The 12th International Conference on Analysis of Images, Social Networks and Texts

# Kyrgyz NLP: Challenges, Progress, and Future

**Anton Alekseev & Timur Turatali**

PDMI RAS & SPbU (St. Petersburg, Russia), KFU (Kazan, Russia)

KSTU n.a. I. Razzakov, The Cramer Project (Bishkek, Kyrgyzstan)


**Bishkek, 2024**

# Let us introduce ourselves

## Anton Alekseev

PDMI RAS, SPbU, KFU, KSTU

Research, consulting & teaching
PhD student at KSTU n . a. I. Razzakov
9y NLProc experience in industry & academia
Best NLP Paper @ AIST-2023
2 Kyrgyz datasets in 2023, more to appear

## Timur Turatali

The Cramer Project

MLE and Data Scientist with over 8y of experience in industry and startups
Founder of AI community of Kyrgyzstan
Co-founder an open-source initiative called **AkylAI** and The Cramer Project.

# What to Expect from This Talk

1. **Data annotation advocacy** — even in the age of LLMs
2. An **overview** of modern Kyrgyz language processing
   (scientometrics, recent works, our contributions, other initiatives)
3. **Current performance** of the LLMs on a Kyrgyz benchmark (work in progress)
4. Our proposal of a roadmap for Kyrgyz NLP development

## Disclaimers

- many works cite much earlier books, e.g. (Sadykov, 1987), yet we have decided to mostly focus on **works no earlier than 2011**: finding and highlighting machine-readable resources mattered the most
- we realize that there is a large body of **relevant reseach on other Turkic languages and on Turkic languages' common features**; yet it is almost impossible to cover everything in one talk

**31 Oct 1949 — 10 Nov 2023**

Last year, Kyrgyz linguistics has suffered a great loss with the passing of **Dr. Tashpolot Sadykov**, a distinguished Turkologist and computational linguist, who passed away in November 2023

Dr. Sadykov's remarkable contributions and pioneering works have left an enduring legacy. In two weeks time, he would have turned 75, and his absense is deeply felt

# Outline

1. **Introduction**: LLMs Turning All the Tables
2. **Kyrgyz Language**: What, Why, and How
3. **Kyrgyz NLP Resources**: Diving Ourselves
4. **How LLMs Perform**: Bonus Slide
5. **Community Efforts**: El Pueblo Unido
6. **Roadmap for Kyrgyz NLP**: Humble Proposal
7. **Conclusion**: Don't Say Goodbye

# Introduction

LLMs Turning ~~All~~ Some of the Tables

# Shakeup in NLProc [1/2]

Sometimes a meme is worth a thousand words
about the 'New AI Spring' and all the talk of the sort

It is obvious that Generative AI and LLMs in general are already transforming education, job market, etc.

- To some, NLP seemed *an obscure field* a few tens of years ago…
- …now the texts in natural language serve as a medium between humans and 'AI', no one wants to be left out, and everyone is pumped (for a good reason, indeed)

# Shakeup in NLProc [2/2]

**NLP has changed**: many high-resource languages enjoy a dramatic increase in quality on most common tasks thanks to the LLMs

Less-Resourced Languages (LRLs)… not so much

- insufficient **volume** of training data
- insufficient **quality** of training data
- **underdeveloped** preprocessing **instruments** and other **resources** affect the overall data quality

The demand for automatic text processing for the LRLs such as Kyrgyz (millions of speakers) only grows

ATTENTION IS ALL YOU NEED

Transformers

PRE-TRAINING OF DEEP BIDIRECTIONAL TRANSFORMERS FOR LANGUAGE UNDERSTANDING

BERT & Co

LANGUAGE MODELS ARE FEW-SHOT LEARNERS

GPT-3

INTRODUCING CHATGPT

oh no

# Less-Resourced Languages Processing Methods

1. Data Collection and Augmentation
2. Leveraging Machine Translation and Multilingual Models
3. Sampling and Transfer Learning
4. Manual and Unsupervised Tricks
5. Utilizing External Tools (Apertium and Rule-Based MT)

# Importance of Data in NLP

Carefully curated data is the **backbone of NLP,** its evolution has continuously been redefining the field, e. g.

- **Penn Treebank (1990s)**

  Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, *19*(2), 313-330.

  - Introduced syntactic and part-of-speech annotations for millions of words
  - Shifted NLP from rule-based to probabilistic, data-driven approaches
  - Enabled breakthroughs in syntactic parsing, machine translation, and statistical models

- **BioBERT (2020s)**

  Lee, J. et al. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.

  - Leveraged annotated biomedical texts to outperform general NLP models
  - Specialized data was key to success in domain-specific tasks (e.g., NER, relation extraction)
  - Triggered a trend in fine-tuning models on annotated, domain-specific datasets

# Before We Move on…

https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#ethics-safety

Gemini 1.5 Pro also shows impressive "in-context learning" skills, meaning that it can learn a new skill from information given in a long prompt, without needing additional fine-tuning. We tested this skill on the Machine Translation from One Book (MTOB) benchmark, which shows how well the model learns from information it's never seen before. When given a grammar manual for Kalamang, a language with fewer than 200 speakers worldwide, the model learns to translate English to Kalamang at a similar level to a person learning from the same content.

| Context | GPT-4 Turbo | Claude 3 Haiku | Claude 3 Sonnet | Claude 3 Opus | Gemini 1.5 Flash | Gemini 1.5 Pro | Human language le |
|---|---|---|---|---|---|---|---|
| 0-shot | 0.14<br>(30.0) | 0.24<br>(33.4) | 0.14<br>(30.0) | 0.18<br>(32.7) | 0.14<br>(31.5) | 0.18<br>(30.0) | -<br>- |
| half book | 2.04<br>(49.7) | 2.80<br>(53.5) | 3.40<br>(58.5) | 3.74<br>(58.3) | 3.00<br>(55.1) | 4.14<br>(63.9) | -<br>- |
| full book | -<br>- | -<br>- | -<br>- | -<br>- | 3.14<br>(57.4) | 4.00<br>(64.6) | 5.52<br>(70.3) |

Table 4 | Quantitative results for Kalamang→English translation on MTOB (Tanzer et al., 2023). We present human evaluation scores on a scale of 0 to 6, with 6 being an excellent translation. We include automatic metrics (BLEURT) in parentheses.

| Context | GPT-4 Turbo | Claude 3 Haiku | Claude 3 Sonnet | Claude 3 Opus | Gemini 1.5 Flash | Gemini 1.5 Pro | Human language learner |
|---|---|---|---|---|---|---|---|
| 0-shot | 0.08<br>(15.0) | 0.08<br>(15.3) | 0.08<br>(17.3) | 0.12<br>(18.7) | 0.08<br>(15.4) | 0.00<br>(12.0) | -<br>- |
| half book | 3.90<br>(45.4) | 4.46<br>(51.7) | 4.64<br>(49.2) | 5.18<br>(55.5) | 4.94<br>(54.6) | 5.38<br>(59.1) | -<br>- |
| full book | -<br>- | -<br>- | -<br>- | -<br>- | 4.66<br>(52.0) | 5.46<br>(59.0) | 5.60<br>(57.0) |

Table 5 | Quantitative results for English→Kalamang translation on MTOB (Tanzer et al., 2023). We present human evaluation scores on a scale of 0 to 6, with 6 being an excellent translation. We include automatic metrics (chrF) in parentheses.

