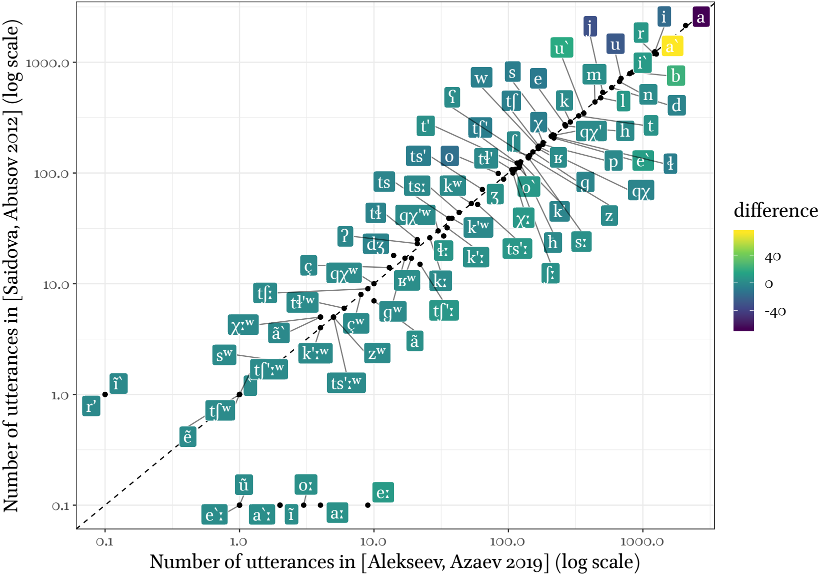
Comparing two Botlikh dictionaries: results

(Alekseev and Azaev 2019)	(Saidova and Abusov 2012)	
ãhajr	*ãhar*	’message’
beʒajr	*beʒir*	’roasting’
mik’kujr	*mik’ːur*	’swallowing’
reqχujr	*reqχwir*	’fight’
reʃkujr	*reʃkur*	’overnight stay’
rikwajr	*rikwar*	’lighting’
...	...	...
χwardar	*χwardir*	’digging’
miʔar	*miʔar*	’nose’
...	...	...
ʃːalaj	*ʃːallaj*	’silt’
inuʕala	*inuʕalla*	’everywhere’
ʕila	*ʕilla*	’reason’

