ЖОЖдордо мамлекеттик тил маселелерин талкуулоо боюнча форум

Кыргыз тилиндеги тексттерди иштетүүнүн эсептөө ыкмалары (KyrgyzNLP): көйгөйлөрү, прогресс жана келечеги

Тимур Туратали, Антон Алексеев

due to unforeseen organizational and communication issues,
the planned presentation could not take place as scheduled

Бишкек
24-октябрь, 2024-жыл

Plan of the Short Talk

1. **Natural Language Processing**
   What, Why, and Why Now

2. **How This is Achieved Elsewhere**
   Research, Integration, Models, Data

3. **What Has Been Done**
   Academia, Activism, and Business

4. **Hopes and Aspirations**
   How the Situation Could Be Changed

Natural Language Processing: What, Why, and...

Natural language processing (NLP, NLProc) is a subfield of computer science and artificial intelligence (AI) that uses machine learning to enable computers to understand and communicate with human language

NLP = Linguistics + Computer Science + Artificial Intelligence

Why would anyone care?

KyrgyzNLP

AI for Kyrgyz speakers

Heritage/culture preservation

…and When?

Now.

Because now **we people communicate with AI through language** (prompting ChatGPT, training BERT/Llama/etc.)

AI is changing the markets, and **no one wants to be left out**

Kyrgyz Language is a Less-Resourced Language

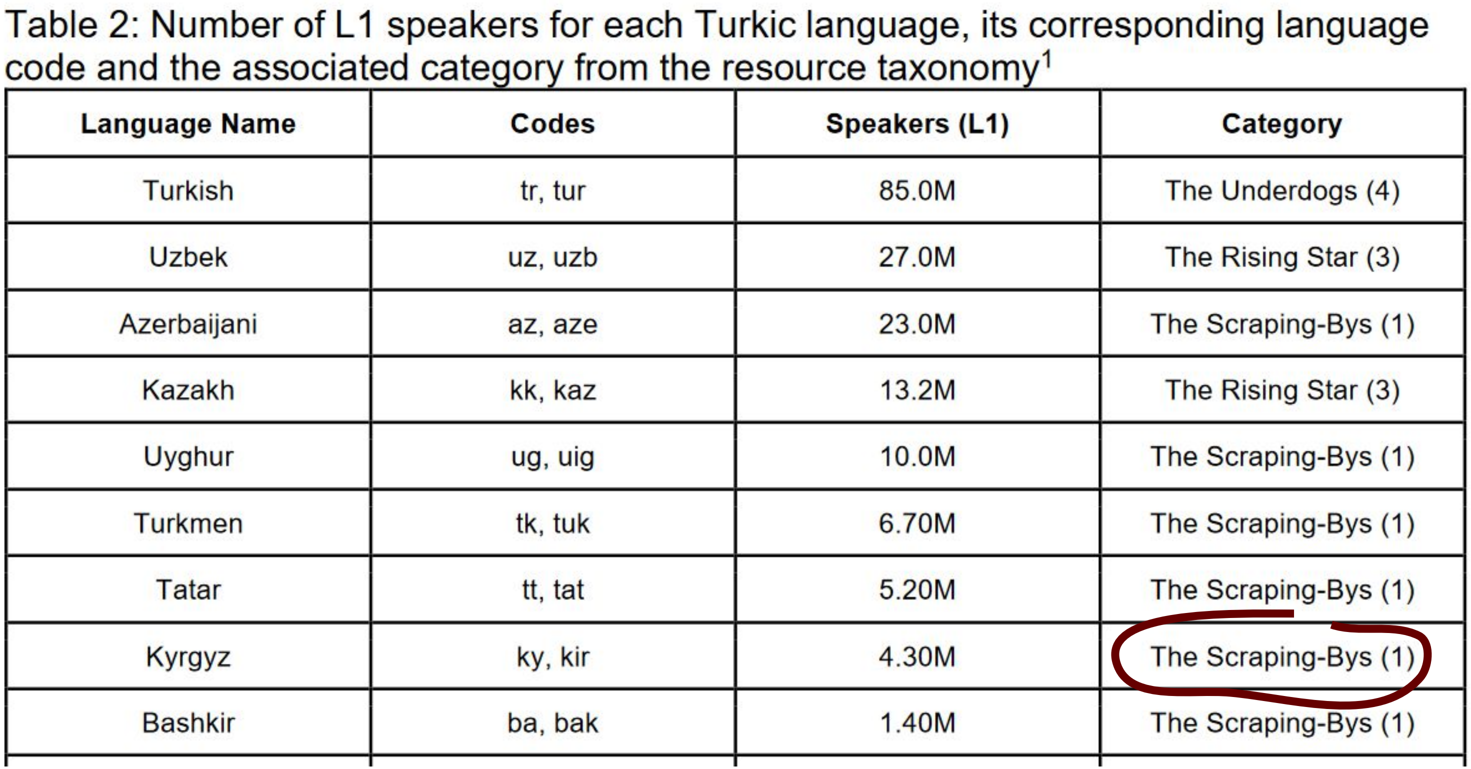


Table 2: Number of L1 speakers for each Turkic language, its corresponding language code and the associated category from the resource taxonomy[1]

Language Name	Codes	Speakers (L1)	Category
Turkish	tr, tur	85.0M	The Underdogs (4)
Uzbek	uz, uzb	27.0M	The Rising Star (3)
Azerbaijani	az, aze	23.0M	The Scraping-Bys (1)
Kazakh	kk, kaz	13.2M	The Rising Star (3)
Uyghur	ug, uig	10.0M	The Scraping-Bys (1)
Turkmen	tk, tuk	6.70M	The Scraping-Bys (1)
Tatar	tt, tat	5.20M	The Scraping-Bys (1)
Kyrgyz	ky, kir	4.30M	The Scraping-Bys (1)
Bashkir	ba, bak	1.40M	The Scraping-Bys (1)

'Scraping-Bys'

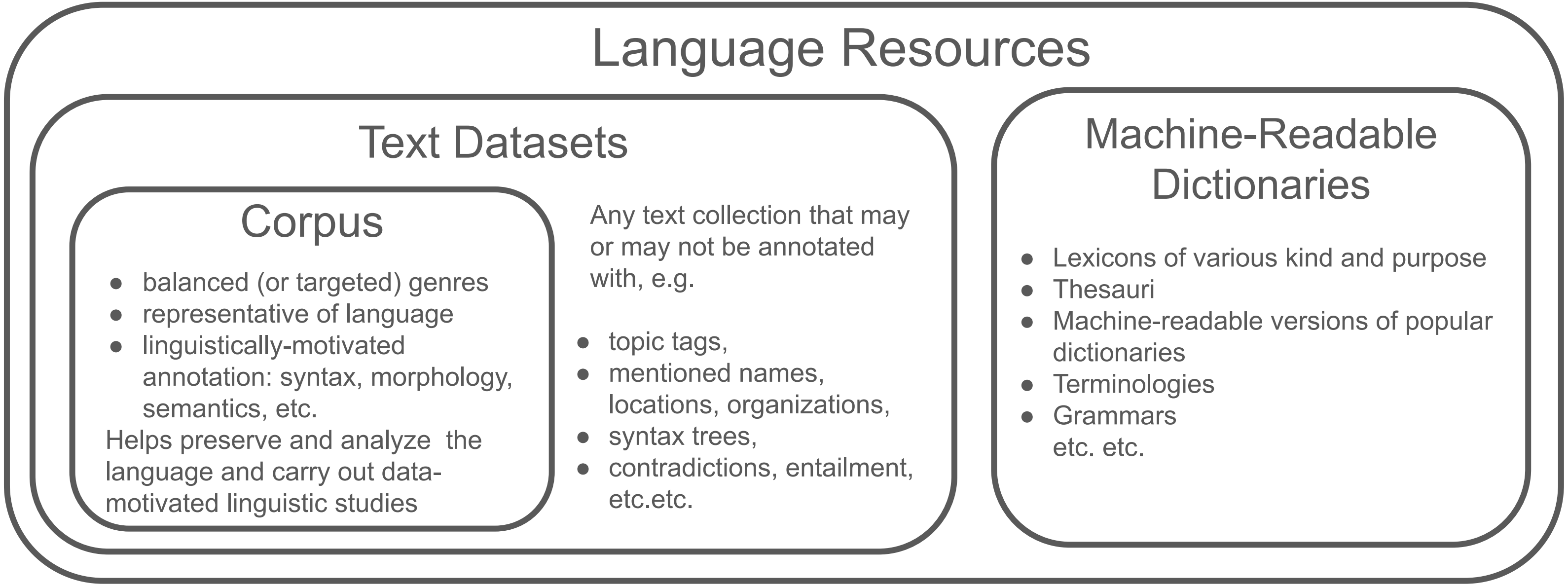
With some amount of unlabeled data, there is a possibility that they could be in a better position in the 'race' in a matter of years. However, this task will take a solid, organized movement that increases awareness about these languages, and also sparks a strong effort to collect labelled datasets for them, seeing as they have almost none

