

A corpus of Andi field recordings

Aigul Zakirova¹

George Moroz¹

Elena Sokur¹

Samira Verhees²

Neige Rochant³

¹HSE University, Russia; ²Independent researcher, the Netherlands;

³Sorbonne Nouvelle University / CNRS

20 September 2022

Tbilisi, VI International Conference: Language and Modern Technologies

Outline of the talk

- Corpora of Linguistic Convergence Laboratory
- An example of Meadow Mari corpus
- Andi language and its data
- Our path towards the corpus

Corpora of Linguistic Convergence Laboratory

Corpora of Linguistic Convergence Laboratory:

<http://lingconlab.ru/>

- Russian Dialect corpora (14)
- Corpora of Bilingual Russian (7)
- Corpora of minority languages of Russia (6)

Corpora of Linguistic Convergence Laboratory:

<http://lingconlab.ru/>

Linguistic Convergence Laboratory eng ru
Resources

All resources

Corpora ▾

Dictionaries ▾

Other Resources ▾

Khislavichi
dialect

Tokens: 260 793

[View](#)

Keba dialect

Tokens: 54 535

[View](#)

Luzhnikovo
dialect

Tokens: 68 666

[View](#)

Lukh and Teza
river basins
dialects

Tokens: 146 350

[View](#)

Malinino
dialect

Tokens: 138 943

[View](#)

Nekhochi
dialect

Tokens: 88 965

[View](#)

OPOCHETSKY
dialects

Tokens: 68 741

[View](#)

Rogovatka
dialect

Tokens: 100 047

[View](#)

Spiridonova
Buda dialect

Tokens: 70 565

[View](#)

Shetnevo and
Makeevo
dialect

Tokens: 58 003

[View](#)

Tserkovnoe
dialect

Tokens: 19 960

[View](#)

Upper Pinega
and Vyga river
basins dialect

Tokens: 70 803

[View](#)

Ustja River
Basin dialects

Tokens: 959 782

[View](#)

Zvenigorod
dialect

Tokens: 68 324

[View](#)

Corpora of Linguistic Convergence Laboratory:

<http://lingconlab.ru/>

Linguistic Convergence Laboratory eng ru

All resourcesCorpora ▾Dictionaries ▾Other Resources ▾

Resources

Dialect corporaCorpora of bilingual RussianCorpora of minority languages of Russia

Bashkortostan Russian

Tokens: ND

View

Beserman Russian

Tokens: 97 216

View

Chuvash Russian

Tokens: 46 307

View

Daghestanian Russian

Tokens: 227 885

View

Karelian Russian

Tokens: 74 014

View

Romani Russian

Tokens: 41 767

View

Yakut Russian

Tokens: 15 139

View

5

Corpora of Linguistic Convergence Laboratory: <http://lingconlab.ru/>

Linguistic Convergence Laboratory eng ru All resources Corpora ▾ Dictionaries ▾ Other Resources ▾
Resources

Dialect corpora	Corpora of bilingual Russian	Corpora of minority languages of Russia	
<div>Abaza</div> <div>Tokens: 3 636</div> <div>View</div>	<div>Adyghe</div> <div>Tokens: ND</div> <div>View</div>	<div>Bashkir</div> <div>Tokens: ~25 000</div> <div>View</div>	<div>Kabardian</div> <div>Tokens: ND</div> <div>View</div>
<div>Khakas</div> <div>Tokens: ~58 000</div> <div>View</div>	<div>Meadow Mari</div> <div>Tokens: ND</div> <div>View</div>		

The spoken corpus of Meadow Mari

- Meadow Mari < Mari < Finno-Ugric < Uralic
- spoken in the Mari El republic and adjacent areas
- ~360.000 speakers (2010 Census)
- endangered, Educational (Ethnologue)
- relatively well-described: grammars, e.g. (Pengitov et al. 1961, Alhoniemi 1993), dictionaries, e.g. <http://marlamuter.com/ru/>

The spoken corpus of Meadow Mari

- Meadow Mari < Mari < Finno-Ugric < Uralic
- spoken in the Mari El republic and adjacent areas
- ~360.000 speakers (2010 Census)
- endangered, Educational (Ethnologue)
- relatively well-described: grammars, e.g. (Pengitov et al. 1961, Alhoniemi 1993), dictionaries, e.g. <http://marlamuter.com/ru/>
- texts come from the village of Saryj Torjal, Mari El republic

The spoken corpus of Meadow Mari: credits

- Anna Volkova
- Aigul Zakirova
- Mikhail Voronov
- Maria Dolgodvorova
- Zinaida Klyucheva
- Svetlana Kokoreva
- Ilya Makarchuk
- Irina Khomchenkova
- Timofey Arkhangelskiy
- Elena Sokur

Example of Meadow Mari corpus

[Meadow Mari Spoken Corpus](#)

EN | RU | ?