Comparing two Botlikh dictionaries: without stress



Comparing two Botlikh dictionaries: with stress

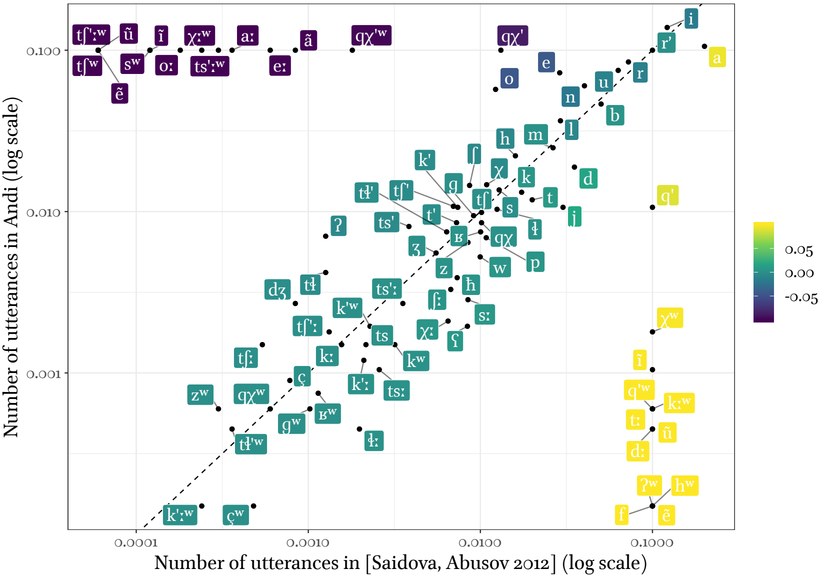


Zilo Andi data

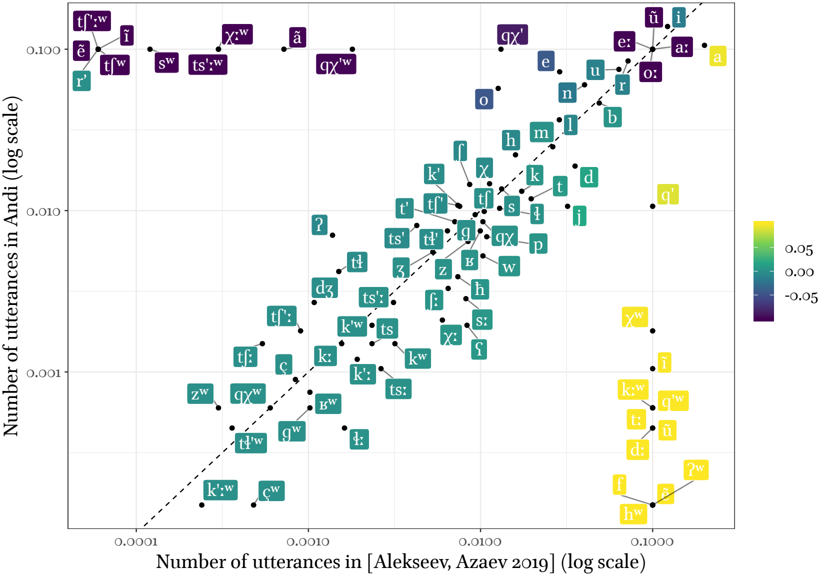
Dictionary data for Zilo were collected during fieldtrips to Zilo in 2016–2019 with N. Rochant, S. Verhees, A. Martynova and A. Zakirova who contributed to the same FieldWorks project.

- Contain morphological affixes
- Doesn't contain additional affixes in a lemma form
- Contain different stems of the same lexeme (e. g. SG.ABS, SG.OBL, PL.ABS, PL.OBL, PST, NPST). Those forms were removed during the analysis.
- No information about stress

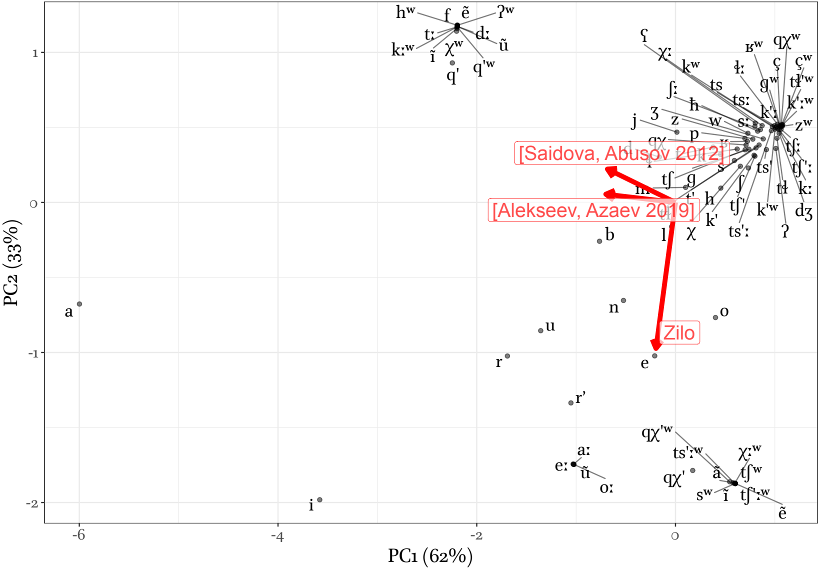
Comparing (Saidova and Abusov 2012) and Zilo



