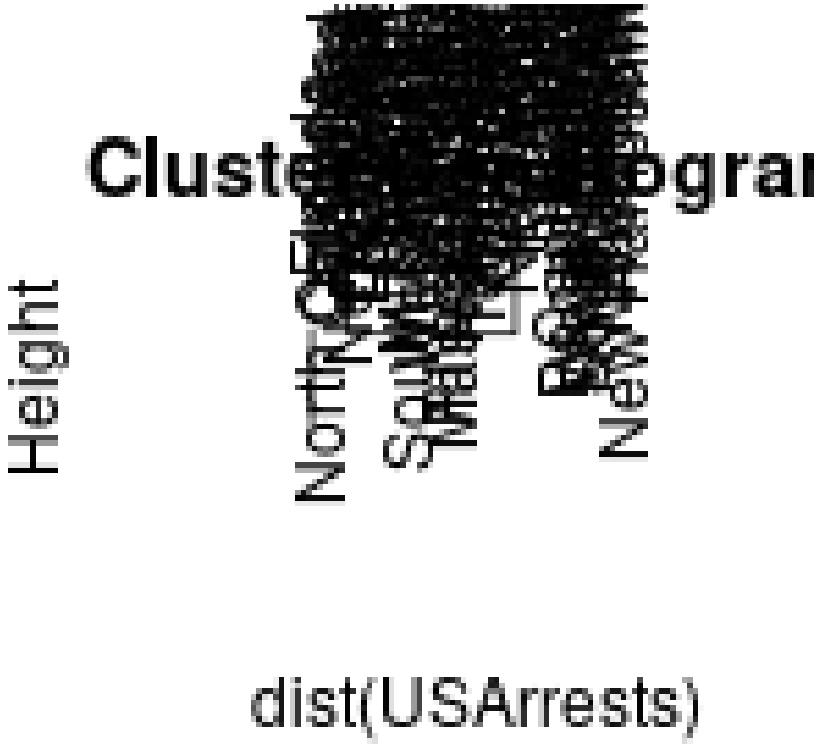
Comparative Andic dictionary database: about its creation and prelimanary results

George Moroz

Linguistic Convergence Laboratory (HSE University)

June 15, 2021

presentation is available here: tinyurl.com/yfbjgcc7



Overview

About the database

Phonological distances between languages

About the database: participants

On May 23 we released a first version of Comparative Andic dictionary database — collection of digitized dictionaries of Andic languages.

https://github.com/phon-dicts-project/comparative_andic_dictionary_database

About the database: participants

On May 23 we released a first version of Comparative Andic dictionary database — collection of digitized dictionaries of Andic languages.

https://github.com/phon-dicts-project/comparative_andic_dictionary_database

We worked during the last two years combining the database creation with term paper and other projects:

- Arseniy Averin (2 years)
- Anastasia Davidenko (2 years)
- Zlata Shkutko (2 years)
- Grigory Kuznetsov (2 years)
- Ilya Sadakov (1 year)
- Anna Tsysova (1 year)
- Wanshu Zhang (1 year)
- Samira Verhees (1 year)
- Chiara Naccarato (1 year)

About the database: contents

Number of lemmata per dictionary:

language	glottocode	reference	lemmata
Akhvakh	akhv1239	Magomedova, Abdulayeva 2007	7796
Andi	andi1255	Salimov 2010	5852
Bagvalal	bagv1239	Magomedova 2004	7881
Botlikh	botl1242	Alekseev 2006	8286
Botlikh	botl1242	Saidova, Abusov 2012	6601
Chamalal	cham1309	Magomedova 1999	7025
Godoberi	ghod1238	Saidova 2006	5640
Karata-Tukita	kara1474	Magomedova, Khalidova 2001	5154
Tindi	tind1238	Magomedova 2003	7779
Tokita	toki1238	Magomedova, Khalidova 2001	217

About the database: contents

Number of meanings per dictionary[1]:

language	glottocode	reference	meaning
Akhvakh	akhv1239	Magomedova, Abdulayeva 2007	14007
Andi	andi1255	Salimov 2010	6144
Bagvalal	bagv1239	Magomedova 2004	12706
Botlikh	botl1242	Alekseev 2006	10612
Botlikh	botl1242	Saidova, Abusov 2012	9068
Chamalal	cham1309	Magomedova 1999	8496
Godoberi	ghod1238	Saidova 2006	7423
Karata-Tukita	kara1474	Magomedova, Khalidova 2001	6651
Tindi	tind1238	Magomedova 2003	12724
Tokita	toki1238	Magomedova, Khalidova 2001	217

[1]Except Chamalal and Botlikh: those dictionaries have not been split yet.

