Low-Resource ASR: Dialog Evaluation – 2021

Daniil Grebenkin, Elena Klyachko, Daria Nosenko, Oleg Serikov

Dialogue-21 18.06.2021

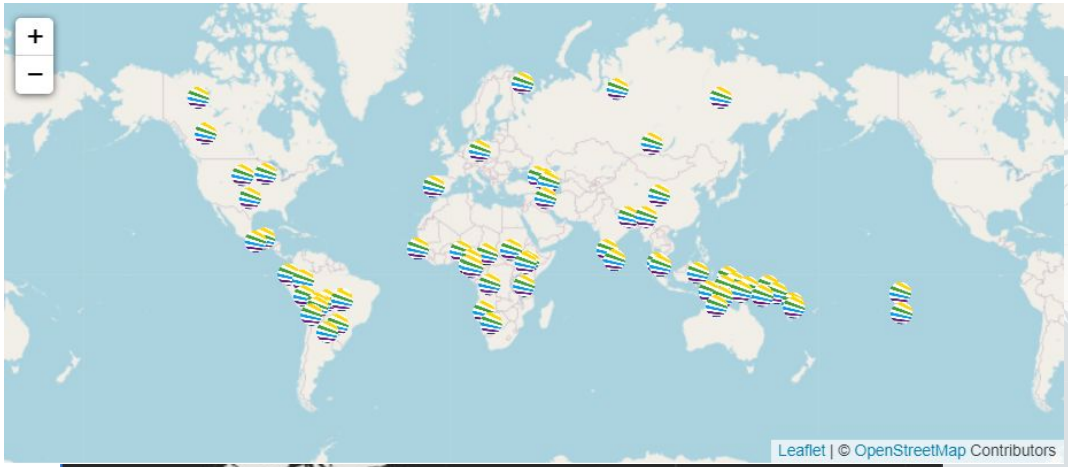
20-012-00520 A

Motivation

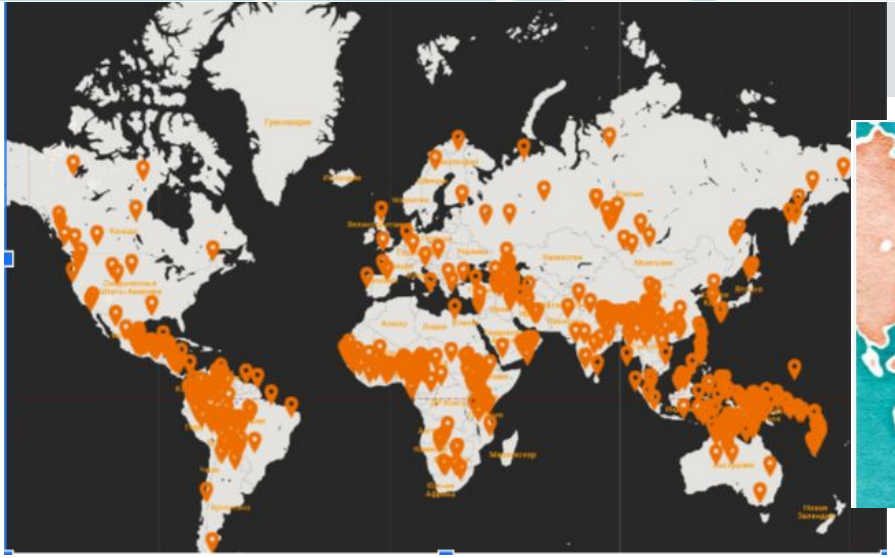
- ➔ lack of resources for most of the world's languages
- ➔ test modern approaches in low-resource settings?
- ➔ field collections used mainly for purely linguistic tasks and locked in archives
- ➔ help linguists annotate data collections

Field linguistics collections

- ELAR SOAS: https://www.elararchive.org/
- Pangloss: https://pangloss.cnrs.fr/?lang=en
- Paradisec: https://www.paradisec.org.au/
- Dobes: https://dobes.mpi.nl
- **Lingvodoc: http://lingvodoc.ispras.ru/**
- etc


