A corpus of Andi field recordings

Samira Verhees[1]
Aigul Zakirova[2]
George Moroz[2]
Elena Sokur[2]
[1]Independent researcher, the Netherlands; [2]HSE University, Russia

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

Outline of the talk

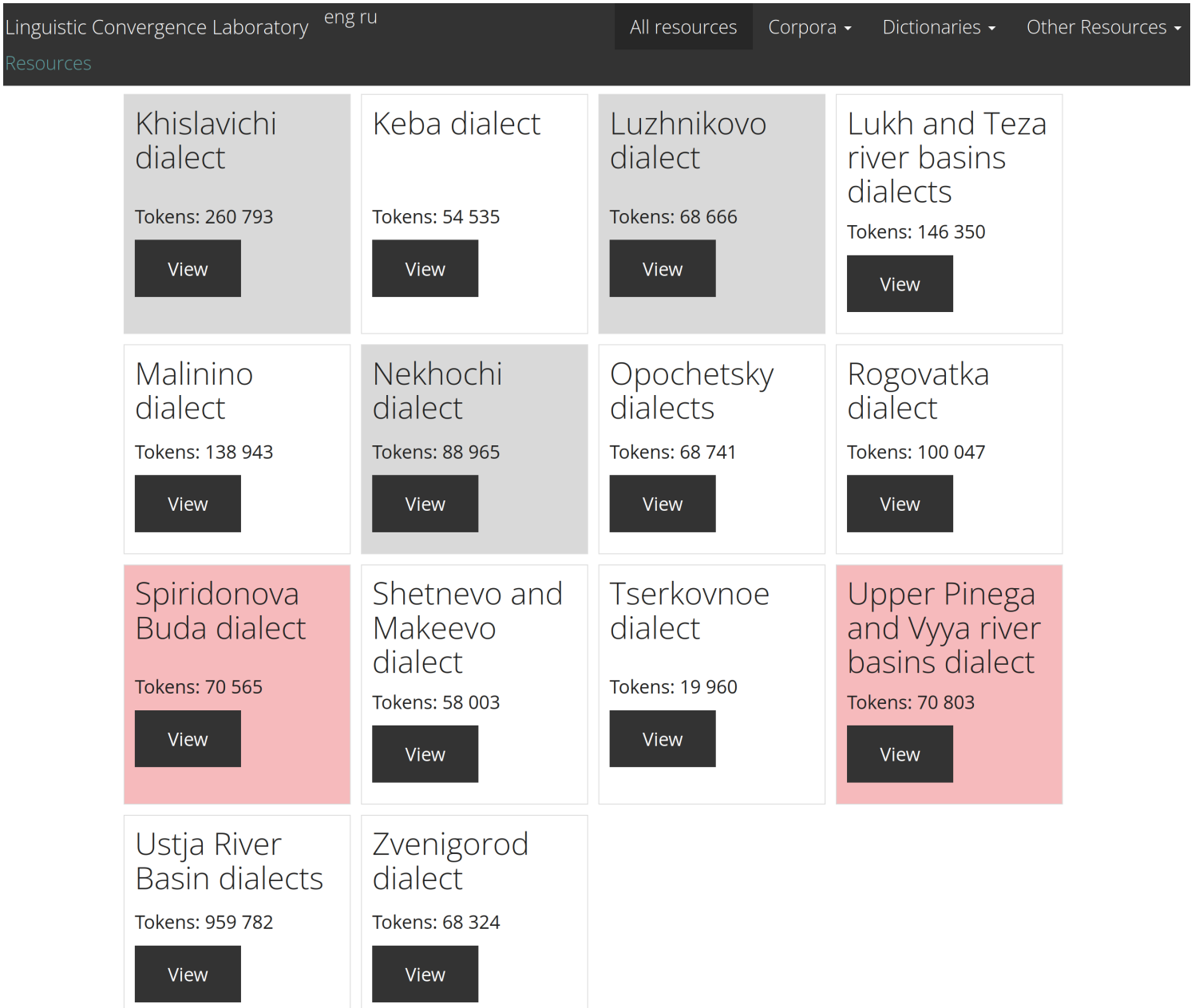
- Corpora of Linguistic Convergence Laboratory
- An example of Meadow Mari corpus
- Andi language and its data
- Our path towards the corpus