**Kalamang language:** 250K tokens in total

- field linguistics documentation, ~500 p grammar
- ~2000-entry bilingual wordlist
- ~400 additional parallel sentences

**Gemini 1.5 be like**


https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf

# …So, Should We Even Bother?

Indeed, if a large enough LM can solve such tasks given just a textbook —
in a couple of years our LRL-related efforts will be for nothing? *

- **Language Diversity**: each language encodes unique cultural knowledge; success with one doesn't ensure success with others (though one should check!)
- **Cultural Preservation**: less-resourced languages preserve cultures and histories
- **Model Limitations**: generalization of one model (e.g., Gemini 1.5) does not guarantee performance across diverse languages (though one should check!)
- **Innovation & Research**: tackling LRLs drives novelties that benefit all NLP
- **Human Oversight**: all NLP tasks still need human review, e. g. in edge cases and cultural nuances


We should constantly evaluate whatever people call 'AI' or even good old 'intelligent systems'

Try Pitch

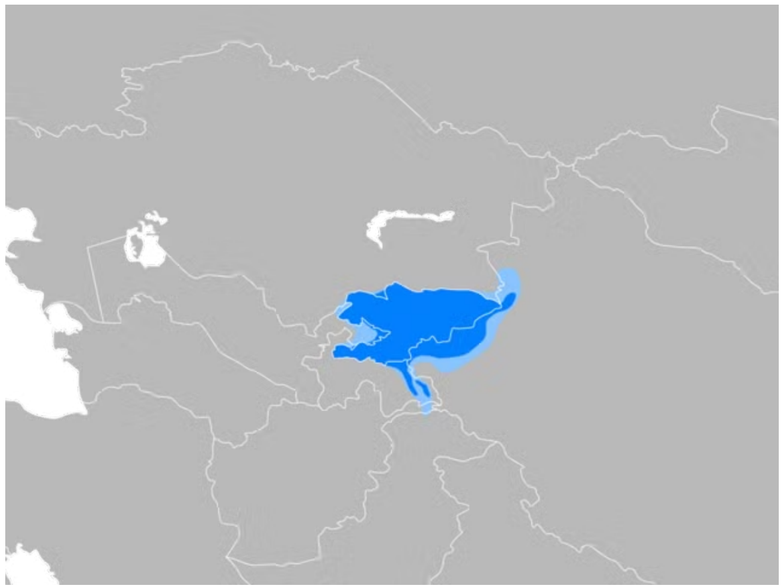
*** expert's opinion one of the speakers has actually heard recently**

# Kyrgyz Language

## What, Why, and How

# Kyrgyz Language

| | |
|---|---|
| **Native to** | Kyrgyzstan, Afghanistan, Tajikistan, Pakistan, China |
| **Region** | Central Asia |
| **Ethnicity** | Kyrgyz |
| **Native speakers** | 5.15 million* (Institutional) |
| **Language family** | Turkic<br>• Common Turkic<br>  • Kipchak<br>  • Kyrgyz–Kipchak<br>  • **Kyrgyz** |
| **Dialects** | Norhern, SoutherPamiri Kyrgyz |
| **Official language in** | Kyrgyzstan, China (Kizilsu Kyrgyz Autonomous Prefecture) |
| **Closest languages** | *Kazakh (91%), Tatar(79%), Uyghur(77%), Uzbek(76%), Altai(73%) |



*Kyrgyz at Ethnologue (26th ed., 2023)

Results of lexical-statistical analysis of the relationship of the Kyrgyz language with individual Turkic languages based on the 215-word Swadesh list

# Kyrgyz Language | Кыргыз тили | قیرغیزچا

**Kyrgyz alphabets** Cyrillic script, Perso-Arabic script (China, Afghanistan, Pakistan)

| Cyrillic alphabet (1938-present) | Old Latin alphabet (1928-1938) | Arabic alphabet (pre 1928) | Old Turkic alphabe (8th - 10th centuries) | English translation |
|---|---|---|---|---|
| Бардык адамдар өз беделинде жана укуктарында эркин жана тең укуктуу болуп жаралат. Алардын аң-сезими менен абийири бар жана бири-бирине бир туугандык мамиле кылууга тийиш. | Bardьq adamdar өz вedelinde çana uquqtarьnda erkin çana teң uquqtuu вolup çaralat. Alardьn aң-sezimi menen авijiri вar çana вiri-вirine вir tuuoŋandьq mamile qьluuoŋa tijiş. | باردىق ادامدار وز بەدەلىنده جانا ۇقۇقتارىندا مركئن جانا تەڭ ۇقۇقتۇ ۇ بولۇپ جارالات. الاردىن اڭ-سەزئمئ مەنەن ابئيئري بار جانا بئرئ-بئرئنده بئر تۇ ۇعاندىق مامئله قىلۇ ۇعا تئيئش. | ، - - ، | All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood. |

# Challenges

1. Machine-readable language resources and evaluation/training datasets are scarce
2. Most machine-readable data for training are news articles, a very *special* domain
3. Kyrgyz language exists in multiple scripts: Perso-Arabic (China) and Cyrillic (Kyrgyzstan) and dialects
4. Agglutinative languages sometimes require complex preprocessing techniques, e. g
   - a verb in Russian can have a maximum of ~150 written forms
   - in Kyrgyz, this number can achieve **thousands**

   **камсызда →камсыздандырылбагандардыкындагылардансыздарбы?**
5. Several *decentralized* corpora-building initiatives so far; a large National Corpus is yet to be built

# Earlier Surveys of Kyrgyz NLP [1/2]

## 2012: Problems and Prospects for the Development of Computational Linguistics in Kyrgyzstan by Мусаев, С. Ж., Карабаева, С. Ж., Иманалиева, А.И. in TurkLang-2013

Mentioned interest groups:

- Institute of Theoretical and Applied Mathematics of the NAS of KR
- КГУСТА (KSU of Construction, Transportation and Architecture); later merged with KSTU: computational linguistics department
- Research under Prof. P. S. Pankov
- Work under E. D. Asanov: "Tamga-KIT", "KyrSpell", etc.

The overview starts from 1990

The following linguistic problems are mentioned:
(1) dev. of a unified word formation algorithm for the Kyrgyz language,
(2) creation of comp. language models and linguistic resources at the Department of CL,
(3) the need for language-independent representations of objects in various subject areas,
(4) challenges in handling Kyrgyz spelling and orthography, including spell-checking, hyphenation rules, and managing synonyms and antonyms in the Kyrgyz language,
(5) standardization of Kyrgyz orthography,
(6) the need for the transition to a Latin alphabet and the design of a unified word formation algorithm for all Turkic languages

# Earlier Surveys of Kyrgyz NLP [2/2]

## 2024. Recent Advancements and Challenges of Turkic Central Asian Language Processing

## by Y. Veitsman (Saarland University) on arxiv

lack of labeled data for the downstream tasks. Uzbek, given its recent developments and greater variability in terms of the linguistic resources available, would be categorized as "The Hopeful," given that in the next years the efforts for collecting the datasets will not fade. Kyrgyz and Turkmen, unfortunately, not being sufficiently backed up by streamlined research efforts, would be classified as "The Scraping-Bys," with the future of their data collection processes yet unclear.