Turkic Interlingua: A Case Study of Machine Translation in Low- resource Languages
Jamshidbek Mirzakhalov
University of South Florida

Language Resources

Language Resource refers to a set of speech or language data and descriptions in machine readable form, used for building, improving or evaluating natural language and speech algorithms or systems, or, as core resources for the software localisation and language services industries, for language studies, electronic publishing, international transactions, subject-area specialists and end users

ELRA Language Resources Association website



Language Resources

Text Datasets

Corpus

- balanced (or targeted) genres
- representative of language
- linguistically-motivated annotation: syntax, morphology, semantics, etc.

Helps preserve and analyze the language and carry out data-motivated linguistic studies

Any text collection that may or may not be annotated with, e.g.

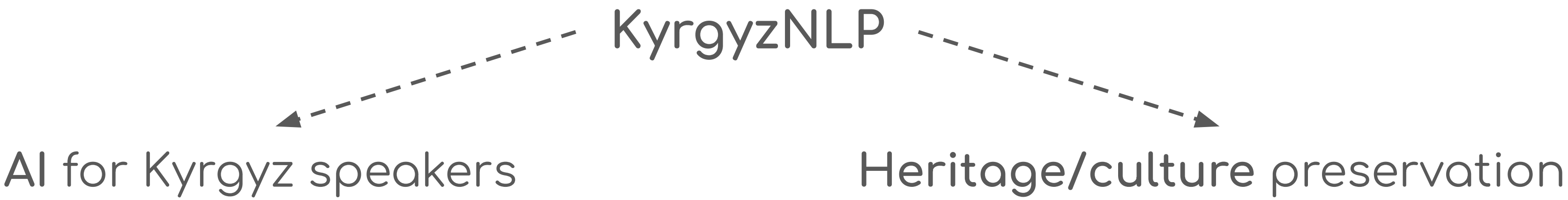
- topic tags,
- mentioned names, locations, organizations,
- syntax trees,
- contradictions, entailment, etc.etc.

Machine-Readable Dictionaries

- Lexicons of various kind and purpose
- Thesauri
- Machine-readable versions of popular dictionaries
- Terminologies
- Grammars

etc. etc.

Wait, Why Language Resources? Examples



KyrgyzNLP

AI for Kyrgyz speakers

Heritage/culture preservation

Evaluation of AI models
Need to compare the outputs of the models to gold standards = humans' responses

No other reliable way!

Data-motivated linguistic studies
Check old hypotheses, find out new regularities, update the dictionaries

Try out a new look on the tales and epics together with manasologists!

Requires **expertise**, **time** and **money**

How This is Achieved Elsewhere

Progress in NLP is usually driven

- by academic research
  (grant programmes, stimulation of top-tier publications at universities)

- by applied corporate research
  (commonly in interaction with universities and research institutions)

- by community of activists
  (but **never** limited to the community; linguists' expertise matters)

More on that later in [How to Stimulate Progress]

What Has Been Done: Open Models and Language Resources

Academic Research	Community Efforts (!)	Businesses' Contribution
Research in BSU, KTU Manas, KSTU, OshSU, Ala-too Un. and elsewhere apertium-kir by Washington et al. 2012 + more Open **syntax/morph. corpora** by Manas University (2021 – now; Benli, Kasieva, Dzhumalieva, etc.) mGPT Kyrgyz **model** by Fenogenova et al. 2024 **Datasets** for topic classification and similarity evaluation by Alekseev et al. 2023 Multiway Turkic machine translation by Turkic Interlingua **and** SIGTURK researchers (2021) and **more**!	**The Cramer Project** • KyrgyzNER model (by volunteers and KSTU students, **thanks to G. Kabaeva and G. Jumalieva**) • Alpaca (dataset), Kyrgyz News corpus, TTS model (text&voice) etc. • **AkylAI** assistant (LLM, Voice) Aya by **Cohere** (volunteers!) KyrgyzNER by **M. Jumashev** Kyrgyz MNIST: handwritten alphabet letters annotated by schoolchildren, dataset curated by **I. Jumaev**	**Ulutsoft** & **Til Commission** • Kyrgyz Text Completion (Mistral-7B-v0.1) • Text-to-Speech

Some more?