Corpora of Linguistic Convergence Laboratory

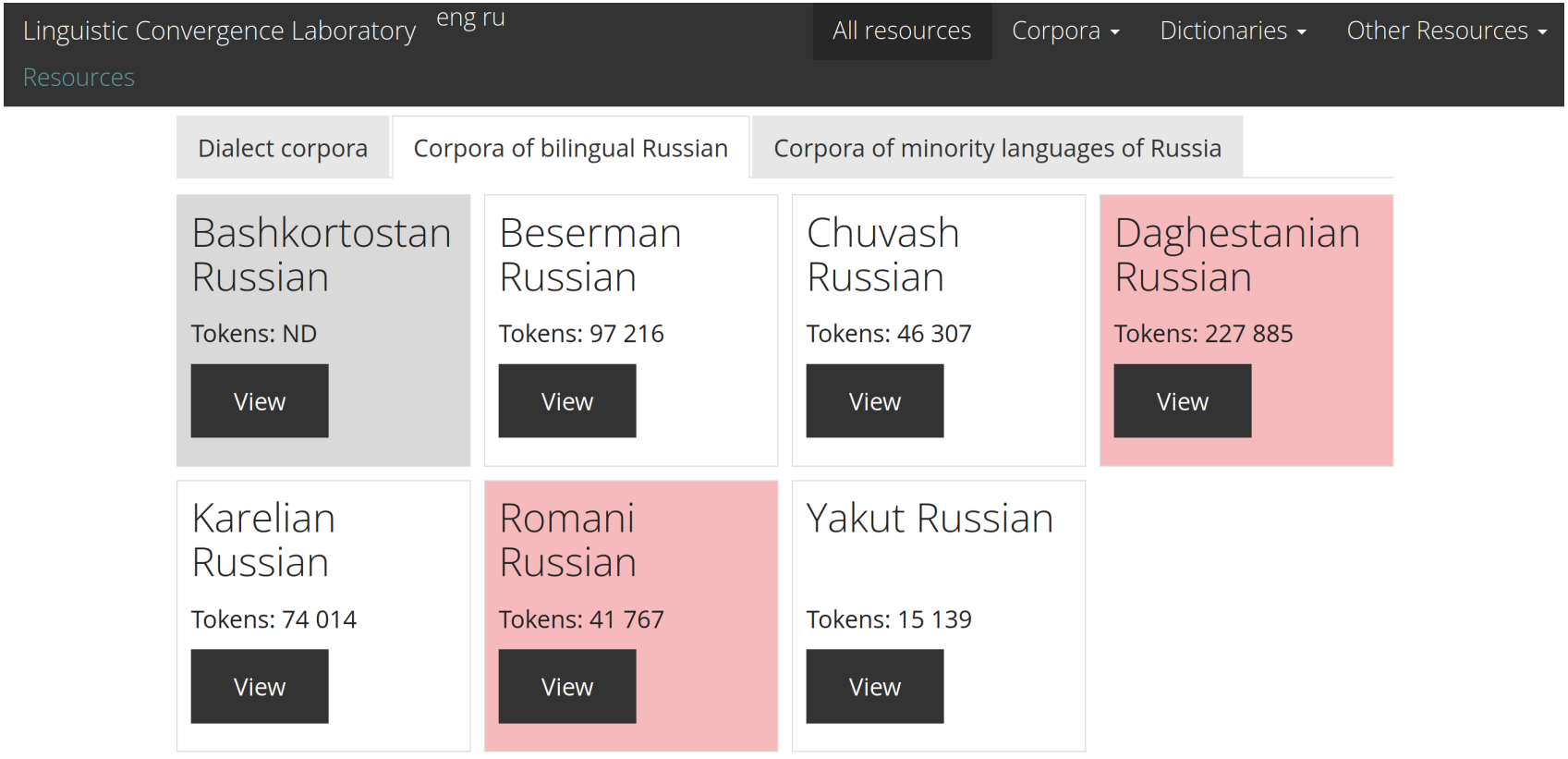
Corpora of Linguistic Convergence Laboratory: http://lingconlab.ru/

