Comparative Andic dictionary database

George Moroz

Linguistic Convergence Laboratory, HSE University, Moscow

4 October 2022, Surrey Morphology Group

Preamble

Outline of the talk

About me

Andic languages

References for Andic languages

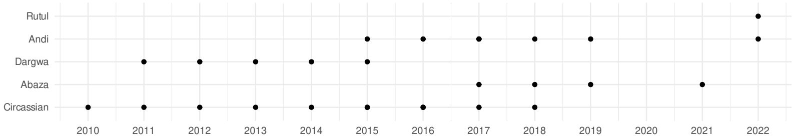
Comparative Andic dictionary database

Preliminary results

Conclusions

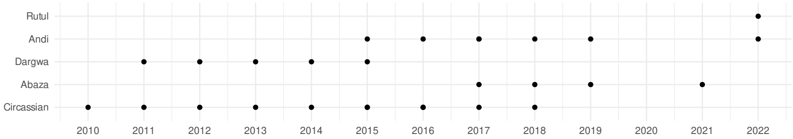
About me

- Field Linguistics
  - Language documentation
  - Phonetics
  - Phonology
  - East and West Caucasian languages



About me

- Field Linguistics
    - Language documentation
    - Phonetics
    - Phonology
    - East and West Caucasian languages



- Computer linguistics
    - R packages: `lingtypology`, `phonfieldwork`, `lingglosses`
    - Quantitative Linguistics
    - Morphological Transducers

Outline of the talk

About me

Andic languages

References for Andic languages

Comparative Andic dictionary database

Preliminary results

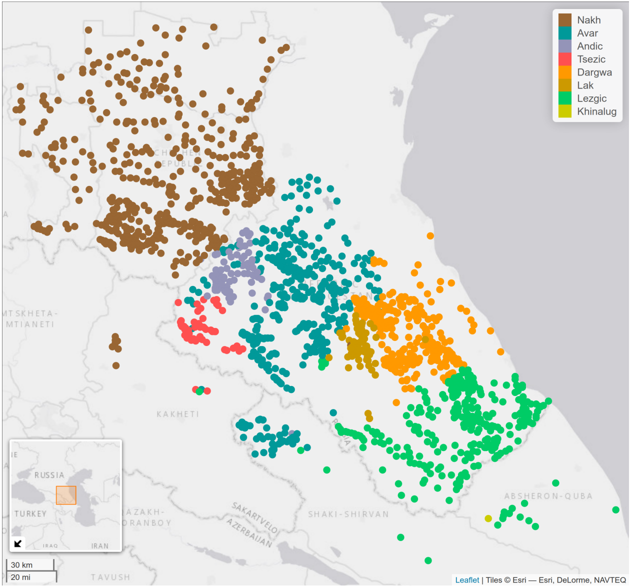
Conclusions

East Caucasian (Nakh-Dagestanian)

- Nakh languages
- Khinalug language
- Lezgic languages
- Lak language
- Dargwa (Dargic) languages
- Tsezic (Didoic) languages
- Avar language
- Andic languages (Dobrushina et al. 2020; Koryakov et al. 2022):

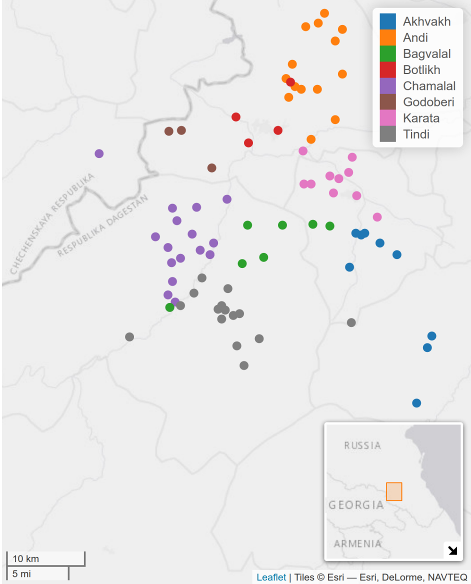
Andi (and Lower Andi)	andi1255	22 500	ɢʷannab mits’ːi
Botlikh	botl1242	7 400	bujχaɬi mits’ːi
Godoberi	ghod1238	3 200	ʁibditɬi mitsːi
Karata	kara1516	11 000	k’ːirtɬi mats’ːi
Tukita	toki1238	1 300	
Northern Akhvakh	nort3330	9 500	aʃʷatɬi mits’ːi
Southern Akhvakh	sout3319	8 000	
Bagvalal	bagv1239	5 500	bagʷalal mis’ː
Tindi	tind1238	9 300	idarab mitsːi
Chamalal (and Gigatli)	cham1309	9 600	tʃ’amalaldub mits’ː

