A corpus of Andi field recordings

Aigul Zakirova[1]
George Moroz[1]
Elena Sokur[1]
Samira Verhees[2]
Neige Rochant[3]

[1]HSE University, Russia; [2]Independent researcher, the Netherlands;
[3]Sorbonne Nouvelle University / CNRS

20 September 2022
Tbilisi, VI International Conference: Language and Modern Technologies

Outline of the talk

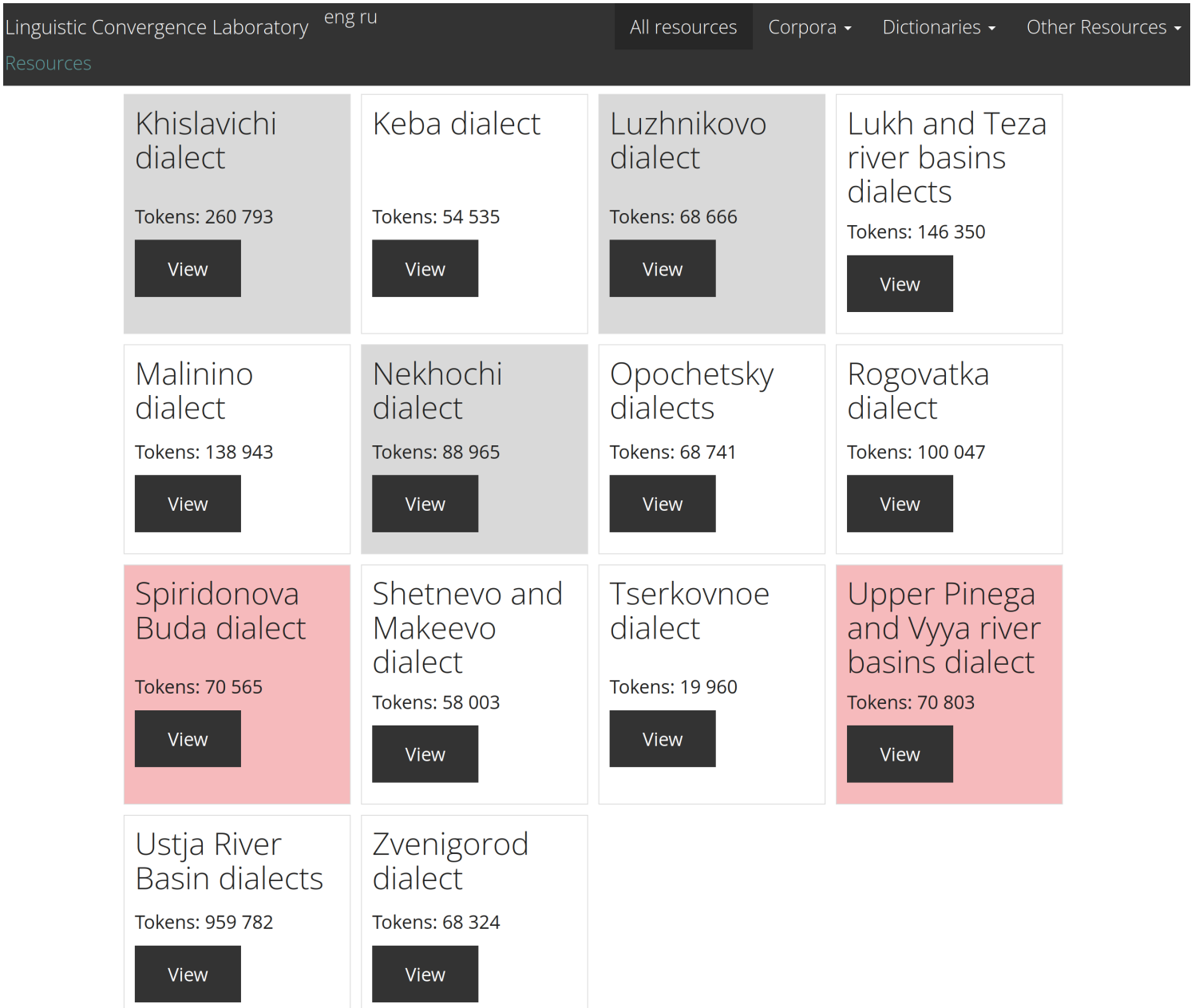
- Corpora of Linguistic Convergence Laboratory
- Corpora of minority languages of Russia: the Meadow Mari corpus
- The Andi language and the descriptive data available
- Our path towards the corpus and the challenges we have faced