Just as in the paper (Mirzakhalov et al., 2021), the author diagnoses Kyrgyz as 'Scraping By' (Joshi et al., 2020)

The author blames

- lack of initiatives (though HTP is mentioned)
- Russian language as lingua franca in Central Asia
- limited Internet access

**Summary**: some progress in data collection and NLP, yet challenges remain (more resources and data required)

**Suggestions**: (1) collect more high-quality data,
(2) focus on **transfer learning** from Kazakh and Turkish,
(3) employ LLMs for **data augmentation**

**1 - The 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.

# Kyrgyz NLP Resources

Diving Ourselves

# Kyrgyz NLP Scientometrics [1/5]: the Collection Procedure

Set of **66** research works on Kyrgyz language processing

1. *TurkLang* papers: word 'kyrgyz' in the title or abstract
2. Relevant works cited in these papers
3. Papers from the publications lists of the corresponding *TurkLang* authors
4. Filtering out irrelevant works
   (e.g. non-NLP papers, or works with neither computational/ data-based approaches nor quantitative justifications)
5. Directly *google-scholar-ed* the rest
   (journal publications, other notable venues)

# Kyrgyz NLP Scientometrics [1/5]: the Collection Procedure

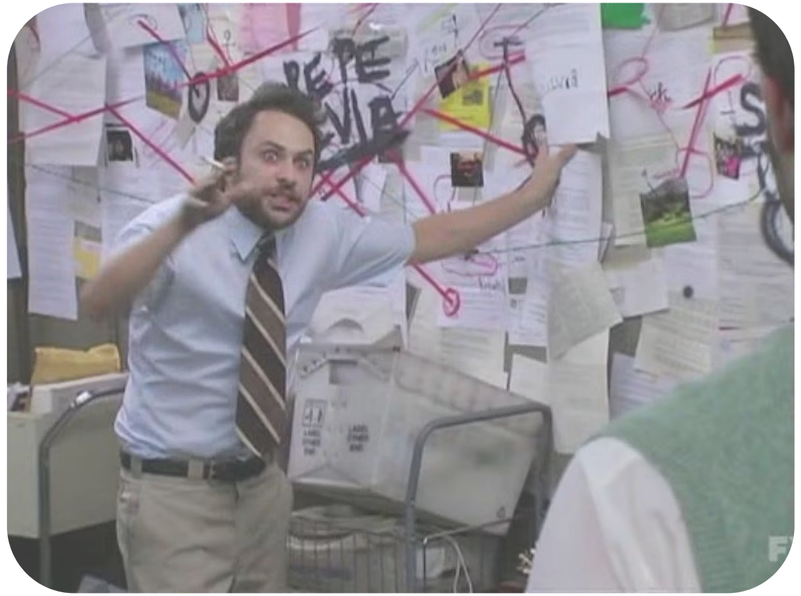
The survey we're preparing is the first review of Kyrgyz NLP of this scale

**Again**: works on tasks common *for Turkic languages in general* were excluded

**Again**: works …-2011 are poorly represented in our list, especially books

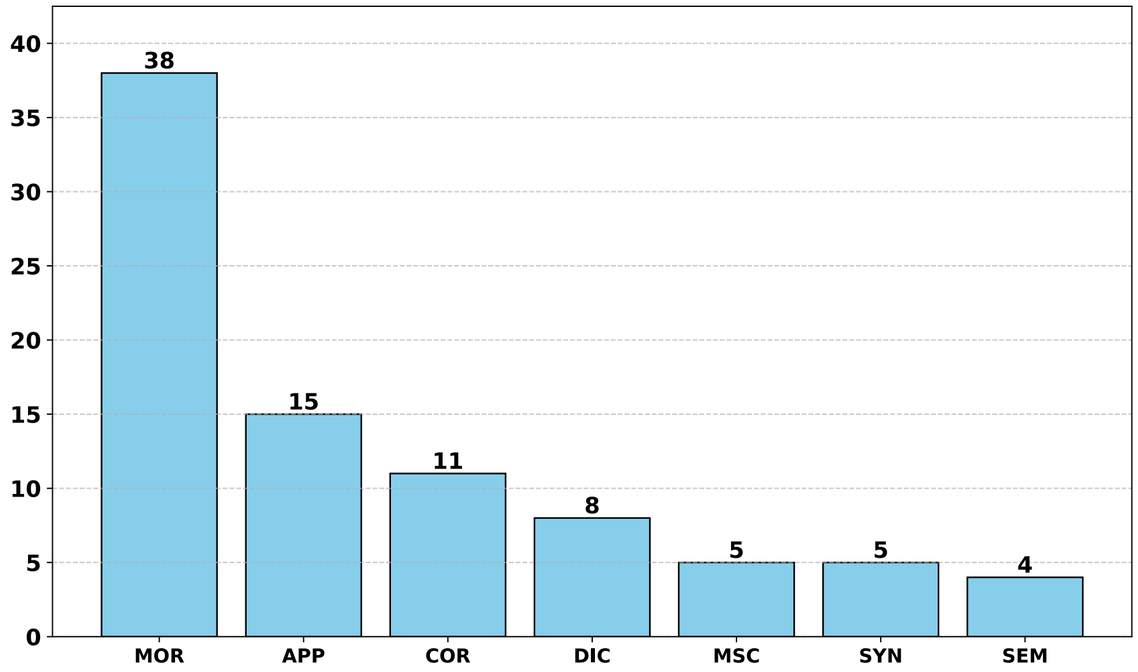
**Disclaimer**: the **list may not be comprehensive**, likely missed a few papers (sorry in advance!), yet it

- provides insights into the structure of the field,
- includes the works of those who presented some of their works on undoubtedly relevant conferences