East Caucasian (Moroz and Verhees 2020)



Created with `lingtypology` (Moroz 2017).

Andic languages (Moroz and Verhees 2020)

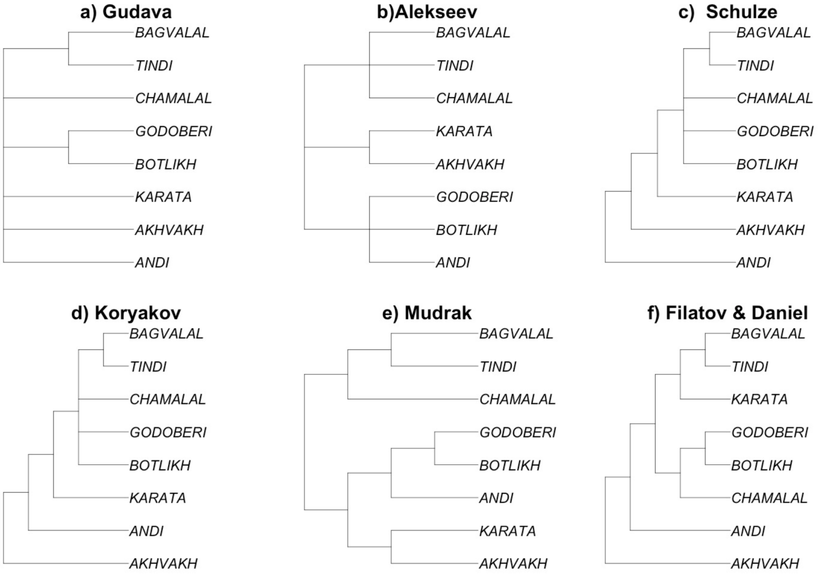


Created with `lingtypology` (Moroz 2017).

Classification of Andic languages

- It has been suggested that Andic languages as a whole may in fact represent a language continuum (Gudava 1967).
- Some village varieties of what is traditionally considered one language may be highly divergent from its other varieties.

Classification of Andic languages (Koile et al. 2022)



Outline of the talk

About me

Andic languages

References for Andic languages

Comparative Andic dictionary database

Preliminary results

Conclusions

References for Andic languages

language	grammar	dictionary	(Kibrik et al.)	(Key et al.)
Andi	+	-	+	+
Lower Andi	-	-	-	+
Botlikh	+	++	-	+
Godoberi	+	+	-	+
Karata	+	+	-	+
Tukita	-	∓	-	+
N. Akhvakh	+	+	+	+
S. Akhvakh	-	-	-	+
Bagvalal	+	+	-	+
Tindi	+	+	-	+
Chamalal	+	+	+	+
Gigatli	-	-	+	+

- all Andic dictionaries for IDS were provided by Madzhid Khalilov.

