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

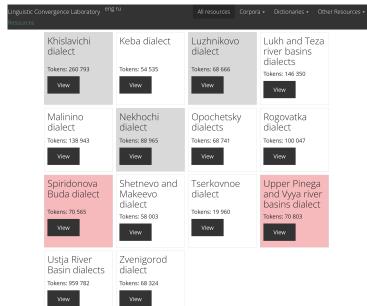
- Corpora of Linguistic Convergence Laboratory
- An example of Meadow Mari corpus
- Andi language and its data
- Our path towards the corpus

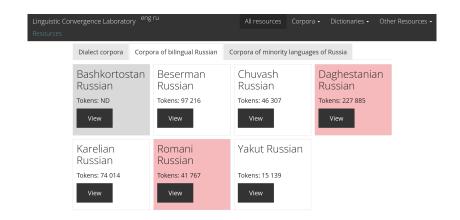
Corpora of Linguistic Convergence

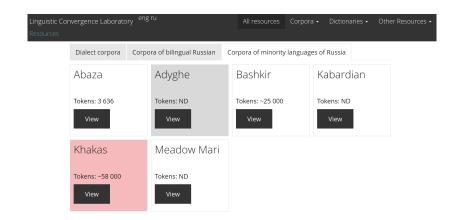
Laboratory

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



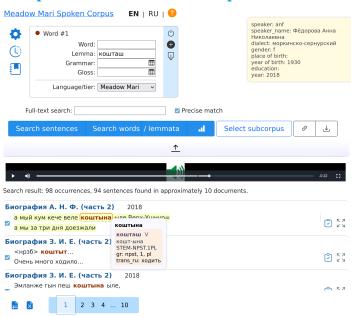








Example of Meadow Mari corpus

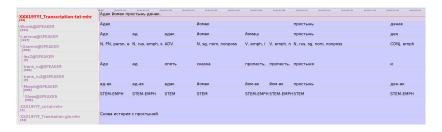


Example of Meadow Mari corpus

- translation
- glosses
- audio/video
- export to .xlsx
- sociolinguistic information



Example of required ELAN .eaf file



- hierarchy of tiers
- time alignment
- stem forms

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);	(Sulejmano 1957)	ov(Salimov 2010)	(Kaye et al. rthc)
grammar	+	+	+	+
sketch dictionary morphological	+ ± ¹	-	± -	± -
parser				

 $^{^{\}mbox{\tiny 1}}\!A$ 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 my 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 my different researchers, with different approaches to recording and processing
- We have approximately 8.27 hours of recordings
 - 7.77 hours is transcribed
 - 3 hours are aligned with sound
 - 17.35 minutes are fully glossed



Problems

- Andi has several dialects and no cross-dialectal standard;
 - no full-fledged dictionary
 - 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.



Solution

The material has to be converted to a singular format using phonfieldwork (Moroz 2020). For the Andi dialectal corpus the pipeline is as follows:

- we preprocess the annotation files, converting them to ELAN
 .eaf format (Wittenburg et al. 2006),
- align them with the sound
- gloss them manually or correct mistakes and ambiguities left by morphological parser
- publish online using the Tsakorpus platform (Arkhangelskiy 2019)
- repeat all previous steps



Conclusions

Conclusions:



Thank you for your attention!

References

- 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 Testelets, 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 morfologičeskogo 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.