Corpora of Linguistic Convergence Laboratory

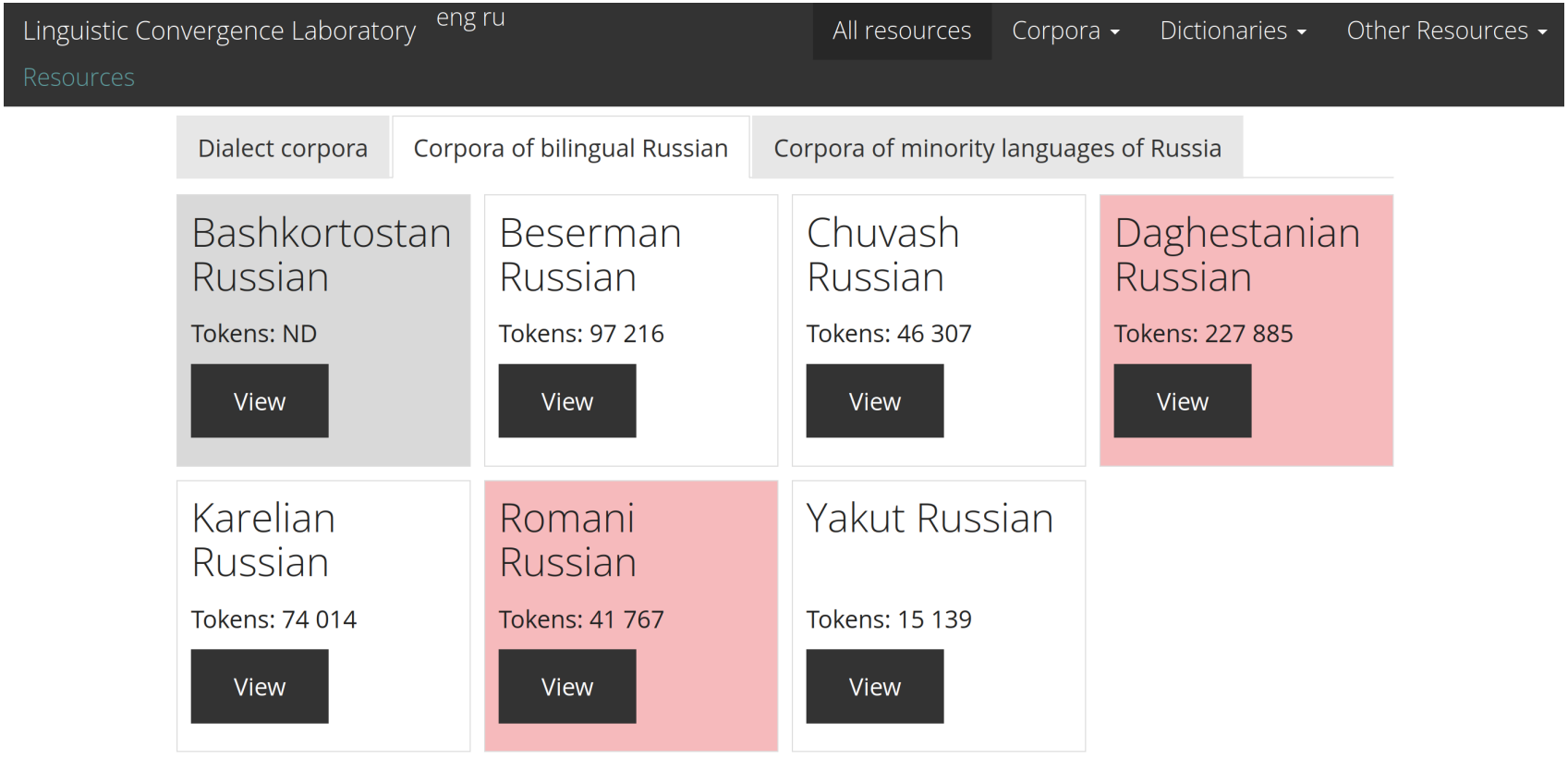
Corpora of Linguistic Convergence Laboratory: http://lingconlab.ru/