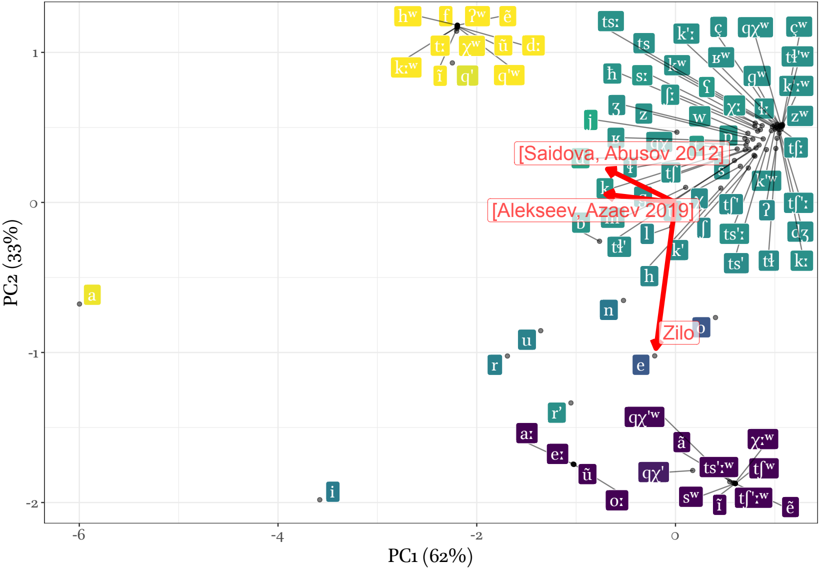
Comparing (Alekseev and Azaev 2019) and Zilo



Comparing Botlikh and Zilo: PCA analysis

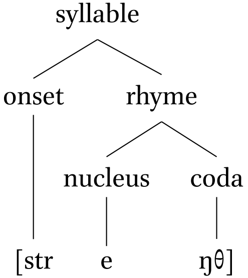


Comparing Botlikh and Zilo: PCA analysis



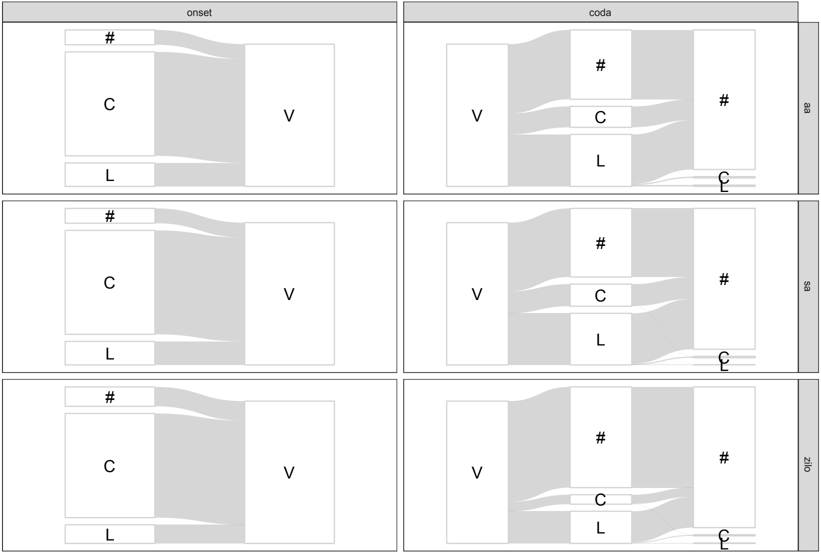
Lets color segments according difference between [aa] and Zilo

Syllable structure: algorithm



- analyse all onsets of initial syllables in corpus
- analyse all codas of final syllables in corpus
- generalize obtained initials and codas into a syllable model
- check, whether this model describes all intervocal consonant clusters