Pros and cons of Andic part of IDS (Key and Comrie 2021)

- the best language and lexicon coverage (more than 1300 meanings)

Pros and cons of Andic part of IDS (Key and Comrie 2021)

- the best language and lexicon coverage (more than 1300 meanings)
- unified transcription

Pros and cons of Andic part of IDS (Key and Comrie 2021)

- the best language and lexicon coverage (more than 1300 meanings)
- unified transcription
- absence of anything but translation

Pros and cons of Andic part of IDS (Key and Comrie 2021)

- the best language and lexicon coverage (more than 1300 meanings)
- unified transcription
- absence of anything but translation
- a lot of typos (p. c. by several East Caucasian researchers)
  - Andi: ɢannni ‘crow’ vs. ɢʷaɢʷal ‘nut’ (Key and Comrie 2021)
  - Andi: ʁannni ‘crow’ vs. ɢʷaɢʷal ‘nut’ (Kibrik and Kodzasov 1990) and my personal fieldwork data
  - Botlikh: ts’ik’ːu ‘sour’ (Key and Comrie 2021) vs ts’ːik’ːu ‘sour’ (Saidova and Abusov 2012; Alekseev and Azaev 2019)

Pros and cons of Andic part of IDS (Key and Comrie 2021)

- the best language and lexicon coverage (more than 1300 meanings)
- unified transcription
- absence of anything but translation
- a lot of typos (p. c. by several East Caucasian researchers)
  - Andi: ɢannni ‘crow’ vs. ɢʷaɢʷal ‘nut’ (Key and Comrie 2021)
  - Andi: ʁannni ‘crow’ vs. ɢʷaɢʷal ‘nut’ (Kibrik and Kodzasov 1990) and my personal fieldwork data
  - Botlikh: ts’ik’ːu ‘sour’ (Key and Comrie 2021) vs ts’ːik’ːu ‘sour’ (Saidova and Abusov 2012; Alekseev and Azaev 2019)
- contradictions with conventional dictionaries
  - Botlikh: beχiq’al ‘hip’ (Key and Comrie 2021) vs aqχu ‘hip’ (Saidova and Abusov 2012)

Pros and cons of Andic part of IDS (Key and Comrie 2021)

- the best language and lexicon coverage (more than 1300 meanings)
- unified transcription
- absence of anything but translation
- a lot of typos (p. c. by several East Caucasian researchers)
    - Andi: ɢannni ‘crow’ vs. ɢʷaɢʷal ‘nut’ (Key and Comrie 2021)
    - Andi: ʁannni ‘crow’ vs. ɢʷaɢʷal ‘nut’ (Kibrik and Kodzasov 1990) and my personal fieldwork data
    - Botlikh: ts’ik’ːu ‘sour’ (Key and Comrie 2021) vs ts’ːik’ːu ‘sour’ (Saidova and Abusov 2012; Alekseev and Azaev 2019)
- contradictions with conventional dictionaries
    - Botlikh: beχiq’al ‘hip’ (Key and Comrie 2021) vs aqχu ‘hip’ (Saidova and Abusov 2012)
- one word per meaning (in most cases)
    - Godoberi: χʷaji ‘dog’ (Key and Comrie 2021) vs. χʷaji ‘dog’ and baħri ‘dog’ (Saidova 2006)

Outline of the talk

About me

Andic languages

References for Andic languages

Comparative Andic dictionary database

Preliminary results

Conclusions

Comparative Andic dictionary database v. 0.6

— collection of digitized dictionaries of Andic languages.

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

A lot of people contributed and continue contributing to the database: A. Averin, D. Chistyakova, K. Chuprinko, A. Davidenko, M. Dolgodvorova, N. Fedorenko, T. Gnedina, G. Kuznetsov, C. Naccarato, G. Moroz, I. Sadakov, Z. Shkutko, A. Tsyzova, S. Verhees.

There are 93875 rows in the database.