+
−
Leaflet | © OpenStreetMap Contributors


Leaflet | © OpenStreetMap contributors, © CARTO



Field linguistics collections

–

- Non-uniform data labelling
- Often single annotator
- Not always available online
- Troubles with licensing

+

- Rich annotation
- Dialect, age, gender variation
- The only available source

Task	Language	Track	Training set (h)	Test set (h)	N of teams
Interspeech2018	Tamil	ASR	45	4.2	14
Interspeech2018	Telugu	ASR	45	4.2	18
Interspeech2018	Gujarati	ASR	45	5	18
GermEval2020	Swiss German	ASR	70	4	3
Interspeech2020	non-native English	ASR	≈51	≈2.5	7—9
Sigtyp2021	16 languages	LID	≈5.5 / lang.	≈0.7 / lang.	3

Tracks

- ~~Number of speakers detection~~
  - useful for linguists
- Automatic language and family detection
  - typological features?
- IPA transcription

Data

- Lingvodoc: http://lingvodoc.ispras.ru/
- Sounding vocabularies and texts
  - mostly field data
  - mostly Uralic and Altaic languages
- A virtual laboratory for historical linguistics studies

Data preparation

- Scraping
- Normalization
  - IPA: variations → `ipapy`
  - non-IPA (UPA etc) transcription
  - standard language alphabet
- Different approaches to vocabulary data
  - Stimulus in Russian + words/phrases
  - Separate words/phrases
  - Repetitions