https://en.meming.world/wiki/File:Pepe_Silvia.jpg/

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




**MOR**: morphology-related works, character-level

**SYN**: syntax-related papers

**SEM**: works related to semantics

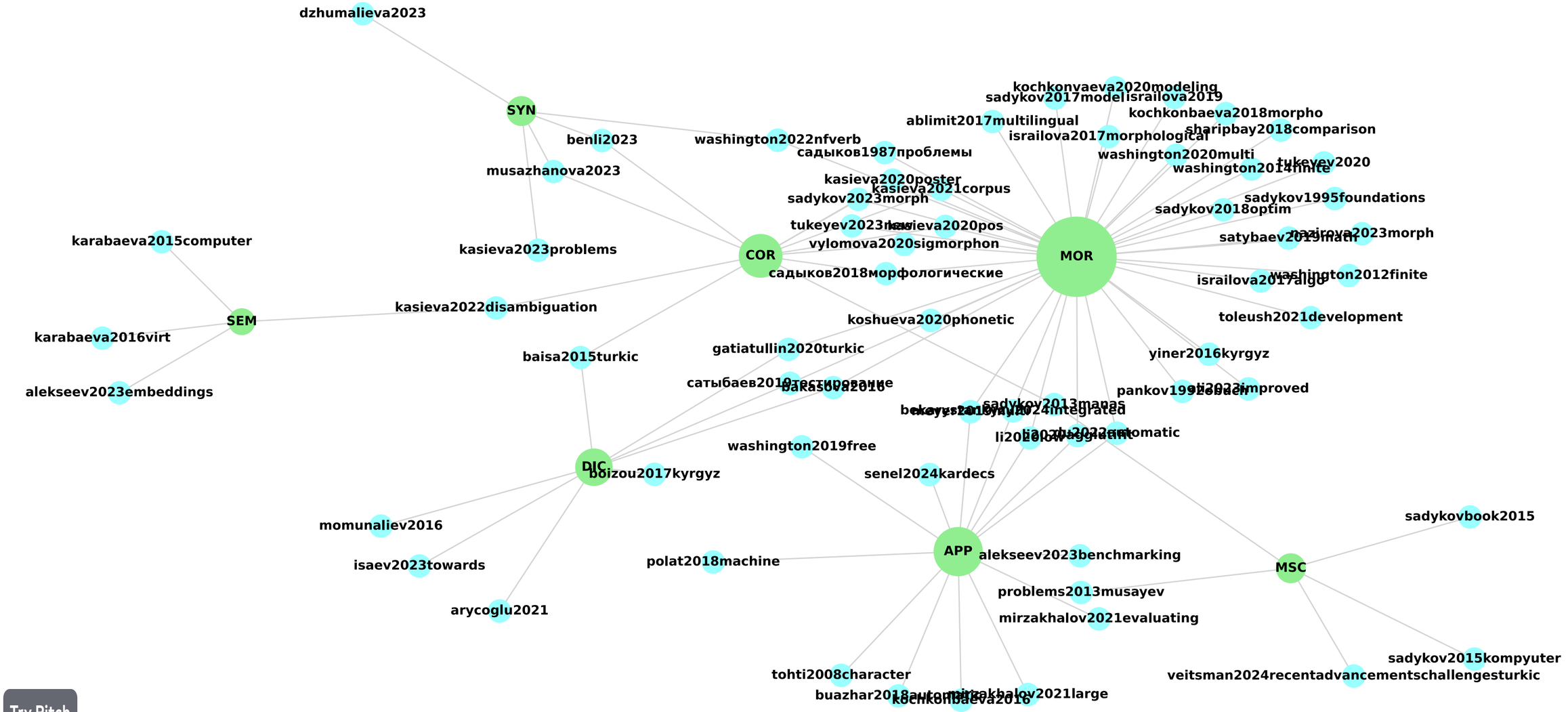
**COR**: corpus studies

**DIC**: dictionaries construction and usage

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

**MSC**: miscellaneous (general surveys, etc.)

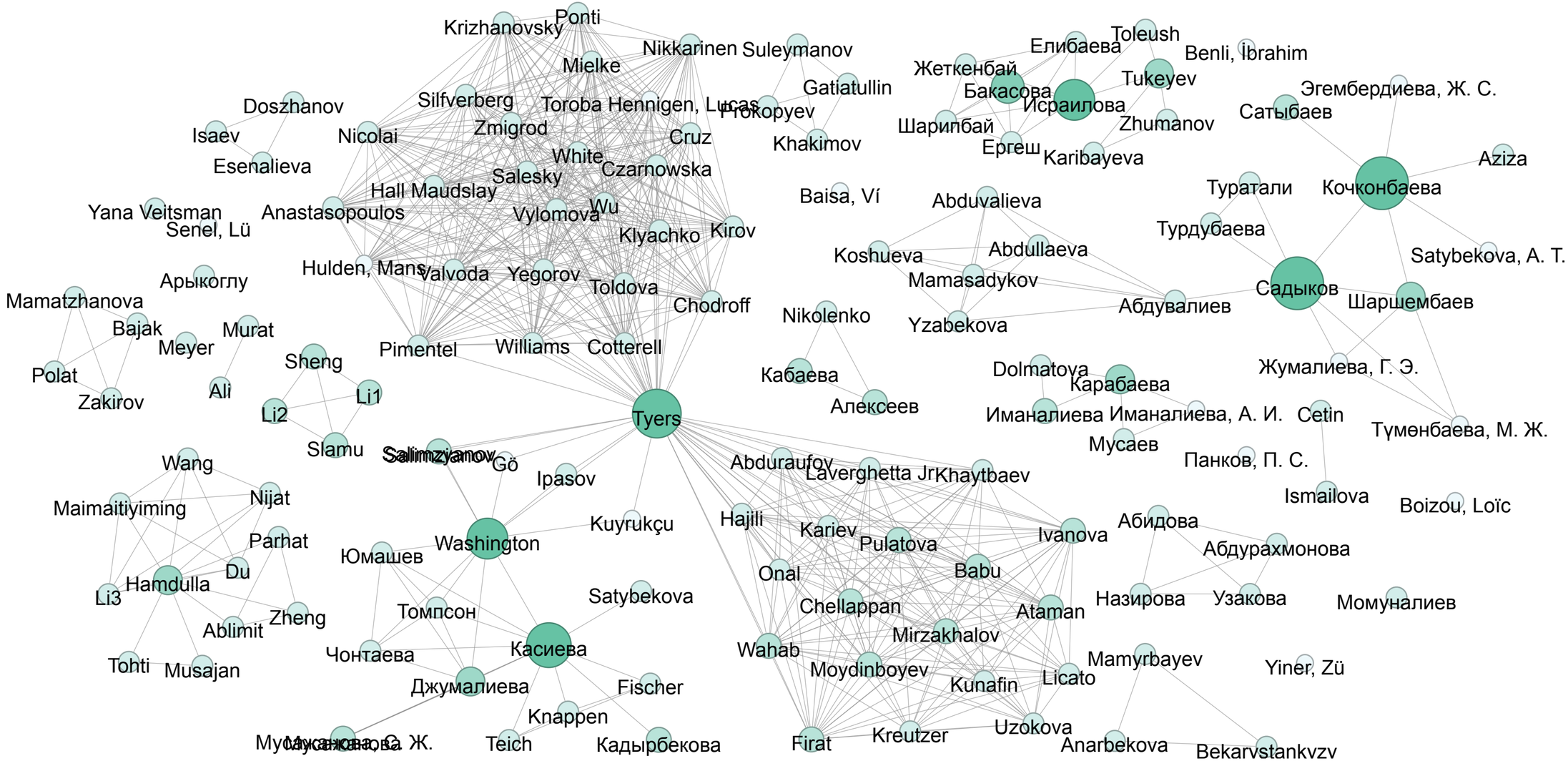
# Kyrgyz NLP Scientometrics [3/5]: Topics vs Papers




Try Pitch

# 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




Try Pitch

# Kyrgyz NLP Scientometrics [5/5]

Much more can be analyzed here!

- connecting the position (e.g. betweenness centrality) in the network to the indicators of performance
- carrying out a more comprehensive search of relevant works and possible collaborators
- build a citation graph and find the most impactful works
- compute statistics of venues where works are presented

# MOR: Finite-State Transducers for Morphology Modeling

Washington J. N., Ipasov M., Tyers F. M. A finite-state morphological transducer for Kyrgyz //LREC. – 2012. – С. 934-940.

Washington J. N., Salimzyanov I., Tyers F. M. Finite-state morphological transducers for three Kypchak languages //LREC. – 2014. – С. 3378-3385.

Washington J. N., Tyers F. M., Kuyrukçu O. Multi-script morphological transducers and transcribers for seven Turkic languages //Proceedings of the Workshop on Turkic and Languages in Contact with Turkic. – 2020. – Т. 5. – С. 173-185.

# MOR: Rule-based Modeling of Different Aspects of Morphology


Панков П. С. Обучающая и контролирующая программа по словоизменению в кыргызском языке на ПЭВМ //Бишкек: Мектеп. – 1992. – С. 20.

Yiner Z. et al. Kyrgyz orthography and morphotactics with implementation in NUVE //Proceedings of International Conference on Engineering and Natural Sciences. – 2016. – С. 1650-1658.

Sadykov, T., & Kochkonbayeva, B. (2017). Model of morphological analysis of the Kyrgyz language. In *Proceedings of the V International Conference on Computer Processing of Turkic Languages Turklang* (Vol. 2, pp. 135–154)

Кочконбаева Б. О. О морфологическом анализе в приложениях автоматической обработки текст //Бюллетень науки и практики. – 2018. – Т. 4. – №. 12. – С. 608-612.

Садыков Т., Кочконбаева Б. Об оптимизации алгоритма морфологического анализа //Шестая Международная конференция по компьютерной обработке тюркских языков «Turklang-2018».(Труды конференции)–Ташкент. – 2018

Israilova, N. A., & Bakasova, P. S. (2019). Ontological models of morphological rules of the Kyrgyz language. In Proceedings of the Seventh International Conference on Computer Processing of Turkic Languages "TurkLang 2019" (Simferopol, Crimea, Russia, October 3–5, 2019)

Сатыбаев А. Д., Кочконбаева Б. О. Тестирование программы морфологического анализатора естественного языка //Бюллетень науки и практики. – 2019. – Т. 5. – №. 3. – С. 215-219.