Comparative Andic dictionary database v. 0.6

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

Comparative Andic dictionary database v. 0.6

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

Comparative Andic dictionary database v. 0.6

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

Comparative Andic dictionary database v. 0.6

Number of lemmata per dictionary:

language	reference	lemmata
Andi	Salimov 2010	5850
Bagvalal	Magomedova 2004	7881
Botlikh	Alkseev, Azayev 2019	8273
Botlikh	Saidova, Abusov 2012	6597
Chamalal	Magomedova 1999	7018
Godoberi	Saidova 2006	5640
Karata	Magomedova, Khalidova 2001	5153
N. Akhvakh	Magomedova, Abdulayeva 2007	7795
Tindi	Magomedova 2003	7779
Tukita	Magomedova, Khalidova 2001	218

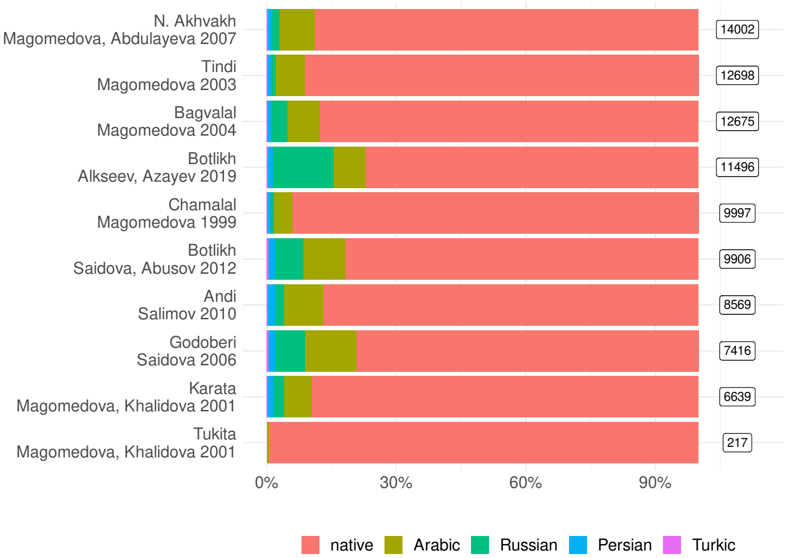
Comparative Andic dictionary database v. 0.6

Number of meanings per dictionary:

language	reference	meanings
Andi	Salimov 2010	8585
Bagvalal	Magomedova 2004	12707
Botlikh	Alkseev, Azayev 2019	11547
Botlikh	Saidova, Abusov 2012	9963
Chamalal	Magomedova 1999	10009
Godoberi	Saidova 2006	7450
Karata	Magomedova, Khalidova 2001	6650
N. Akhvakh	Magomedova, Abdulayeva 2007	14020
Tindi	Magomedova 2003	12726
Tukita	Magomedova, Khalidova 2001	218

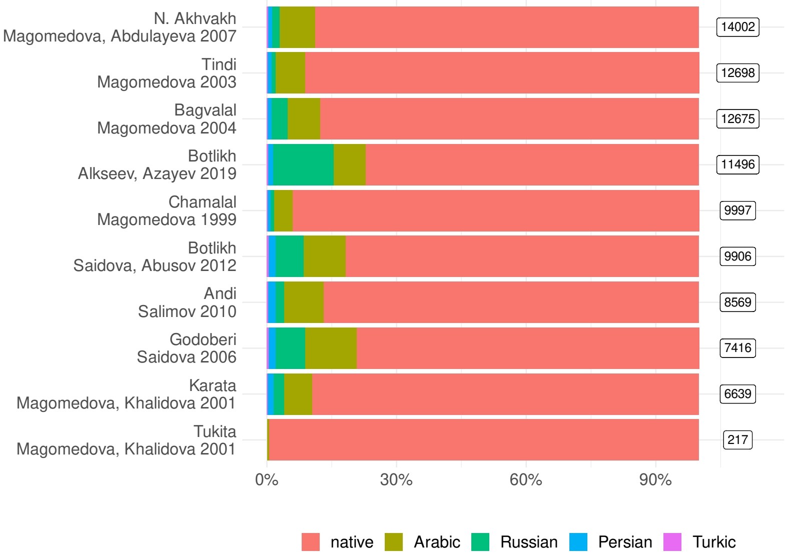
Comparative Andic dictionary database v. 0.6

Number of borrowings per dictionary:



Comparative Andic dictionary database v. 0.6

Number of borrowings per dictionary: (Avar annotation needed)



Outline of the talk

About me

Andic languages

References for Andic languages

Comparative Andic dictionary database

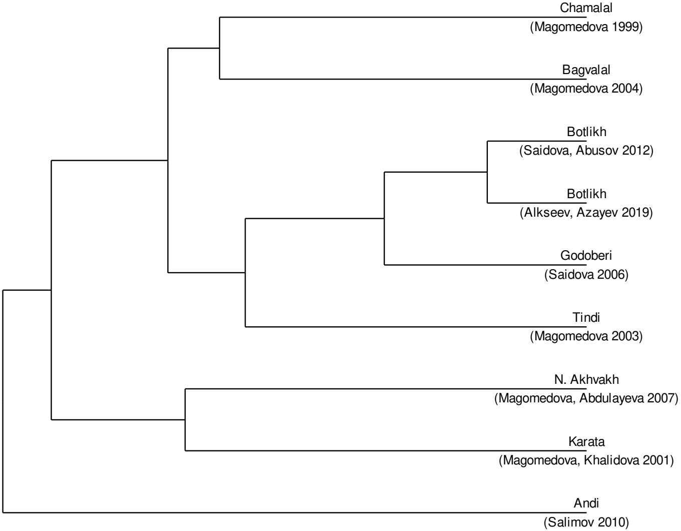
Preliminary results

Conclusions

Preliminary results: phonological distance

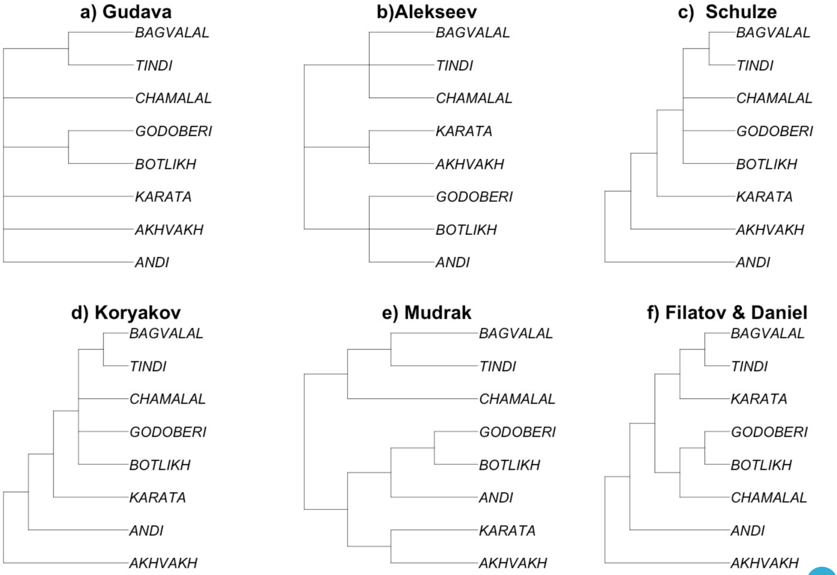
- remove Tukita
- remove borrowings
- remove the stress sign
- select meanings shared across all dictionaries
- calculate frequencies of each segment
- use the obtained frequencies for distance calculation
- apply hierarchical clustering to the gathered distances