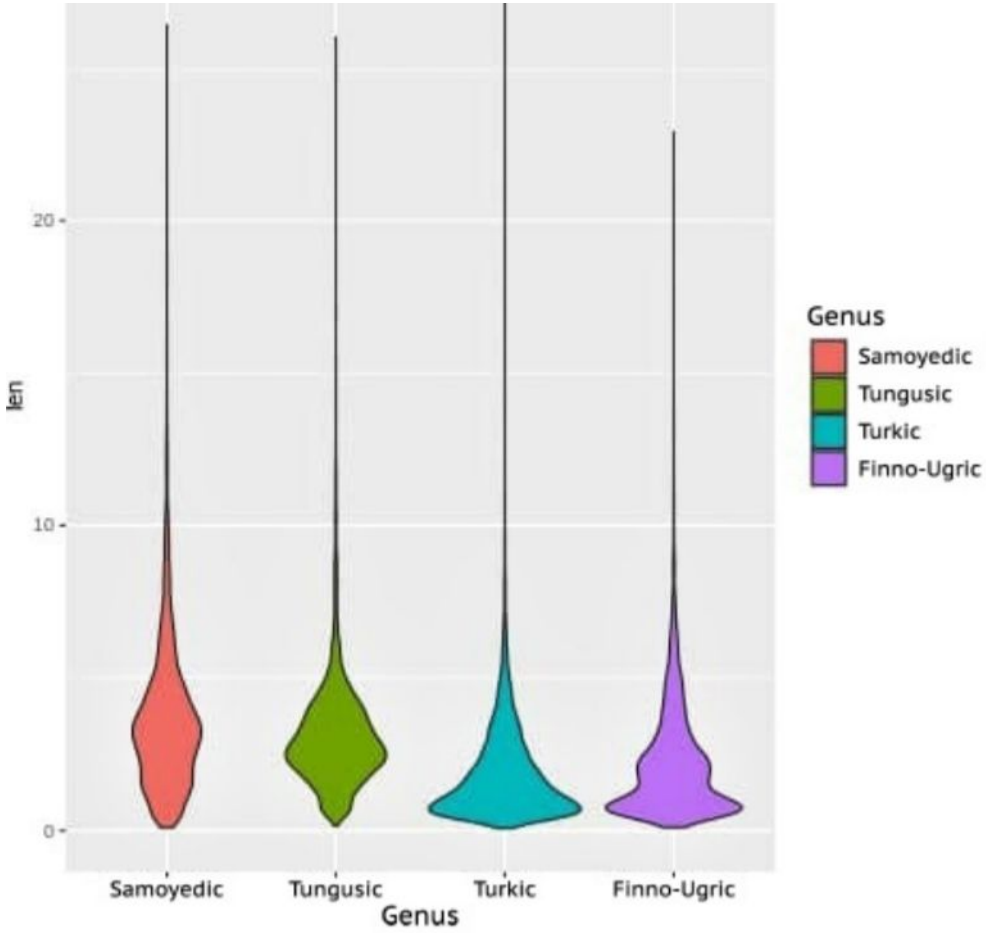


Figure 1: Recordings lenghts (in seconds) distribution across language groups

Language group	N of utterances, training set	N of utterances, test set
Tungusic	4812	2846
Turkic	10185	5799
Samoyedic	4753	628
Finno-Ugric	4281	1172

Evaluation

https://lowresource-lang-eval.github.io/content/shared_tasks/asr2021_en.html

https://competitions.codalab.org/competitions/30008

Leaderboard still active, welcome!

Automatic language and family detection

NTR/TSU:

-MFCC

-CNN with a self-attentive pooling layer for the classification task (QuartzNet ASR)