Кочконбаева Б. О., Эгембердиева Ж. С. Modeling of Morphological Analysis and Synthesis of Word Forms of the Natural Language // Бюллетень науки и практики. – 2020. – Т. 6. – №. 9. – С. 435-439.

Nazirova, E., Abdurakhmonova, N., Abidova, Sh., & Uzakova, M. (2023). Morphological analysis of word forms in Uzbek, Karakalpak, and Kyrgyz languages, which belong to the Turkic language family. In TurkLang (Bukhara, 2023).

# MOR: Other Algorithmic Aspects of Morphological Analysis

Бакасова П. С., Исраилова Н. А. Алгоритм образования словоформ для автоматизации процедуры пополнения базы данных словаря //Известия Кыргызского государственного технического университета им. И. Раззакова. – 2016. – №. 2. – С. 23-27.

Israilova N. A., Bakasova P. S. Morphological analyzer of the Kyrgyz language //Proceedings of the V International Conference on Computer Processing of Turkic Languages Turklang. – 2017. – Т. 2. – С. 100-116.

Исраилова Н. А. Алгоритм морфологического анализа и синтеза в трансляторе //Современные проблемы механики. – 2017. – №. 28. – С. 11-19.

Шарипбай А. А. и др. Сравнение онтологических моделей существительных казахского и кыргызского языков //Сборник содержит материалы Шестой Международной конференции по компьютерной обработке тюркских языков «TurkLang-2018»(Ташкент, Узбекистан, 18–20 октября 2018 г.). – 2018.

Tukeyev U., Karibayeva A., Zhumanov Z. Morphological segmentation method for Turkic language neural machine translation //Cogent Engineering. – 2020. – Т. 7. – №. 1. – С. 1856500.

Toleush A., Israilova N., Tukeyev U. Development of morphological segmentation for the Kyrgyz language on complete set of endings //Intelligent Information and Database Systems: 13th Asian Conference, ACIIDS 2021, Phuket, Thailand, April 7–10, 2021, Proceedings 13. – Springer International Publishing, 2021. – С. 327-339.

Tukeyev U. A NEW COMPUTATIONAL MODEL FOR TURKIC LANGUAGES MORPHOLOGY AND PROCESSING //Journal of Problems in Computer Science and Information Technologies. – 2023. – Т. 1. – №. 1.

# MOR: Working with Perso-Arabic Script

Li Z. et al. AgglutiFiT: Efficient low-resource agglutinative language model fine-tuning //IEEE Access. – 2020. – Т. 8. – С. 148489-148499.

Li X. et al. Low-resource text classification via cross-lingual language model fine-tuning //China National Conference on Chinese Computational Linguistics. – Cham : Springer International Publishing, 2020. – С. 231-246.

Ablimit M. et al. A multilingual language processing tool for Uyghur, Kazak and Kirghiz //2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). – IEEE, 2017. – С. 737-740.

# SYN: Recent Works on Computational Syntax

| | | |
|---|---|---|
| **2023** | UD_Kyrgyz-KTMU: UD for Kyrgyz by İbrahim Benli (on GitHub) | A **treebank** annotated with **dependency** relations: 781 sentences from unspecified sources (Hemingway translation to Kyrgyz, news articles, etc.); no paper yet (as of October 2024). |
| **2023** | Г.К. Джумалиева, А.А. Касиева, С.Ж. Мусажанова. АДАПТАЦИЯ ТЕРМИНОВ ВЕБ-ПРОЕКТА УНИВЕРСАЛЬНЫЕ ЗАВИСИМОСТИ НА КЫРГЫЗСКИЙ ЯЗЫК // Вестник КРСУ. 2023. Т. 23. № 6. С. 71-75. | Adapts Universal Dependencies (UD) for Kyrgyz, focusing on syntactic annotation. Provides examples of how **UD's structures apply to Kyrgyz**. Based on UD's existing platforms. |
| **2023** | Musazhanova, S., Kasieva, A., & Dzhumalieva, G. (2023). Syntactic Annotation of the Newly-Created Kyrgyz Corpus. Bulletin of the Issyk-Kul State University named after K. Tynystanov, 54(2), 139-148. IGU. | Syntactic annotation of the newly created Kyrgyz corpus, highlights issues in adapting Kyrgyz grammar to the UD framework. Provides dependency trees for parsing Kyrgyz sentences. No code provided, but the Kyrgyz corpus is manually annotated and will support further research and development. |
| **2023** | Kasieva, A., Dzhumalieva, G., Thompson, A., Jumashev, M., Chontaeva, B., & Washington, J. (2023). Issues of Kyrgyz Syntactic Annotation within the Universal Dependencies Framework. In TurkLang, Buxoro. | Examines issues in Kyrgyz syntactic annotation within UD framework: **copula tokenization, small words, null-headed clauses, and distinguishing inflection vs derivation**. In-depth analysis to improve Kyrgyz TB quality. |
| **2024** | UD_Kyrgyz-TueCL: UD for Kyrgyz by Chontaeva, Bermet; Çöltekin, Çağrı (with the support of UD Turkic Group and Kyrgyz team: J. Washington, A. Kasieva, G. Dzumalieva, A. Tursunova, M. Ryspakova, A. Kadyrbekova for their weekly informative meetings and discussions and for all the support we have received.) | 145 sentences including 20 Cairo sentences, and ~ 100 sentences suggested by UD Turkic Group. |

# SEM: Recent Works on Computational Semantics (and Around)

| | | |
|---|---|---|
| **2015** | Karabaeva S., Dolmatova P., Imanalieva A. Computer-mathematical modeling of national specificity of spatial models in Kyrgyz language // Proceedings of TurkLang-2015. – 2015. – С. 416-422 | Models **spatial semantics in Kyrgyz**, focusing on spatial terms and grammar. Models for representing these concepts, useful for NLP and educational tools for Turkic languages. No code provided, but models can be implemented. |
| **2016** | Карабаева С. Ж. Виртуальные геометрические объекты, создаваемые глаголами в кыргызском языке // В мире науки и искусства: вопросы филологии, искусствоведения и культурологии. – 2016. – №. 11 (66). – С. 74-79. | Models **virtual geometrical spaces created by verbs in Kyrgyz**, focusing on how these verbs define spatial relations and motion. Models for representing these spaces, useful for Kyrgyz LP and linguistic analysis. No code provided, models can be experimentally tested and implemented. |
| **2022** | Касиева А. А., Кадырбекова А. К. КЫРГЫЗ ТИЛИНДЕГИ ЭТИШТЕРДИН КОШ МААНИЛҮҮЛҮГҮН ЖОЮУ (жаңы түзүлгөн кыргыз тилинин корпусунун негизинде)//ИЗВЕСТИЯ ВУЗОВ КЫРГЫЗСТАНА. — 2022. — №6. — С. 341-345 | Models verb sense disambiguation (VSD) in Kyrgyz. **Presents rule-based methods for VSD automation**, useful for NLP and corpus linguistics in agglutinative languages. No code provided, but approaches are implemented using the newly created Kyrgyz corpus. |
| **2023** | А. М. Алексеев, Г. Д. Кабаева. HJ-Ky-0.1: набор данных для оценки качества векторных представлений слов кыргызского языка // Известия КГТУ им. И. Раззакова. – 2023. – № 4(68). – С. 1806-1814. | Word **embeddings evaluation dataset** construction via direct translation of RUSSE benchmark; several models are evaluated. Prepared prior to the releases of Kyrgyz-focused contextual models such as the next one. Waits to be re-annotated, available on request. |
| **2024** | Shliazhko, O., Fenogenova, A., Tikhonova, M., Kozlova, A., Mikhailov, V., & Shavrina, T. (2024). mGPT: Few-Shot Learners Go Multilingual. Transactions of the Association for Computational Linguistics, 12, 58-79. | ***LLMs go brrrr!*** GPT-3 model for Kyrgyz language, freely available. https://huggingface.co/ai-forever/mGPT-1.3B-kirgiz |