Preliminary results: phonological distance

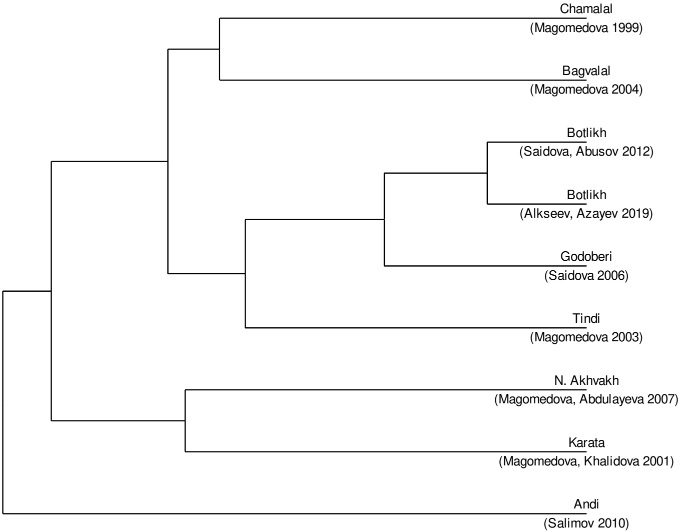


Based on 453 shared meanings
from Comparative Andic dictionary database v. 0.6

Classification of Andic languages (Koile et al. 2022)

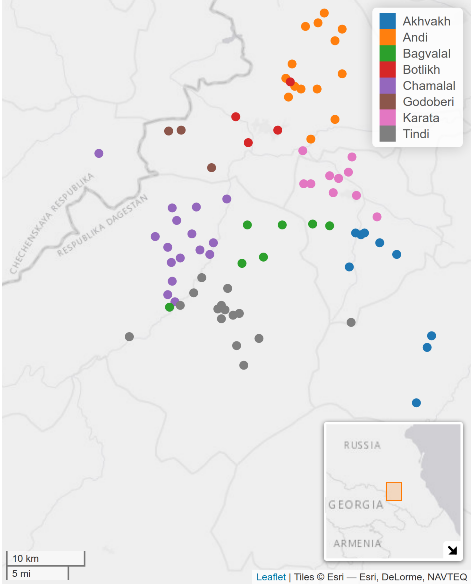


Preliminary results: phonological distance



Based on 453 shared meanings
from Comparative Andic dictionary database v. 0.6

Andic languages (Moroz and Verhees 2020)



Created with `lingtypology` (Moroz 2017).

Can dictionary data be trusted?

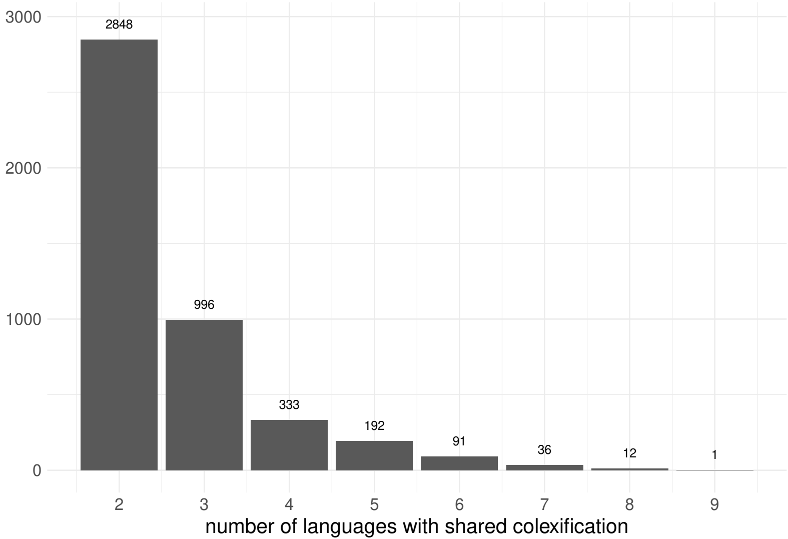
- There are some morphemes that could increase frequency of certain segments (inf, adjectivizer).
- There are a lot of Avar borrowings that are not annotated.
- However, in (Davidenko 2021) we compared dictionary and corpora segment frequencies of Andi, Botlikh and Bagvalal phonological segments and found a linear relation between them:

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

Preliminary results: colexification (see the CLICS project (Rzymski et al. 2020))

- ‘COTTON’ and ‘COTTON WOOL’ (9 languages)
- ‘WRITE’ and ‘PLAY (musical instruments)’ (8 languages)
- ‘NATURE’ and ‘CHARACTER, TEMPER’ (8 languages)
- ‘HOLD’ and ‘TRAP (CATCH)’ (8 languages), see CLICS
- ‘REFUSE’ and ‘DENY’ (8 languages), see CLICS
- ‘PAPER’ and ‘LETTER’ (8 languages), see CLICS
- ‘(STRING) THREAD’ and ‘WHETSTONE’ (8 languages)
- ‘DRESS’ and ‘SHIRT’ (8 languages), see CLICS
- and about 4500 other cases.