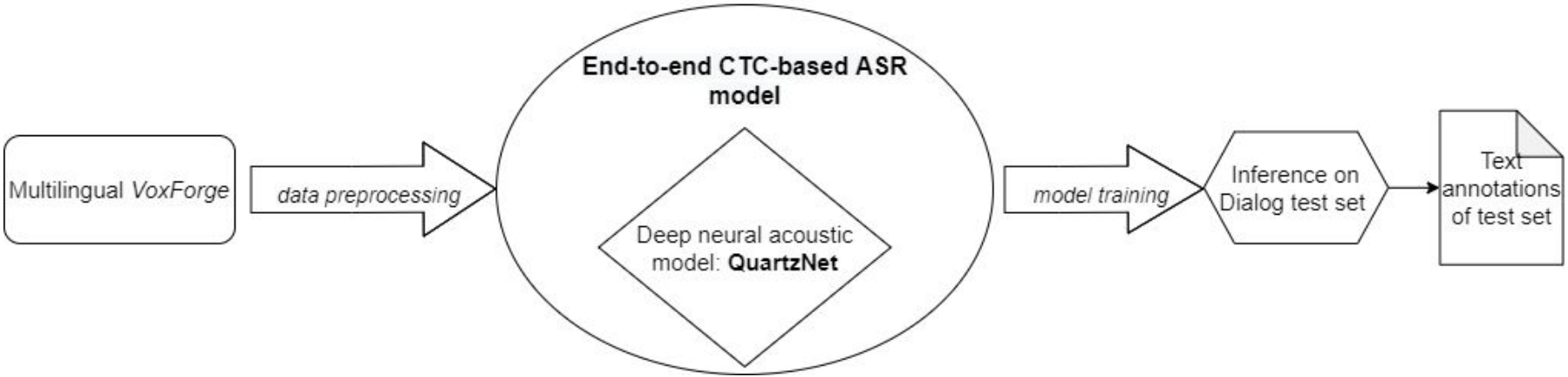
-augmentation techniques

Automatic language and family detection

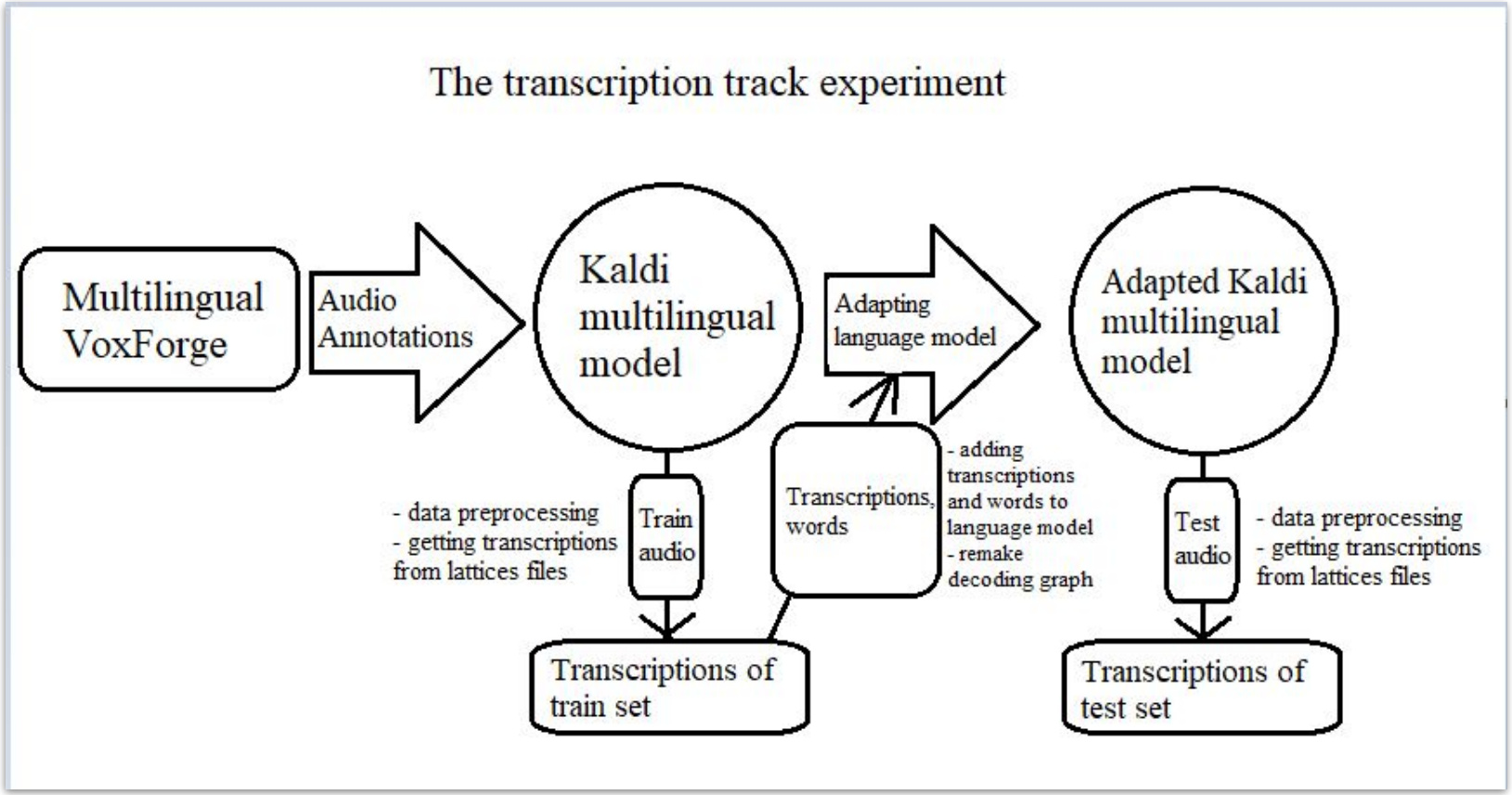
Team	LId	GId	FId
NTR	**0.06**	**0.34**	0.61
baseline	0.01	0.22	**0.82**

Team DN

The transcription task experiment: multilingual end-to-end approach



Team DG


The transcription track experiment
Multilingual VoxForge
Audio Annotations
Kaldi multilingual model
- data preprocessing
- getting transcriptions from lattices files
Train audio
Transcriptions of train set
Adapting language model
Transcriptions, words
- adding transcriptions and words to language model
- remake decoding graph
Adapted Kaldi multilingual model
Test audio
- data preprocessing
- getting transcriptions from lattices files
Transcriptions of test set

Results for the IPA transcription task

Team	Total test set (files)	Not recognized (files)	normalized CER
DG	10445	438	1.0828
DN	10445	412	1.572267

After the deadline

Team	LId	GId	FId
alumae	**0.24**	**0.63**	0.79
NTR	0.06	0.34	0.61
baseline	0.01	0.22	**0.82**

Automatic language and family detection scores

Team alumae

- architecture: ResNet-50 w. attention + dense classification layer
- pretrained (!) on **VoxLingua107** (arxiv.org/abs/2011.12998)

Acknowledgements

We would like to thank the participants

and the Lingvodoc team for sharing their data