# COR: Corpora-Related Works

A few works related to the National Corpora and other resources has already been mentioned earlier

| | | |
|---|---|---|
| **2013** | Садыков Т., Шаршембаев Б. Манас” эпосунун улттук корпусун түзүү жөнүндө //Компьютерная обработка тюркских языков. Первая меж дународная конференция: Труды.–Астана: ЕНУ им. ЛН Гумилева.–148–154 б. – 2013. | Discusses the creation of a national corpus of the Kyrgyz "Manas" epic; corpus linguistics techniques to analyze and document its vocabulary and grammar. Framework for **building a searchable dictionary of the epic's texts**, useful for linguistic research and preserving Kyrgyz heritage. No code is provided, the corpus is said to be under development and supposed to serve as a key resource for future computational projects. |
| **2015** | Baisa V., Suchomel V. Turkic language support in Sketch Engine //PROCEEDINGS OF THE INTERNATIONAL CONFERENCE" TURKIC LANGUAGES PROCESSING" TurkLang-2015. – 2015. – С. 214-223. | Turkic language support in Sketch Engine, focusing on building corpora and providing tools for analysis in Kazakh, Kyrgyz, and Turkish. Presents methods for web-crawling, concordance searches, and word sketches, useful for advancing corpus linguistics and NLP for under-resourced Turkic languages. |
| **2020** | Kasieva A.A. PARTS-OF-SPEECH ANNOTATION OF THE NEWLY CREATED KYRGYZ CORPUS (Turkic Lexicon Apertium Tools) / A.A. Kasieva, A.T. Satybekova // Herald of KRSU. 2020. T. 20. No 6. S. 67-72. | Focuses on part-of-speech (POS) annotation for the newly created Kyrgyz language corpus, utilizing the Turkic Lexicon Apertium platform. Provides detailed examples of tagging and morphological analysis, useful for further development of Kyrgyz linguistic resources and corpus-based NLP applications. |

# APP: Applied NLP (Analysis and Rule-Based)

| | | |
|---|---|---|
| 2018 | Polat, Y., Zakirov, A., Bajak, S., Mamatzhanova, Z., & Bishkek, K. (2018). Machine Translation for Kyrgyz Proverbs—Google Translate Vs. Yandex Translate-From Kyrgyz into English and Turkish. In *Сборник материалов Шестой Международной конференции«TurkLang-2018»(Ташкент, Узбекистан, 18–20 октября 2018 г.)* | Compares the accuracy of Google Translate and Yandex Translate for translating Kyrgyz proverbs into English and Turkish, focusing on lexical, semantic, and syntactic performance. Concludes that Google Translate performs better overall, particularly in handling Kyrgyz's agglutinative structure, though both systems struggle with proverbs. No code is provided, but error analysis reveals key challenges for improving Kyrgyz machine translation. |
| 2016 | Кочконбаева Б. О. Табигый тилдеги тексттерди орус тилинен кыргыз тилине машиналык которууда сездерду анализдве^ н алгоритмин тузуу //Известия Кыргызского государственного технического университета им. И. Раззакова. – 2016. – №. 2. – С. 52-54. | Develops an algorithm for machine translation from Russian to Kyrgyz, focusing on morphological and syntactic analysis. Provides a framework for processing affixes and word formation in Kyrgyz, useful for improving translation accuracy between these languages in automated systems. No specific code is provided, but a morphological analyzer was developed using Delphi, with a database of common Russian words. |
| 2019 | Washington, J. N., Salimzianov, I., Tyers, F. M., Gökırmak, M., Ivanova, S., & Kuyrukçu, O. (2019). Free/open-source technologies for Turkic languages developed in the Apertium project. In *Proceedings of the International Conference on Turkic Language Processing (TURKLANG 2019)* (pp. 30-71). | Describes free and open-source technologies for Turkic languages developed in the Apertium project, including morphological transducers and machine translation systems for languages like Kazakh, Kyrgyz, and Tatar. Provides tools to support language revitalization and linguistic rights, useful for developing NLP resources in under-resourced Turkic languages. All code is available, and technologies are production-ready for several languages. |
| 2018 | Kochkonbaeva, Buazhar and Aldosova Aziza. "Automatic processing of text in natural language." *Бюллетень науки и практики* 4.7 (2018): 216-221. | Focuses on automatic processing of Kyrgyz text, proposing an algorithm for wordform analysis and morphological tagging. Presents a method to identify word stems and affixes, useful for developing NLP tools like machine translation and information retrieval for the Kyrgyz language. No code is provided, but the algorithm can be applied in various text processing systems. |

# APP: Applied NLP (MT by Turkic Interlingua)

| | | |
|---|---|---|
| **2021** | Mirzakhalov, J., Babu, A., Kunafin, A., Wahab, A., Moydinboyev, B., Ivanova, S., ... & Chellappan, S. (2021, November). Evaluating Multiway Multilingual NMT in the Turkic Languages. In *Proceedings of the Sixth Conference on Machine Translation* (pp. 518-530).<br><br>Mirzakhalov, J., Babu, A., Ataman, D., Kariev, S., Tyers, F., Abduraufov, O., ... & Chellappan, S. (2021, November). A Large-Scale Study of Machine Translation in Turkic Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 5876-5890). | Evaluates **multiway** multilingual neural machine translation (MNMT) for Turkic languages: 22 languages, most are low-resource. Comparison between bilingual baselines and MNMT models, demonstrating that MNMT performs better **in out-of-domain tasks**, useful for enhancing machine translation in under-resourced Turkic languages.<br><br>**All code and models are made publicly available for further research and development**<br><br>In-domain results: bilingual baselines sometimes outperform MNMT models for Kyrgyz when dealing with domain-specific translations |

# APP: Applied NLP (Other)

| | | |
|---|---|---|
| **2008** | Tohti, T., Musajan, W., & Hamdulla, A. (2008, July). Character code conversion and misspelled word processing in Uyghur, Kazak, Kyrgyz multilingual information retrieval system. In *2008 International Conference on Advanced Language Processing and Web Information Technology* (pp. 139-144). IEEE. | Solutions for character code conversion and spelling error correction in Uyghur, Kazak, and Kyrgyz for multilingual information retrieval. Methods for converting non-Unicode characters to Unicode and root-based query expansion to handle spelling errors, improving search accuracy and recall in these languages. No specific code is provided, but algorithms for character conversion and query correction are tested within the system. |
| **2023** | Alekseev A., Nikolenko S., Kabaeva G. Benchmarking Multilabel Topic Classification in the Kyrgyz Language //International Conference on Analysis of Images, Social Networks and Texts. – Cham : Springer Nature Switzerland, 2023. – С. 21-35. | Presenting a new **manually labeled dataset of news articles from 24.kg**. Provides several baseline models, showing that multilingual neural models like XLM-RoBERTa perform best, useful for advancing Kyrgyz-language NLP tools. No specific code is provided, but the dataset and models are planned for public release. |
| **2024** | Senel, L. K., Ebing, B., Baghirova, K., Schütze, H., & Glavaš, G. (2024, March). Kardeş-NLU: Transfer to Low-Resource Languages with the Help of a High-Resource Cousin–A Benchmark and Evaluation for Turkic Languages. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1672-1688). | Introduces the Kardes-NLU benchmark, focusing on **cross-lingual transfer** from Turkish to less resourced Turkic languages like Kyrgyz. Presents strategies for improving natural language understanding tasks through intermediate training and fine-tuning with Turkish, showing significant gains in accuracy for Kyrgyz, useful for NLP advancements in low-resource languages. **All code and models are publicly available for further research and applications**. |