Preliminary results: colexification



Based on Comparative Andic dictionary database v. 0.6

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 meanings using 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
- compare morphological patterns across languages
- extract language examples and use them as a corpus

Outline of the talk

About me

Andic languages

References for Andic languages

Comparative Andic dictionary database

Preliminary results

Conclusions

Conclusions

- Digitalization provides a lot of new data for possible research
  - However it take a lot of effort to create and curate such a database
- Phonological distance can reveal new information and pose new questions
  - See (Cysouw and Forker 2009) for the similar approach to spatial case marking in Tsezic
- There is a lot that we do not know about colexification in East Caucasian languages

References

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

Michael Cysouw and Diana Forker. Reconstruction of morphosyntactic function: Nonspatial usage of spatial case marking in Tsezic. *Language*, pages 588–617, 2009.

A. V. Davidenko. Sravnenie fonologičeskih sistem, polučennyh na osnove slovarej i korpusov: dannye andijskih yazykov [dictionary and corpora segment frequencies: Andic data]. 2021.

N. Dobrushina, M. Daniel, and Y. Koryakov. Languages and sociolinguistics of the Caucasus. In M. Polinsky, editor, *The Oxford Handbook of Languages of the Caucasus*. Oxford University Press, 2020.

References

T. E. Gudava. *Andiyskiye Yaziki: Vvedenie.* [*Andic Languages: an Introduction*]. Nauka, Moscow, 1967.

M. R. Key and B. Comrie, editors. *IDS*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2021. URL https://ids.clld.org/.

A. E. Kibrik and S. V. Kodzasov. *Sopostavitelnoye izucheniye dagestanskix yazykov: Imya. Fonetika* [*Comparative study of languages of Dagestan: Nouns. Phonetics*], volume 2. Moskovskij Gosudarstvennyj Universitet, 1990.

E. Koile, I. Chechuro, G. Moroz, and M. Daniel. Geography and language divergence: The case of andic languages. *PloS One*, 17 (5), 2022.

References

Y. B. Koryakov, T. I. Davidyuk, A. P. Evstigneeva, V. V. Ivanov, M. A. Kade, A. A. Syuryun, and V. S. Haritonov. Project languages of russia, 2022. URL http://jazykirf.iling-ran.ru/groups/ND.%20Andic.shtml.

Johann-Mattis List. 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, 2017. Association for Computational Linguistics. URL http://edictor.digling.org.

Johann Mattis List, Christoph Rzymski, Simon Greenhill, Nathanael Schweikhard, Kristina Pianykh, Annika Tjuka, Carolin Hundt, and Robert Forkel, editors. *Concepticon 2.5.0*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2021. URL https://concepticon.clld.org/.

References

G. Moroz. *lingtypology: easy mapping for Linguistic Typology*, 2017. URL https://CRAN.R-project.org/package=lingtypology.

G. Moroz and S. Verhees. East Caucasian villages dataset (Version v2.0) [Data set], 2020. URL https://doi.org/10.5281/zenodo.5588473.

Christoph Rzymski, Tiago Tresoldi, Simon J Greenhill, Mei-Shin Wu, Nathanael E Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A Bodt, Abbie Hantgan, Gereon A Kaiping, et al. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific data*, 7(1):1–12, 2020.

References

P. A. Saidova. *Godoberinsko-russkij slovar'* [*Godoberi-Russian dictionary*]. Bespis'mennye Jazyki Dagestana : Serija Nacional'no-Russkie Slovari. Rossijskaja Akademia Nauk, Dagestanskij Naucnyj Centr, Machackala, 2006.

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