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

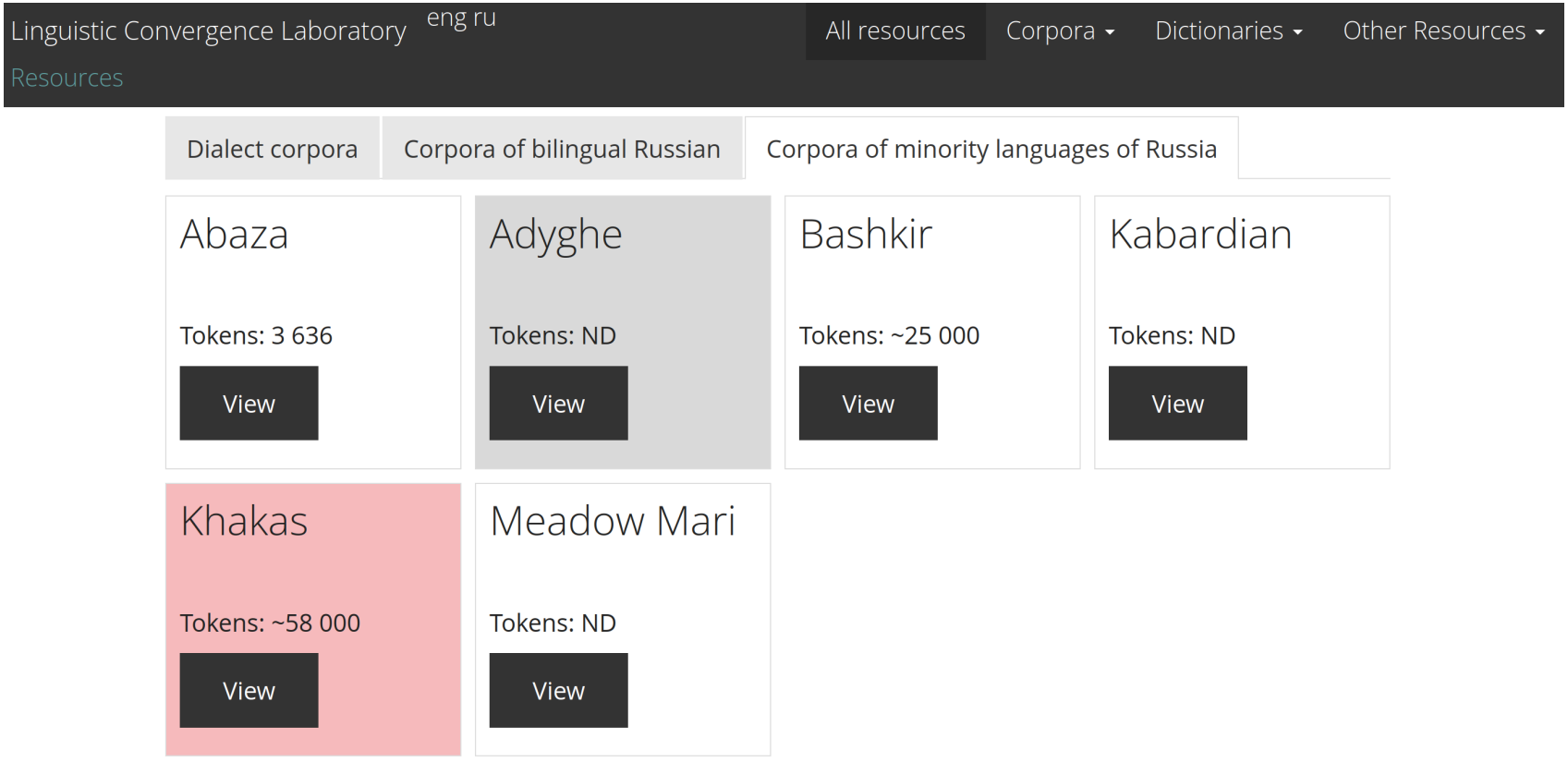
Corpora of Linguistic Convergence Laboratory: http://lingconlab.ru/