● Word #1

Word:

Lemma:

Grammar:

Gloss:

Language/tier: Meadow Mari



speaker: anf
speaker_name: Фёдорова Анна
Николаевна
dialect: моркинско-сернурский
gender: f
place of birth:
year of birth: 1930
education:
year: 2018

Full-text search:

☒ Precise match

Search sentences

Search words / lemmata



Select subcorpus



Search result: 98 occurrences, 94 sentences found in approximately 10 documents.

Биография А. Н. Ф. (часть 2) 2018

а мый кум кече веле **коштына** ыла Реру-Ушчыл...

а мы за три дня доезжали

Биография З. И. Е. (часть 2)

<нрзб> **коштит**...

Очень много ходило...

Биография З. И. Е. (часть 2) 2018

Эмланже гын пеш **коштына** ыле,

коштына
кошташ V
кошт-ына
STEM-NPST.1PL
gr: npst, 1, pl
trans_ru: ходить



1

2

3

4

...

10

Example of Meadow Mari corpus

- created with Tsakorpus (Arkhangelskiy 2019)
- translation (into Russian)
- glosses
- audio
- export to `.xlsx`
- sociolinguistic information

Example of required ELAN .eaf file

	00.000	00:00:01.000	00:00:02.000	00:00:03.000	00:00:04.000
tfp1955f_Transcript [42]	Таче манына ынде		чыла молодёжет погыненна.		
Words@tfp [201]	Таче	манына	ынде	чыла	молодёжет погыненна .
Lemma@tfp [133]	таче	манаш	ынде	чыла	молодёжь погынаш
Gramm@tfp [133]	ADV	V, npst, 1,	ADV	PRO, sg, nom, N, anim, hum, V, pst2, 1, pl	
lex2@tfp [4]					
trans_ru@tfp [133]	сегодня	говорить,	теперь	весь, все	молодёжь собираться
trans_ru2@tfp [5]					
Morph@tfp [133]	таче	ман-ына	ынде	чыла	молодёж-ет погын-енна
Gloss@tfp [133]	STEM	STEM-NP	STEM	STEM	STEM-2SG STEM-PST2.1,
tfp1955f_Translatio [42]	Сегодня теперь		мы, вся молодёжь, собрались.		

- hierarchy of tiers
- time alignment
- stem forms

Pipeline used for the Meadow Mari corpus

- transcribing the texts by a native speaker of Meadow Mari;
- aligning sound and transcription in ELAN;
- proofreading of the texts by linguists;
- automatic glossing of texts with the help of a morphological analyzer
- modifying the analyzer (adding dialect-specific morphemes, correcting mistakes in the previous version of the analyzer)
- manually removing homonymy left after the morphological analysis.

Developing the corpus of Andi

The Andi Language: a sociolinguistic background

- Andic < Avar-Andic < East Caucasian, glottocode [andi1255]
- spoken in several villages of the Botlikh district of Dagestan.
- more than 20,000 speakers of Andi [Aglarov (2002); All-Russian National Census 2010]
- Andi speakers are trilingual in Andi, Avar and Russian
- Avar serves as a lingua franca and is taught in school (Dobrushina et al. 2017)



	Andi	Rikvani	Gagatli	Zilo
sources	(Kibrik and Kodzasov 1988; Alekseev 1999; Dirr 1906; Tsertsvadze 1965);	(Sulejmanov 1957)	(Salimov 2010)	(Kaye et al. rthc)
grammar sketch	+	+	+	+
dictionary	+	-	±	±
morphological parser	± ¹	-	-	-

¹A pilot version of a morphological parser of Andi is presented in (Buntyakova 2022).