About the database: contents

- `id_word`: 9
- `id_meaning`: 1
- `id`: 11
- `lemma`: а’ва
- `ipa`: ’a-w-a
- `morphology`: (-лъилIи / -лIи, /ди)
- `bor`: _
- `pos`: noun
- `meaning_ru`: дом
- `definition`: 1) дом, здание; *ава гурулъIа* строить дом 2) этаж; *кьIасе ава* верхний этаж; *гекьисе ава* нижний этаж; *цег., тлян. авал, ратл. авали*
- `glottocode`: akhv1239
- `reference`: Magomedova, Abdulayeva 2007

About the database: contents

- `id_word`: 9
- `id_meaning`: 2
- `id`: 12
- `lemma`: а’ва
- `ipa`: ’a-w-a
- `morphology`: (-лъилIи / -лIи, /ди)
- `bor`: _
- `pos`: noun
- `meaning_ru`: этаж
- `definition`: 1) дом, здание; *ава гурулъIа* строить дом 2) этаж; *кьIаце ава* верхний этаж; *гекьице ава* нижний этаж; *цег., тлян. авал, ратл. авали*
- `glottocode`: akhv1239
- `reference`: Magomedova, Abdulayeva 2007

About the database: contents

- `id_word`: 17
- `id_meaning`: 1
- `id`: 21
- `lemma`: ава’рийа
- `ipa`: a-w-’a-r-i-j-a
- `morphology`: (-лIи, -ди)
- `bor`: 1
- `pos`: noun
- `meaning_ru`: авария
- `definition`: авария (*дорожное происшествие*); *аварийа-лъ̄ига бухьурулъIа* попасть в аварию
- `glottocode`: akhv1239
- `reference`: Magomedova, Abdulayeva 2007

About the database: purpose

What can be done with this database?

- calculate frequency of phonological units and compare them across Andic languages
- use some modern tools like Edictor [List 2017] for automatic analysis of sound correspondences
- annotate concepts from Concepticon [List et al. 2021] with `pyconcepticon` and use some tools for investigation of the colexification (see the CLICS project [Rzymski et al. 2020])
- the database also could be a good ground for selecting words for a phonetic or any other research

Overview

About the database

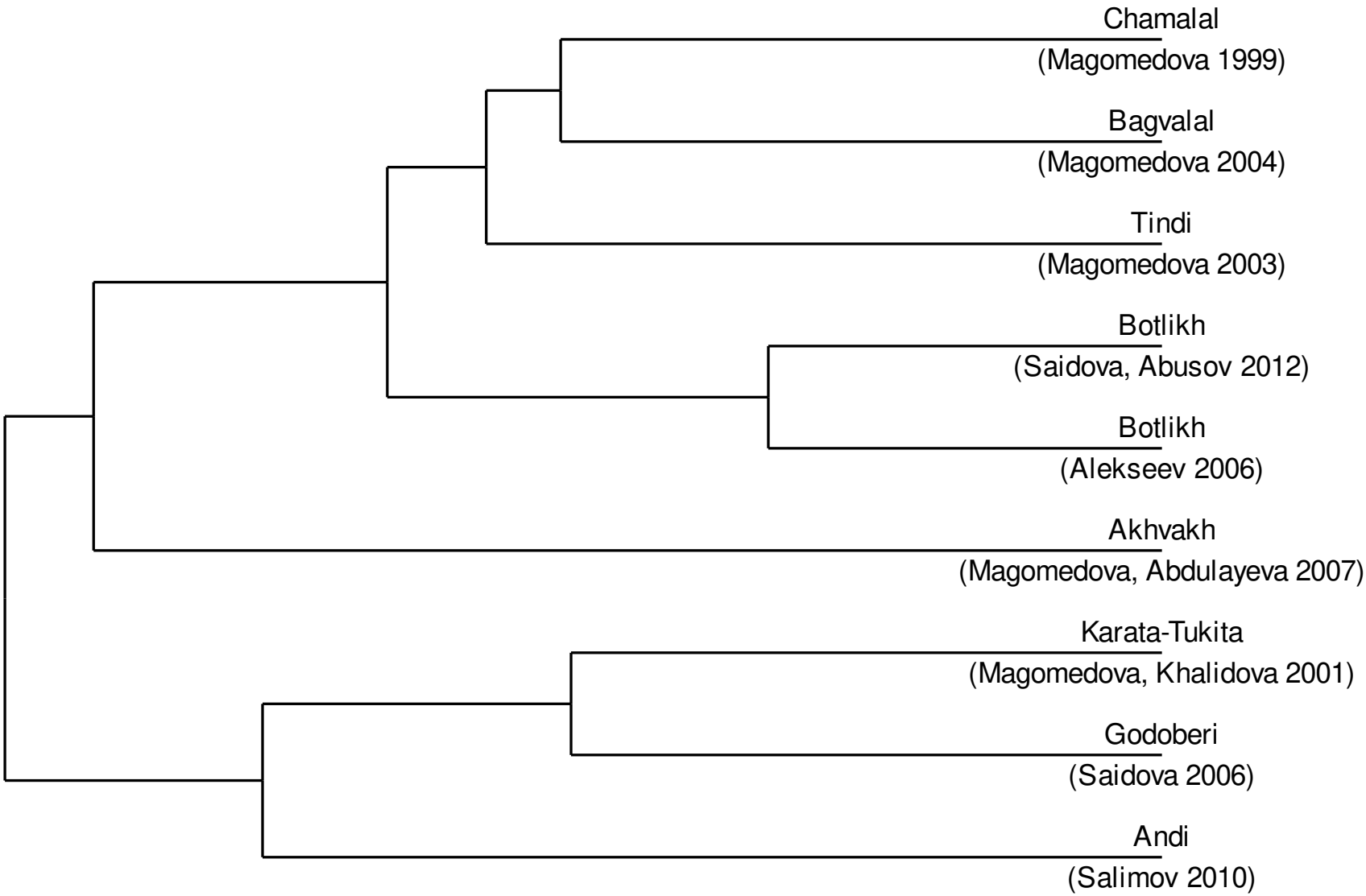
Phonological distances between languages