We're trying to keep track on GitHub: alexeyev/awesome-kyrgyz-nlp

How to Stimulate the Progress

- **The National Corpus** of Kyrgyz Language
  (takes time, requires collab. of multiple research groups, openness is crucial for research and industry)

- **Support** to the community efforts
  (e.g. AkylAI project, kyrgyz-nlp community, etc.)

- **International collaboration**
  + residences for foreign researchers/lecturers
  + internships for local university teachers

- **Encourage research in Kyrgyz NLP**
- **Encourage businesses' adoption of research**

UD Turkic Group

A research group working on harmon

5 followers

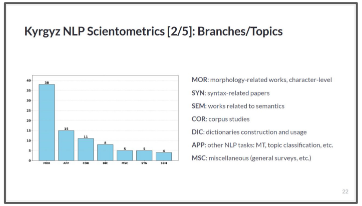
ud-turkic-group@g

Efforts are made (e.g. in KSTU)! Yet we believe the scale of support should be larger **for faster progress**

Where to Find More Information

The analysis of current state of affairs in Kyrgyz NLP

- other surveys of Kyrgyz NLP
- current state scientometrics (numbers and collaboration networks)
- conclusions of our research
- suggestions for the future: **ROADMAP** and timeline



Kyrgyz NLP Scientometrics [2/5]: Branches/Topics

MOR: morphology-related works, character-level

SYN: syntax-related papers

SEM: works related to semantics

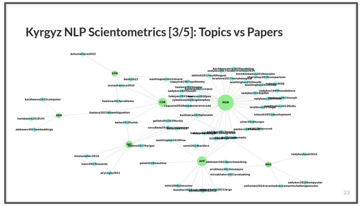
COR: corpus studies

DIC: dictionaries construction and usage

APP: other NLP tasks: MT, topic classification, etc.

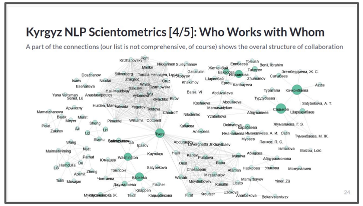
MSC: miscellaneous (general surveys, etc.)

22



Kyrgyz NLP Scientometrics [3/5]: Topics vs Papers

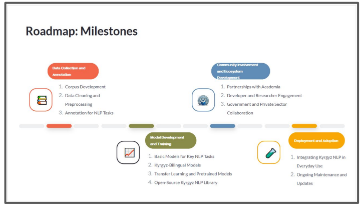
23



Kyrgyz NLP Scientometrics [4/5]: Who Works with Whom

A part of the connections (our list is not comprehensive, of course) shows the overal structure of collaboration

24



Roadmap: Milestones

Data Collection and Annotation

1. Corpus Development
2. Data Cleaning and Preprocessing
3. Annotation for NLP Tasks

Model Development and Training

1. Basic Models for Key NLP Tasks
2. Kyrgyz-Bilingual Models
3. Transfer Learning and Pretrained Models
4. Open-Source Kyrgyz NLP Library

Community Involvement and Ecosystem Development

1. Partnerships with Academia
2. Developer and Researcher Engagement
3. Government and Private Sector Collaboration

Deployment and Adoption

1. Integrating Kyrgyz NLP in Everyday Use
2. Ongoing Maintenance and Updates

More details in our talk given at the AIST-2024 conference (Oct 17–19, 2024)

kyrgyznlp.github.io

Thank you for your attention!

Көңүл бурганыңызга рахмат!

ЖОЖдордо мамлекеттик тил маселелерин талкуулоо боюнча форум

Кыргыз тилиндеги тексттерди иштететүүнүн эсептөө ыкмалары (KyrgyzNLP): көйгөйлөрү, прогресс жана келечеги

Тимур Турамали, Антон Алексеев

due to unforeseen organizational and communication issues, the planned presentation could not take place as scheduled

Бишкек
24-октябрь, 2024-жыл