Syllable structure: results



C — obstruent, L — sonorant, V — vowel, # — syllable boundary

Known problems

- Frequency analysis is not a novel approach: you can find it in (Bloomfield 1933) and probably among other scholars
- Botlikh dictionaries were specially selected for shared meaning, the same procedure for the Andic dictionary was not done
- Botlikh dictionaries contain a lot of borrowings, this is not true for the Andic dictionary
- Lemmata are not the same as wordforms, so the model should be checked with the wordform material
- Lemmata can contain some affix that will shift all frequencies (e. g. INF, PL or =CL) for some types of phonological units
- It would be nice to compare the obtained models with the models built on corpora data, when/if it will be available

Further work

- Now within this project we are adding other Avar-Andic languages (Avar, Godoberi, Karata, Tindi, Chamalal, Bagvalal, Axvax)

Further work

- Now within this project we are adding other Avar-Andic languages (Avar, Godoberi, Karata, Tindi, Chamalal, Bagvalal, Axvax)
- This approach could be typologically scaled:
    - it is possible to apply it to 329 dictionary lists from IDS
    - as a result you will get what is typical and what is rare in languages of the world

Further work

- Now within this project we are adding other Avar-Andic languages (Avar, Godoberi, Karata, Tindi, Chamalal, Bagvalal, Axvax)
- This approach could be typologically scaled:
    - it is possible to apply it to 329 dictionary lists from IDS
    - as a result you will get what is typical and what is rare in languages of the world
- This approach is not restricted to one type of phonological units:
    - segments
    - syllables
    - sound alternations
    - phonotactic rules
    - ...

Further work

- Now within this project we are adding other Avar-Andic languages (Avar, Godoberi, Karata, Tindi, Chamalal, Bagvalal, Axvax)
- This approach could be typologically scaled:
  - it is possible to apply it to 329 dictionary lists from IDS
  - as a result you will get what is typical and what is rare in languages of the world
- This approach is not restricted to one type of phonological units:
  - segments
  - syllables
  - sound alternations
  - phonotactic rules
  - ...
- It is possible to apply known NLP technics:
  - language attribution (< authorship attribution);
  - measure distance between languages and phonological units (< vector models from distributional semantics);
  - model could be extended with Markov Chains (what is probability of the sequence *p-a* in a languages of the world?)

References

Alekseev, M. and Azaev, X. (2019). *Botlixsko-russkij slovar’* [*Botlikh-Russian dictionary*]. Moscow: Academia.

Bloomfield, L. (1933). *Language*. New York: Henry Holt.

Gudava, Togo, E. (1962). *Botlixuri ena* [*The Botlikh language*]. Tbilisi: Mecniereba.

Kibrik, A. E. and Kodzasov, S. V. (1988). *Sopostavitelnoye izucheniye dagestanskikh yazykov* [*Comparative study of Daghestanian languages*]. Moscow State University, Moscow.

Moroz, G., A. (2018). lingphonology: automatic phonological description. R package draft.

Moroz, G., A. (2019). Slogovaya struktura adygeyskogo yazika: ot dannyx k obosheniyam [Adyghe syllable structure: From empirical data to generalizations]. *Voprosy Jazykoznanija*, 2:82–95.

References

Moroz, G. (2017). *lingtypology: easy mapping for Linguistic Typology*.

Romanova, K., I. (2019). Automatic Syllable Structure Extracting From Dictionaries: Slavic Data. Term paper.

Saidova, P. A. and Abusov, M. G. (2012). *Botlixsko-russkij slovar'* [*Botlikh-Russian dictionary*]. Makhachkala: IJaLI.

Salimov, X. S. (2010 (1968)). *Gagatlinskij govor andijskogo jazyka* [*The Gagatli dialect of the Andi language*]. Makhachkala.

Suleymanov, J. G. (1957). *Grammatičeskij očerk andijskogo jazyka* (*po dannim govora s. Rikvani*) [*Grammar sketch of the Andi language* (*based on material from the dialect of the village Rikvani*)]. PhD thesis, Institut Jazykoznania AN SSSR.

Tsertsvadze, I. I. (1965). *Andiuri Ena*. Tbilisi: Metsniereba.