Phonological distances

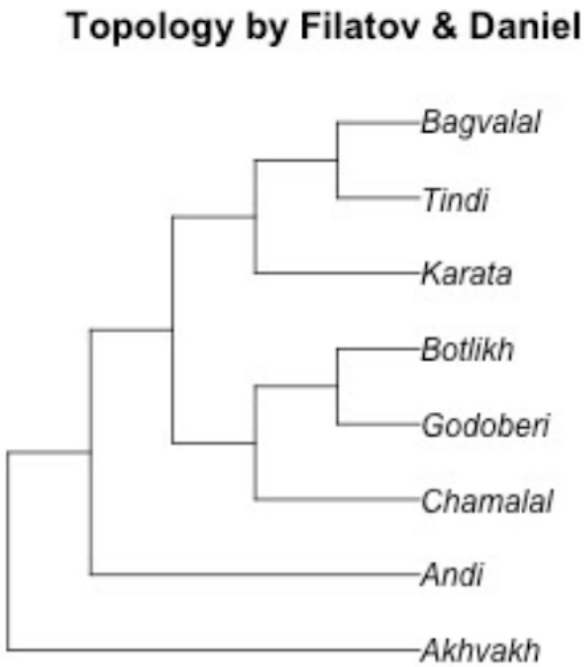
In order to calculate phonological distances I used the following algorithm:

- remove borrowings and Tokita data;
- remove the stress sign;
- remove duplicated IPA transcriptions for each language;
- extract frequencies of each segments and use in hierarchical clusterization

Phonological distances: results



Phonological distances: phylogenetic data



Obtained phonological distances does not correlate with phylogenetic distances.

Can dictionary data be trusted?

- There are some morphemes that could increase freaquency of certain segments
- However, in [Давиденко 2021] we compared dictionary and corpora frequencies of Andi, Botlikh and Bagvalal phonological segments and found a linear relation between them:

$$\text{corpora frequency} = 0.002 + 0.906 \times \text{dictionary frequency}$$

Thank you for your attention!

References

List, J.-M. (2017). A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*, pages 9–12, Valencia. Association for Computational Linguistics.

List, J. M., Rzymski, C., Greenhill, S., Schweikhard, N., Pianykh, K., Tjuka, A., Hundt, C., and Forkel, R., editors (2021). *Concepticon 2.5.0*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

References

Rzymski, C., Tresoldi, T., Greenhill, S. J., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., Gast, V., Bodt, T. A., Hantgan, A., Kaiping, G. A., et al. (2020). The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific data*, 7(1):1–12.

Давиденко, □. □. (2021). Сравнение фонологических систем, полученных на основе словарей и корпусов: данные андийских языков.