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

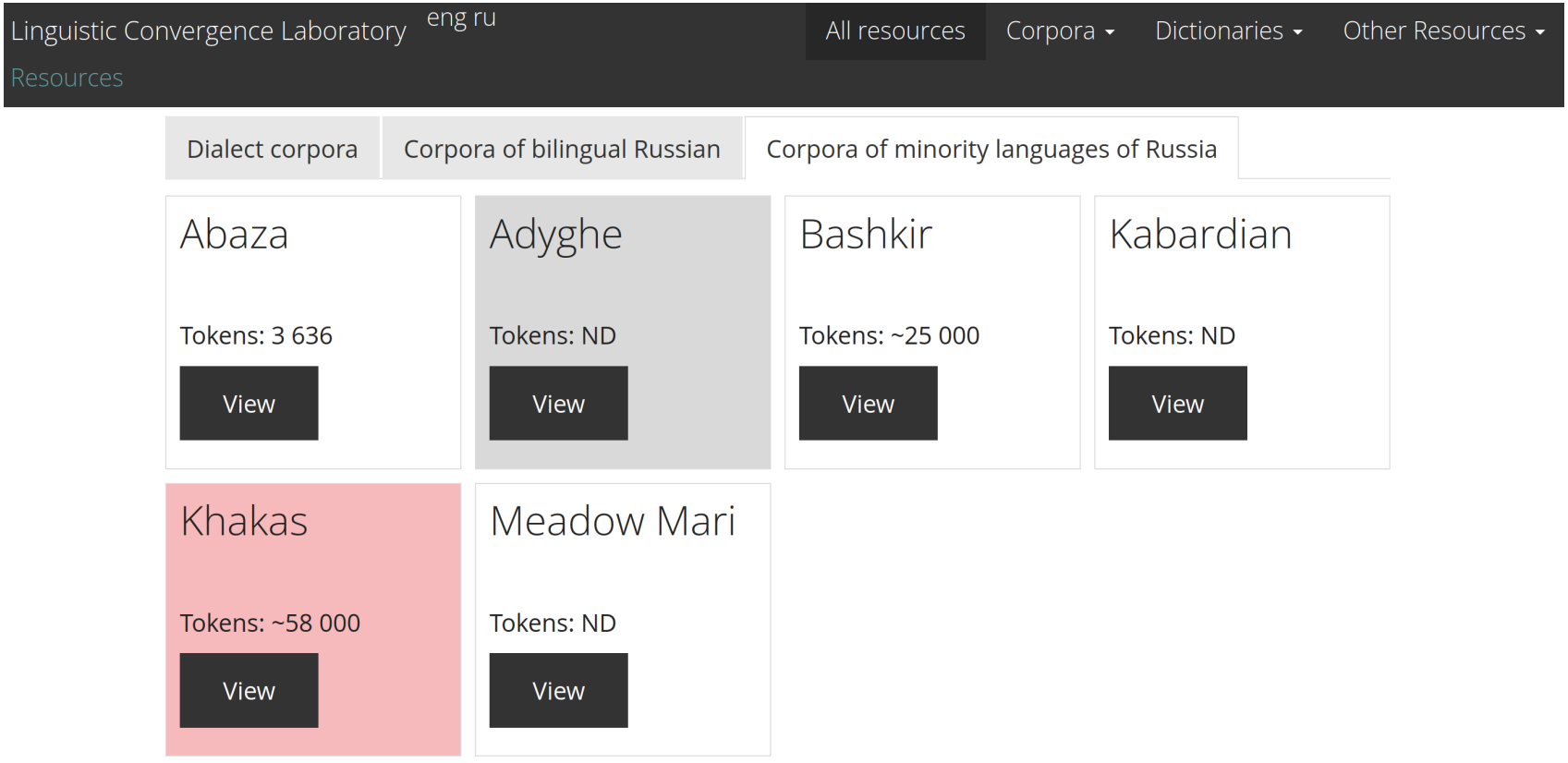
Corpora of Linguistic Convergence Laboratory: http://lingconlab.ru/