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



Corpora of Linguistic Convergence Laboratory: http://lingconlab.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 (Ethnologue)
- written and taught in school
- relatively well-described: grammars, e.g. (Pengitov et al. 1961, Alhoniemi 1993), dictionaries, e.g. http://marlamuter.com/ru/
- several written corpora, e.g. http://corp.marnii.ru/# and http://meadow-mari.web-corpora.net/index_en.html

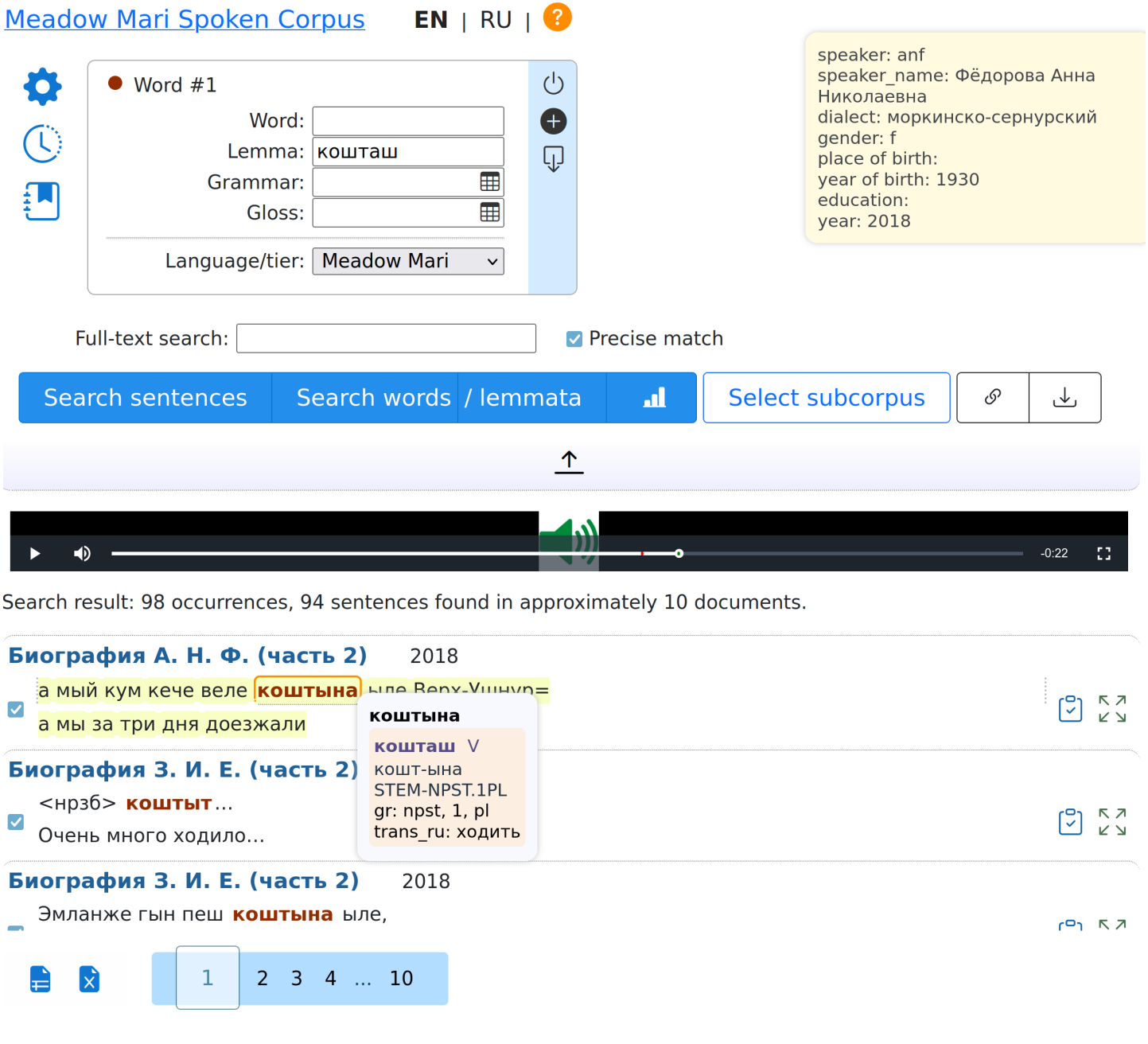
The spoken corpus of Meadow Mari

- Meadow Mari < Mari < Finno-Ugric < Uralic
- spoken in the Mari El republic and adjacent areas
- ~360.000 speakers (2010 Census)
- endangered (Ethnologue)
- written and taught in school
- relatively well-described: grammars, e.g. (Pengitov et al. 1961, Alhoniemi 1993), dictionaries, e.g. http://marlamuter.com/ru/
- several written corpora, e.g. http://corp.marnii.ru/# and http://meadow-mari.web-corpora.net/index_en.html
- texts come from the village of Staryj 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