- Written texts from grammar sketches, some with translation
- Field recordings from 1 trip to Rikvani, a few trips to Muni and Kwankhidatli, and several trips to Zilo
- Especially the Kwankhidatli dialect is endangered; only known record is a few short texts in (Tsertsvadze 1965)

- Written texts from grammar sketches, some with translation
- Field recordings from 1 trip to Rikvani, a few trips to Muni and Kwankhidatli, and several trips to Zilo
- Especially the Kwankhidatli dialect is endangered; only known record is a few short texts in (Tsertsvadze 1965)
- Recordings were made by different researchers, with different approaches to recording and processing

- Written texts from grammar sketches, some with translation
- Field recordings from 1 trip to Rikvani, a few trips to Muni and Kwankhidatli, and several trips to Zilo
- Especially the Kwankhidatli dialect is endangered; only known record is a few short texts in (Tsertsvadze 1965)
- Recordings were made by different researchers, with different approaches to recording and processing
- We have approximately 8.27 hours of recordings
 - 7.77 hours are transcribed
 - 3 hours are aligned with sound
 - 17.35 minutes are fully glossed

- Andi has several dialects and no cross-dialectal standard;
 - no full-fledged dictionary (though see word lists in grammars of Gagatli and Rikvani)
 - no full-fledged grammatical parser (though see a first attempt in (Buntyakova 2022))
- Our recorded data are heterogeneous
 - different dialects
 - different conventions
 - different file formats
- Due to our limited knowledge of the Andi dialects, sometimes we do not know what the correct analysis of a given word form is.

Plans

- collect an extensive word list for Zilo;
- create a morphological parser for Zilo;

- collect an extensive word list for Zilo;
- create a morphological parser for Zilo;
- convert the material to a singular format using `phonfieldwork` (Moroz 2020);

- preprocess the annotation files, converting them to ELAN .eaf format (Wittenburg et al. 2006);
- align them with the sound;
- gloss them manually;
- publish online using the Tsakorpus platform (Arkhangelskiy 2019);
- repeat all previous steps, adding more audio files.

- preprocess the annotation files, converting them to ELAN .eaf format (Wittenburg et al. 2006);
- align them with the sound;
- gloss them manually;
- publish online using the Tsakorpus platform (Arkhangelskiy 2019);
- repeat all previous steps, adding more audio files.
- optionally, continue developing the morphological parser
- should it be cross-dialectal?

Conclusions

Conclusions:

- For a better-described language, we were able to use a morphological parser;
- For underdescribed languages, such as Andi, we still need to find the optimal way.

Thank you for your attention!

- Aglarov, M. A. (2002). *Andijcy: Istoriko-etnografičeskoe issledovanie* [*The Andi people: a historical and ethnographic study*]. Jupiter, Makhachkala.
- Alekseev, M. E. (1999). Andijskie jazyki. In Alekseev, M. E., Starostin, S. A., Klimov, G. A., and Testeleets, J. G., editors, *Jazyki mira. Kavkazskie jazyki*, pages 220–228. Moskva.
- Arkhangelskiy, T. (2019). Corpora of social media in minority Uralic languages. In *Proceedings of the fifth Workshop on Computational Linguistics for Uralic Languages*. Association for Computational Linguistics, Tartu, Estonia.
- Buntyakova, V. A. (2022). Sozdanie morfoložičeskogo parsera dla andijskogo jazyka v sisteme lexd i twol [morphological parser of andi in lexd and twol]. Term paper.

References

- Dirr, A. (1906). Kratkij očerk andijskago jazyka [grammar sketch of andi]. In *Sbornik materialov dlja opisanija mestnostej i plemën Kavkaza*. Upravlenie Kavkazskago Učebnago Okruga, Tbilisi.
- Dobrushina, N., Staferova, D., and Belokon, A. (2017). Atlas of multilingualism in dagestan online. <https://multidagestan.com>.
- Kaye, S., Moroz, G., Rochant, N., Verhees, S., and Zakirova, A. (Forthc.). Andi (Zilo dialect). In Lander, Y., Maisak, T., and Koryakov, Y., editors, *The Caucasian Languages. An International Handbook*. De Gruyter Mouton, Berlin/New York.
- Kibrik, A. E. and Kodzasov, S. V. (1988). *Sopostavitelnoye izucheniye dagestanskikh yazykov* [Comparative study of Daghestanian languages]. Moscow State University, Moscow.
- Moroz, G. (2017). *lingtypology: easy mapping for Linguistic Typology*.

References

- Moroz, G. (2020). *Phonetic fieldwork and experiments with phonfieldwork package*.
- Salimov, H. S. (2010). *Gagatlinskij govor andijskogo jazyka* [*The Gagatli dialect of Andi*]. IJaLI, Makhachkala.
- Sulejmanov, J. G. (1957). Grammatičeskij očerk andijskogo jazyka. na materiale govora s. Rikvani [A sketch of the grammar of Andi. Based on material from the dialect of the village Rikvani].
- Tsertsvadze, I. I. (1965). *Andiuri Ena*. Tbilisi: Metsniereba.
- Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). Elan: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559.