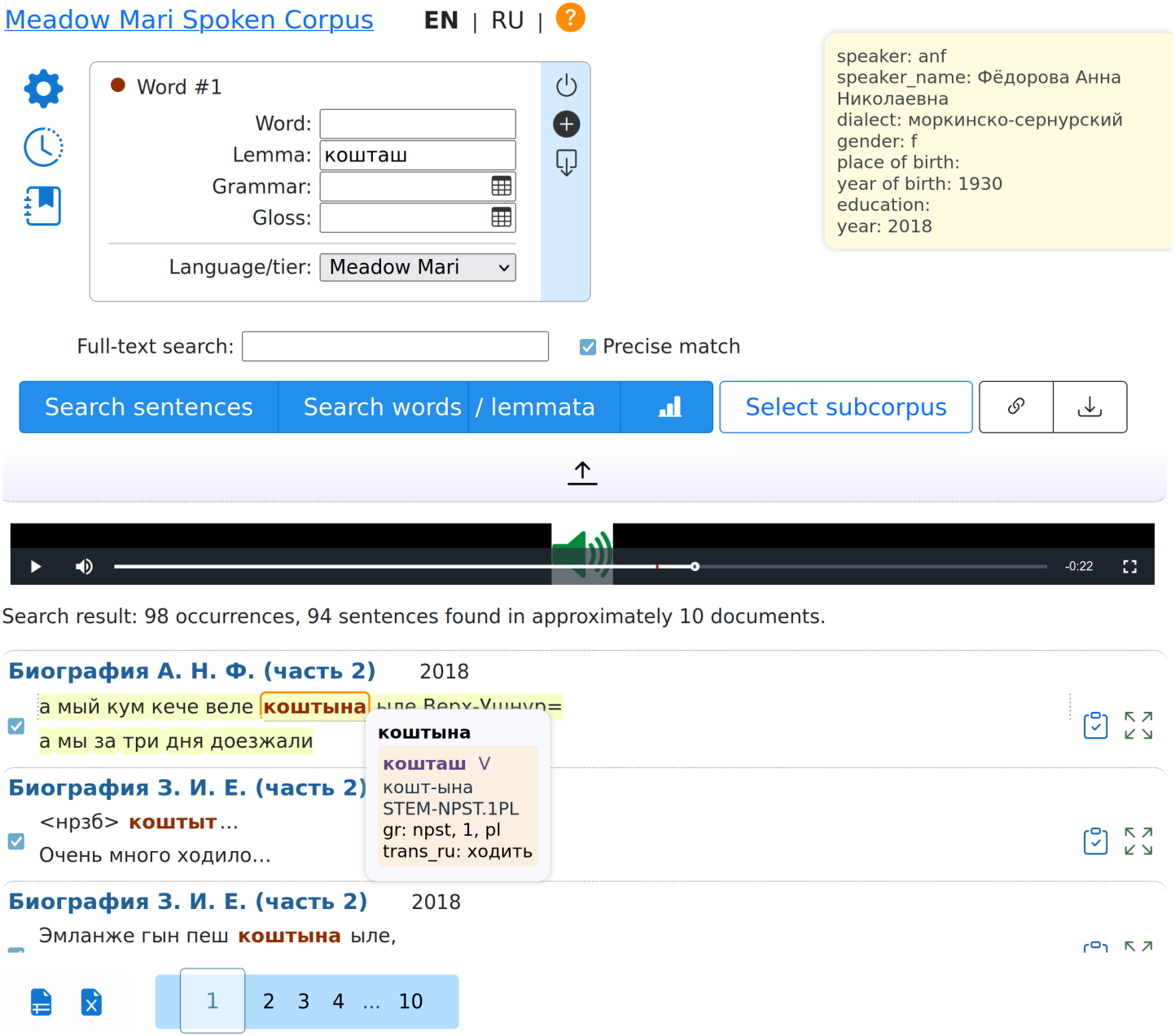
Corpora of Linguistic Convergence Laboratory: http://lingconlab.ru/