# Kyrgyz NLP Analysis: Lessons Learned So Far

*Correlation doesn't imply causation**, however, some ideas are clear

Join SIGTURK

https://sigturk.github.io/

1. **Collaboration** and diverse expertise within the research team **do matter**
   (MT and UD papers would not be possible without Turkic languages interests groups)


https://github.com/ud-turkic

2. We haven't analyzed the citations yet; from what we've seen in references lists:
   the textbooks and other **educational and systematizing materials** are important

3. Data and tools are useful yet scarce, a lot of **opportunities to make something new;**
   instruments and resources are in high esteem ⇒ very **welcome at top-ranked venues** (and cited)

4. **Online presence** is important for research visibility
   (some projects didn't make it into this talk because we could not find any digital traces)

Try Pitch

*Ironically, a misconception abused in the *Other NLP* (the so-called neurolinguistic programming)

# Other Datasets & Language Resources for Kyrgyz

- **Corpora:**

Manas-UdS: 1.2M words, 84 literary texts, 5 genres: novel, novelette, epic, minor epic, and fairy tale; lemmata, PoS tags, rich per-text metadata.

kkWaC: Kyrgyz corpus from the web, 19M words, Jan 2012

Kyrgyz in Leipzig Corpora Collecion: Community data / Newscrawl (1M sentences) / Wikipedia (300K sentences)

TilCorpusu by the Cramer project: Kyrgyz corpus, 300M words, news+fiction, made public in Dec 2023

Kyrgyz language hand-written letters (Kyrgyz MNIST): hand-written Kyrgyz alphabet letters collection for machine learning applications; original images (a total of 80213) have been transformed to 50x50 images, then to CSV format

- **Morphology & Syntax:**

UD project comments on difficulties in Turkish language processing, might bring light to the question why parsing Kyrgyz is hard as well

KTMU's UD Treebank, 781 sentences

- **Named Entity Recognition:**

**WikiANN** has a Kyrgyz language part

**KyrgyzNER**: [not published yet]

- **Word Similarity Data**

Kyrgyz Word Embedding Evaluation: [not published yet]

we're trying to keep up with the development of datasets and models: alexeyev/awesome-kyrgyz-nlp at GitHub

# Publicly Available Models: Specifically for Kyrgyz Language

1. The Cramer Project: Kyrgyz NER, Text-to-Speech model
2. Ulutsoft and Til Comission: Kyrgyz Text Completion (Mistral-7B-v0.1), Text-to-Speech
3. Kyrgyz NER model by Murat Jumashev

Try Pitch we're trying to keep up with the development of datasets and models: **alexeyev/awesome-kyrgyz-nlp** at GitHub

# Our Contributions  So Far [1/4]: Dataset Papers

**Morphological segmentation dataset presented in**

Садыков, Т. С., Туратали, Т., and Турдубаева, А. Б. Морфологический анализ для сбора текстовых данных в Национальном корпусе кыргызского языка. In TurkLang, Buxoro, 2023.

**News texts classification** (on re-annotation; planning to run a data science competition)

Alekseev, A., Nikolenko, S., & Kabaeva, G. (2023). Benchmarking Multilabel Topic Classification in the Kyrgyz Language. In Int. Conf. on Analysis of Images, Social Networks and Texts (pp. 21-35). Cham: Springer Nature CH.

**Word Similarity Dataset** (on re-annotation; planning to run a data science competition)

А. Алексеев, Г. Кабаева. Hj-ky-0.1: набор данных для оценки качества векторных представлений слов кыргызского языка. Herald of KSTU, 4, 2023

**Kyrgyz Named Entity Recognition** (the paper is being prepared)

# Our Contributions  So Far [2/4]: Research in Progress

- **Kyrgyz Named Entity Recognition** (3rd AI Datathon dataset; the paper is being prepared)

- Ongoing student projects @ SPbU:
  - **ML-Based Aid for Kyrgyz Dependencies Annotation**
  - **Transferring NER Labels to the Translated Dataset**
  - **New Results on Kyrgyz News Texts Classification**
    translation into English, Kyrgyz fastText, more tuning (e.g. mGPT-1.3B-kirgiz) and prompting

**STAY TUNED!**

- Plans to **setup up a website**: a curated list of Kyrgyz NLP papers based on this talk:
  bibliographical items, citations, search, analytics, automatic translations into English, etc.
  This will hopefully **increase the visibility** of research and improve the **understanding of the field**

# Our Contributions  So Far [3/4]: Online Resources & Services

**The Cramer Project**:

- Text-to-Speech dataset and model
- Datasets: 300m corpus, Kyrgyz MNIST, Kyrgyz NER, Kyrgyz Alpaca

Try Pitch

# Our Contributions  So Far [4/4]: Online Resources & Services

- LLM (fine-tuning Llama)
- Kyrgyz spell checker model (based on BERT)
- Kyrgyz language evaluation benchmark (MMLU, Reading Comprehension)
- The Kyrgyz National corpus TBA

# How LLMs Perform

Bonus Slide

# How do LLMs for Kyrgyz Perform?

| | MMLU (Math + History) |
|---|---|
| ChatGPT-o | 93% |
| Claude Sonnet 3.5 | 93% |
| Llama3-70B | 71% |
| Llama3.1-70B | 73% |

MMLU in accuracy

# Community Efforts

El Pueblo Unido

# Community Efforts

1. **The Cramer Project**
    a. Kyrgyz NER model :
        volunteers, KSTU students,
        thanks to G. Kabaeva & G. Jumalieva
    b. LLM dataset
2. **Aya** dataset by Cohere:
    community-annotated
3. **Kyrgyz MNIST** dataset:
    handwritten letters of the alphabet
    annotated by schoolchildren, curated by I. Jumaev
4. **AkyAI** Assistant:
    data collection and testing

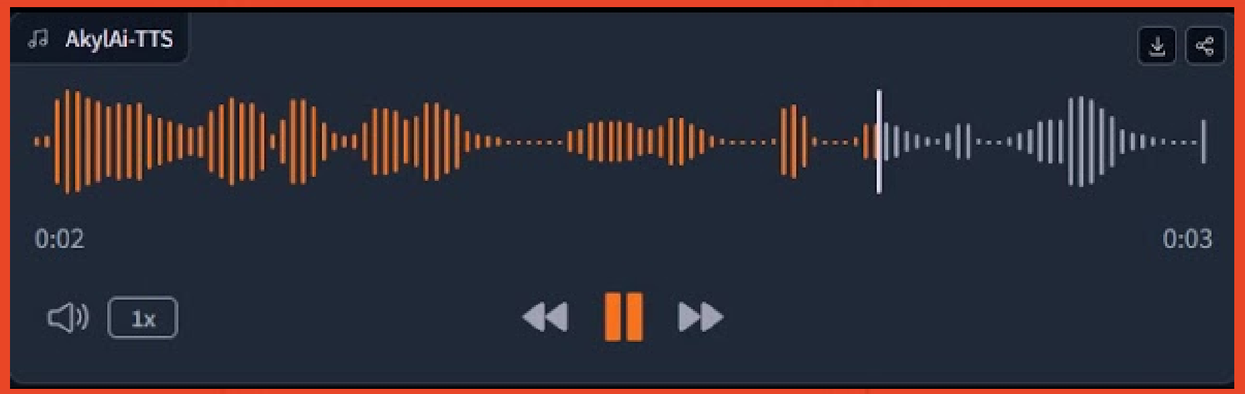
**100+**
volunteers

**15**
AI experts

**50**
linguistics

HIGH TECHNOLOGY PARK of the Kyrgyz Republic
Kyrgyz State Technical University named after I.Razzakov
!nkubasia
European Bank for Reconstruction and Development
accelerate prosperity
tilimpõz

