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

G. Moroz

Linguistic Convergence Laboratory, NRU HSE, 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.)

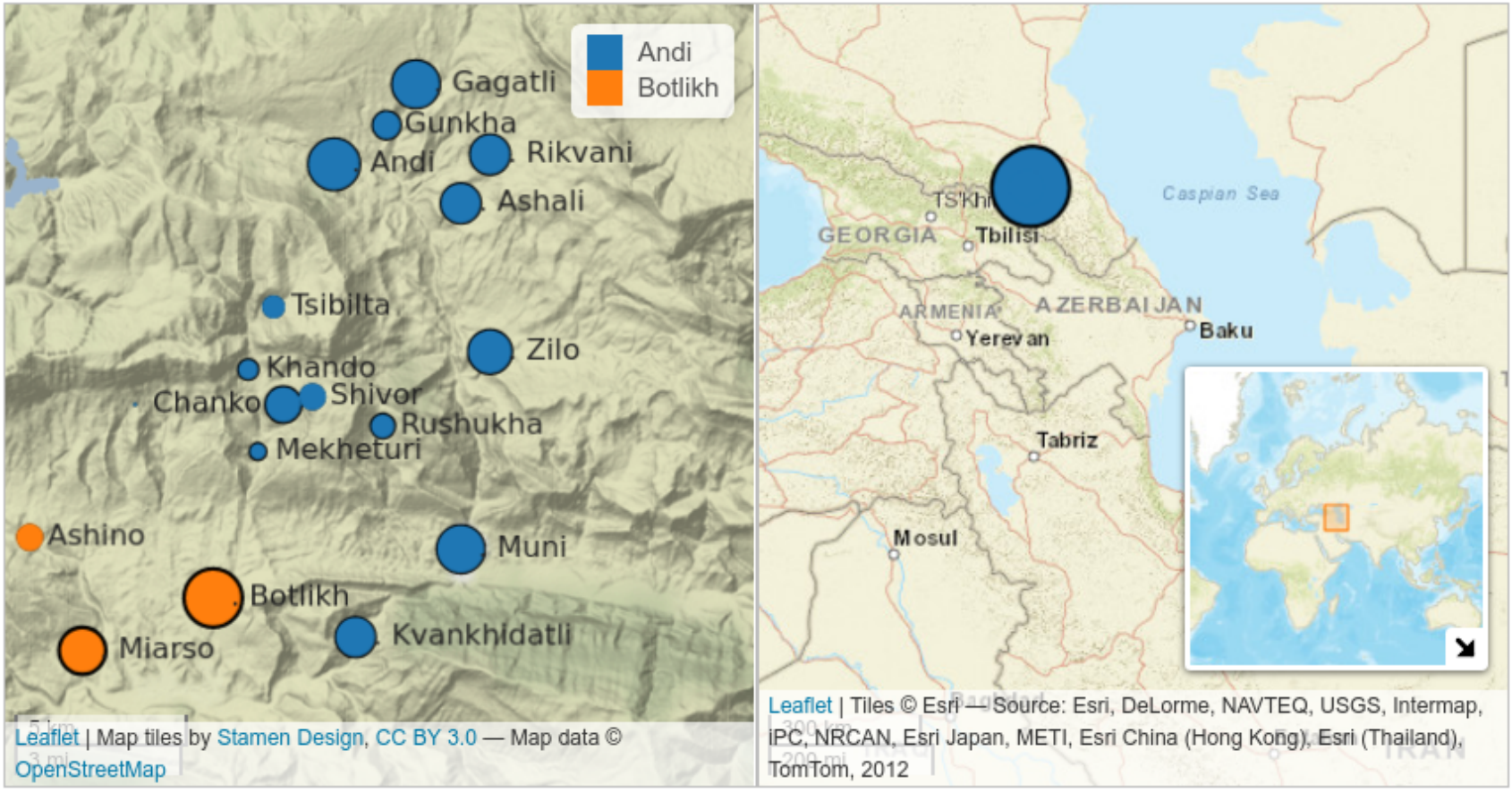
Further steps:

- 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



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

Ссылки

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.

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

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