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



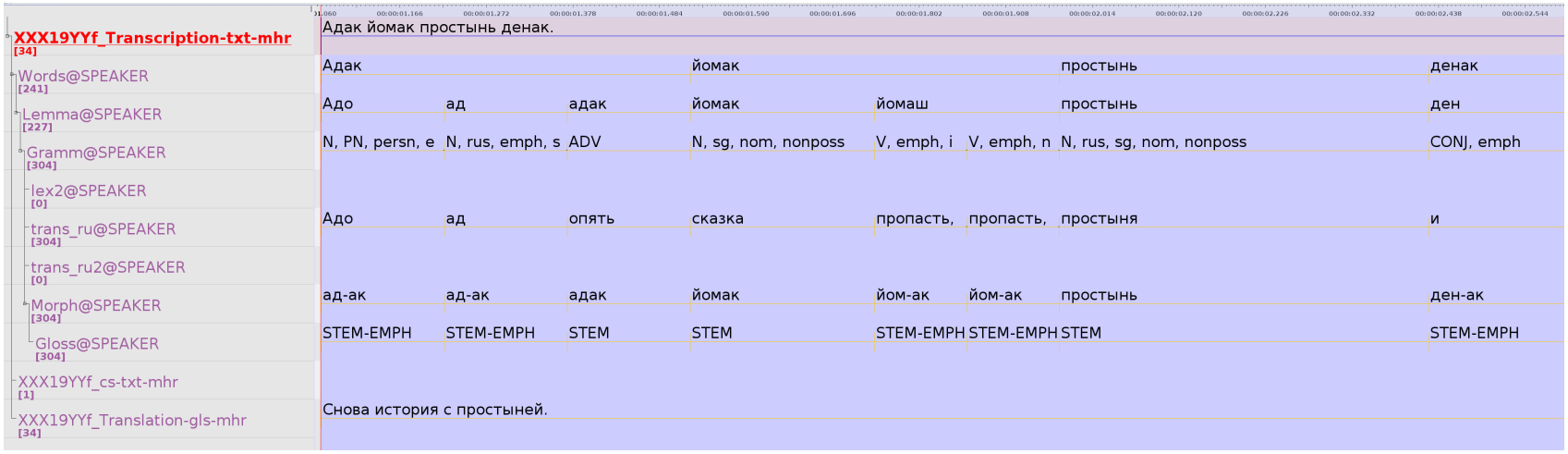
Example of Meadow Mari corpus



Example of Meadow Mari corpus

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

Example of required ELAN .eaf file



XXX19YYf_Transcription-txt-mhr [34]	Адак йомак простынь денак.						
Words@SPEAKER [241]	Адак			йомак		простынь	денак
Lemma@SPEAKER [227]	Адо	ад	адак	йомак	йомаш	простынь	ден
Gramm@SPEAKER [304]	N, PN, persn, e	N, rus, emph, s	ADV	N, sg, nom, nonposs	V, emph, i	V, emph, n	N, rus, sg, nom, nonposs	CONJ, emph
lex2@SPEAKER [0]							
trans_ru@SPEAKER [304]	Адо	ад	опять	сказка	пропасть,	пропасть,	простыня	и
trans_ru2@SPEAKER [0]							
Morph@SPEAKER [304]	ад-ак	ад-ак	адак	йомак	йом-ак	йом-ак	простынь	ден-ак
Gloss@SPEAKER [304]	STEM-EMPH	STEM-EMPH	STEM	STEM	STEM-EMPH	STEM-EMPH	STEM	STEM-EMPH
XXX19YYf_cs-txt-mhr [1]							
XXX19YYf_Translation-gls-mhr [34]	Снова история с простыней.						

- 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)

Created with `lingtypology` (Moroz 2017)

Andic data

	Andi	Rikvani	Gagatli	Zilo
materials	(Kibrik and Kodzasov 1988; Alekseev 1999)	(Sulejmanov 1957)	(Salimov 2010)	(Kaye et al. orth)
grammar sketch	+	+	+	+
dictionary	+	-	±	±
morphological parser	±[1]	-	-	-

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

Andic data

- How many hours do we have?
- How many is annotated?

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

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. (forth). 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*.

Moroz, G. (2020). *Phonetic fieldwork and experiments with phonfieldwork package*.

References

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].

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.