# EXAMPLE OF COMMUNITY EFFORTS

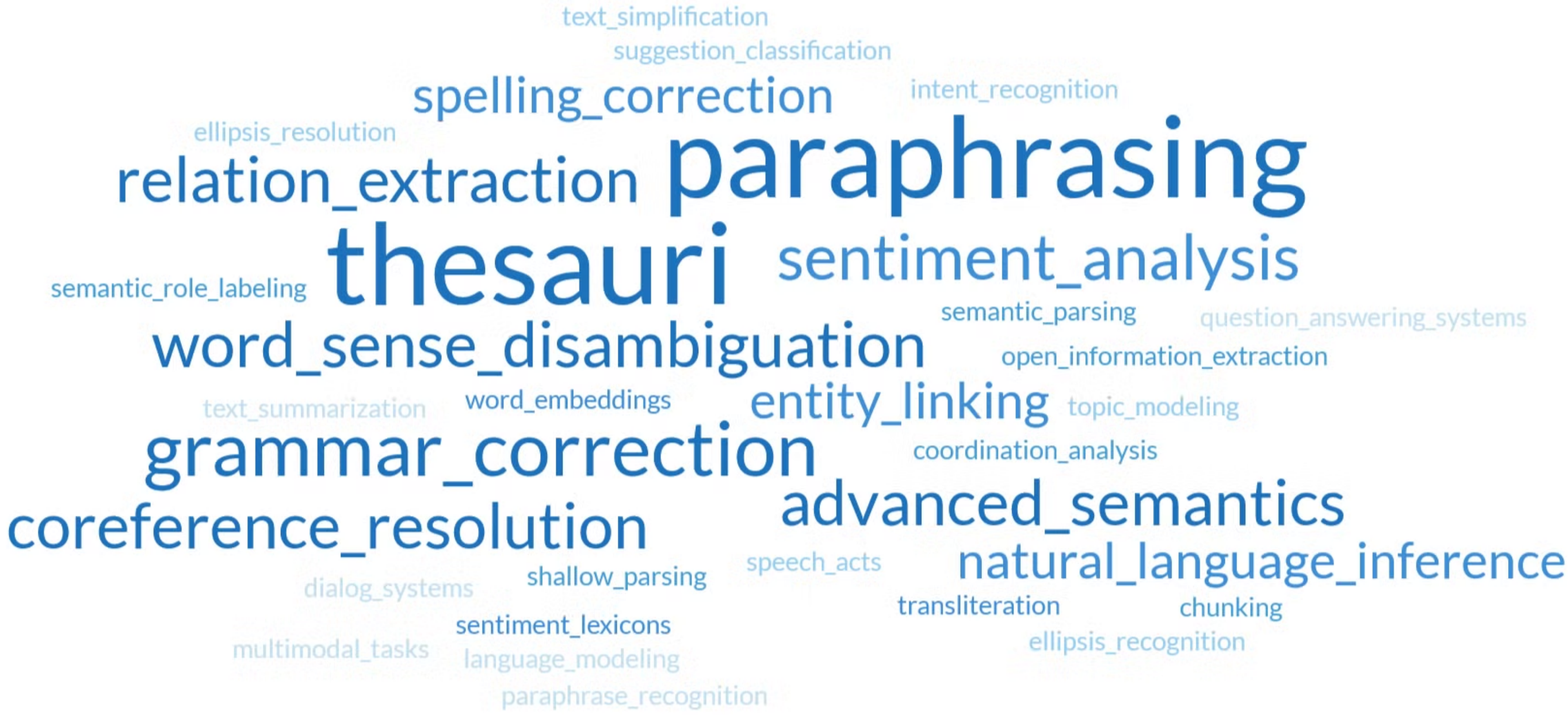




Try Pitch

# Roadmap for Kyrgyz NLP

Humble Proposal

# The Blue Ocean of the Unexplored: Enough Room for Everyone



# Roadmap

## What we suggest to pay attention to most closely

# Roadmap: Milestones

**Data Collection and Annotation**

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

**Community Involvement and Ecosystem Development**

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

**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

**Deployment and Adoption**

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

# Roadmap: Timeline

| | 2025 | 2026 | 2027 | 2028 |
|---|---|---|---|---|
| Best Scenario (with Government Support) | Creation of Kyrgyz corpora | Expand dataset with spoken language | Kyrgyz NLP integrated into educational systems | |
| | Annotation work with universities | Advanced models (NER, machine translation) | Full model maturity and wide adoption | |
| | Frist open models for POS tagging and translation | Public sector integration | | |
| Optimal Scenario (without full government involvement, but with private sector support) | Private companies (e.g., tech firms) fund data collection | Expand data collection with volunteers and collaborations with universities | Continuous improvements in models (NER, sentiment analysis) | |
| | Launch a small-scale open-source Kyrgyz NLP platform | Develop Kyrgyz-Russian and Kyrgyz-English translation models | Kyrgyz-language tech products (smart speakers, apps) enter the market | |
| | First models for POS tagging, basic translation | Initial industry adoption: Kyrgyz chatbots and customer support tools for businesses | Limited adoption in the public sector; mostly driven by private companies | |
| Worst Scenario (minimal support) | Data collection is slow and mainly community-driven (crowdsourcing) | Slow development of additional models (NER, sentiment analysis) | Limited tech applications; Kyrgyz remains underrepresented in digital tools | |
| | Limited progress due to lack of funding and minimal resources | Minimal industry adoption, mostly within academic and research circles | Some isolated academic projects continue, but with slow progress | |
| | Development of basic models for POS tagging and translation, but at a slow pace | | Kyrgyz NLP stagnates without sufficient funding or interest | |

# Conclusion

## Don't Say Goodbye

01

*Need More Minerals*

Kyrgyz NLP needs more research, data, and collaborative activity to reach its potential

02

*A Humble Proposal*

We share our view on the directions of Kyrgyz NLP that should be paid attention to for the sake of sustainable advancement of the field and the related applications

03

*El Pueblo Unido*

The only chance for Kyrgyz language to stop being *Scraping By* is the close collaboration of the interested parties; taking this opportunity, we invite you to cooperate

**The good news is that there is enough room for everyone in Kyrgyz NLP!**

# Thank you for your attention!
# Көңүл бургандыңызга рахмат!

# Questions and answers time?

The 12th International Conference on Analysis of Images, Social Networks and Texts

# Kyrgyz NLP: Challenges, Progress, and Future

**Anton Alekseev** & **Timur Turatali**

PDMI RAS & SPbU (St. Petersburg, Russia), KFU (Kazan, Russia)

KSTU n.a. I. Razzakov, The Cramer Project (Bishkek, Kyrgyzstan)


Russian Science Foundation

The work of A. Alekseev was supported by the Russian Science Foundation grant # 23-11-00358

**Bishkek, 2024**