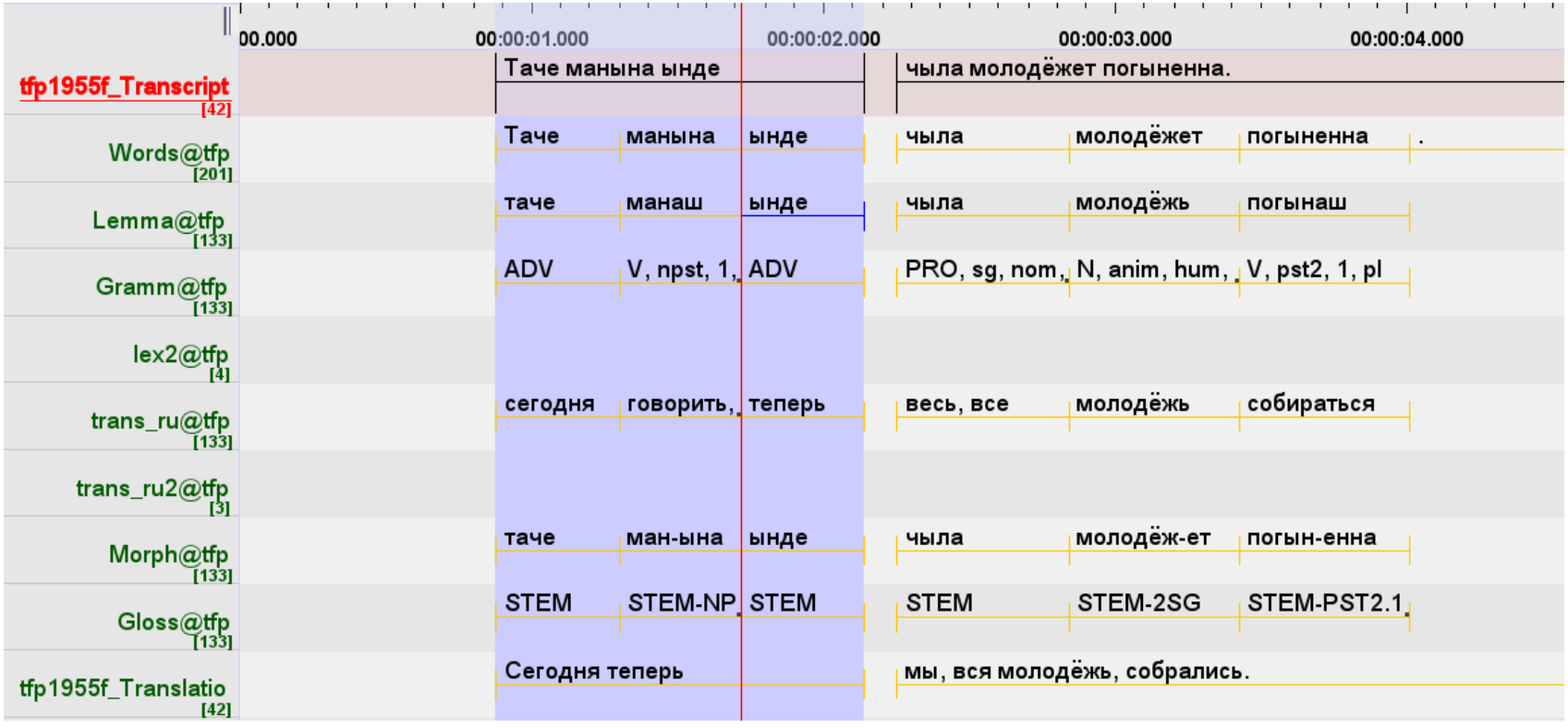
An example from the Meadow Mari corpus



An example from the Meadow Mari corpus

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

An example of a required ELAN .eaf file



- hierarchy of tiers
- time alignment
- lemmatization

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 (Upper dialects: Andi, Rikvani, Gagatli, Zilo, Chanko, Gunkha, Lower dialects: Kvankhidatli, Muni).
- more than 20,000 speakers of Andi [Aglarov (2002); All-Russian National Census 2010]
- rarely written (e.g. almost no books or newspapers, but Andi is used in messengers)

The Andi Language: a sociolinguistic background

- Andic < Avar-Andic < East Caucasian, glottocode [andi1255]
- spoken in several villages of the Botlikh district of Dagestan (Upper dialects: Andi, Rikvani, Gagatli, Zilo, Chanko, Gunkha, Lower dialects: Kvankhidatli, Muni).
- more than 20,000 speakers of Andi [Aglarov (2002); All-Russian National Census 2010]
- rarely written (e.g. almost no books or newspapers, but Andi is used in messengers)
- Andi speakers are trilingual in Andi, Avar and Russian
- Avar serves as a lingua franca in the area and is taught in school (Dobrushina et al. 2017)
- Andi speakers write using the Avar alphabeth

Andic data

	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]	-	-	-

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

Andic data

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

Andic data

- Written texts from grammar sketches, some with translation
- Field recordings from 1 trip to Rikvani, a few trips to Muni and Kvankhidatli, and several trips to Zilo
- Especially the Kvankhidatli 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

Andic data

- Written texts from grammar sketches, some with translation
- Field recordings from 1 trip to Rikvani, a few trips to Muni and Kvankhidatli, and several trips to Zilo
- Especially the Kvankhidatli 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 are 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 (but there exist word lists in Sulejmanov 1961, Salimov 2010, Kibrik, Kodzasov 1990)
  - no full-fledged morphological analyzer (but there exists 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.

Morphological analyzer

(Buntyakova 2022):

- analyzes nominal morphology;
- works for the Andi dialect (based on word lists in Kibrik, Kodzasov 1990)

Morphological analyzer

(Buntyakova 2022):

- analyzes nominal morphology;
- works for the Andi dialect (based on word lists in Kibrik, Kodzasov 1990)
- for two test corpora, the analyzer was able to analyze 7.6% ~ 13.8% tokens (due to the fact that the analyzer only processes nominal morphology + there are different spelling conventions in corpora).

Plans

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

Plans

- 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.

Plans

- 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.
- at the same time, continue developing the morphological analyzer of Upper Andi.

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 to develop the corpus, the dictionary and the parser.

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.