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

Corpora of Linguistic Convergence Laboratory

Developing the corpus of Andi

The Andi Language: a sociolinguistic background

< 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).

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.

Solutions

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

Solutions (2)

a unified and relatively flexible system of morphological annotation has been developed we have assigned the same gloss to formally different morphemes that we analyze as dialectal correspondences When we are not sure about the morphological analysis, we plan to mark it in a separate tier, so that the “dubious” status of some annotations is accessible through search.

Conclusions

Conclusions:

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.

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.

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.

References

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*.

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].

References

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.