Recent Advancements and Challenges of Turkic Central Asian Language Processing

Yana Veitsman, Mareike Hartmann
Department of Language Science and Technology
Saarland University, Germany
{yanav, mareikeh}@lst.uni-saarland.de

Abstract

Research in NLP for Central Asian Turkic languages - Kazakh, Uzbek, Kyrgyz, and Turkmen - faces typical low-resource language challenges like data scarcity, limited linguistic resources and technology development. However, recent advancements have included the collection of language-specific datasets and the development of models for downstream tasks. Thus, this paper aims to summarize recent progress and identify future research directions. It provides a high-level overview of each language's linguistic features, the current technology landscape, the application of transfer learning from higher-resource languages, and the availability of labeled and unlabeled data. By outlining the current state, we hope to inspire and facilitate future research.

1 Introduction

Turkic languages are spoken by approximately 200 million people worldwide, with a significant concentration in Central Asia (see detailed breakdown of the number of speakers in Figure 1). While Turkish is the most resourceful language in the family, this paper focuses on less-resourced Turkic languages that are geographically, historically, and linguistically closer to one another in Central Asia. These languages represent an important subset of the Turkic family, spoken by approximately 80 million people in the region.

Like any other low-resource language speakers, speakers of Central Asian languages would benefit from having reliable language technology, from simple spell checkers to virtual assistants. Such tools would uphold newly adopted language policies and cement the role of local languages in the region. Developing these resources, however, requires the existence of open-source datasets and up-to-date language models. To address these resource limitations, researchers are exploring methods like transfer learning and data augmentation, though both have limitations in task applicability and effectiveness (Chen et al., 2021; Raffel et al., 2019).

This paper aims to provide an overview of existing resources and suggest directions for future research to support both those utilizing current resources and those developing new ones (for the search strategy details see Appendix A). We also seek to highlight current resource needs, addressing which of those could be particularly crucial in advancing the Turkic Central Asian languages toward a higher-resource status.

2 Related Work

Recently, substantial efforts have been made to consolidate domain knowledge for Turkic languages via linguistic analysis tools (Akın and Akın, 2007; Abdurakhmonova et al., 2022), and NLP technology assessments (Mirzakhalov et al., 2021; Maxutov et al., 2024). However, there is still a lack of comprehensive research summarizing the available data and language processing tools, especially for Central Asian Turkic languages. While state-of-the-art advancements in speech recognition and machine translation exist for some languages (Bekarystankyzy et al., 2024; Yeshpanov et al., 2024b), no cross-linguistic comparisons have been conducted. A detailed survey could provide a valuable foundation for comparison and help define future research directions.

3 Difficulties in Processing Turkic Languages

3.1 Overview

Typology has a potential to improve language processing and transfer learning (Ponti et al., 2019), and learning from similar languages is overall beneficial for the latter (Zoph et al., 2016). Therefore, providing an overview of the linguistic features of Turkic languages may help identify potential simi-

Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025), pages 345–360
January 20, 2025.

Feature	Turkish	Kazakh	Kyrgyz	Uzbek	Turkmen
Number of vowels	8	12	8	6	9
Number of plural suffixes	2	12	12	4	4
Number of pronouns	6	8	8	8	6
Number of noun cases	6	7	6	6	6
Number of personal verb suffixes	5	9	6	9	5
Word order	SOV	SOV	SOV	SOV	SOV

Table 1: High-level overview of the Turkic Central Asian languages' differences. Sources: https://ecosystem.education/doc/Turkic%20Diller-SS.pdf, https://www.britannica.com/topic/Turkic-languages

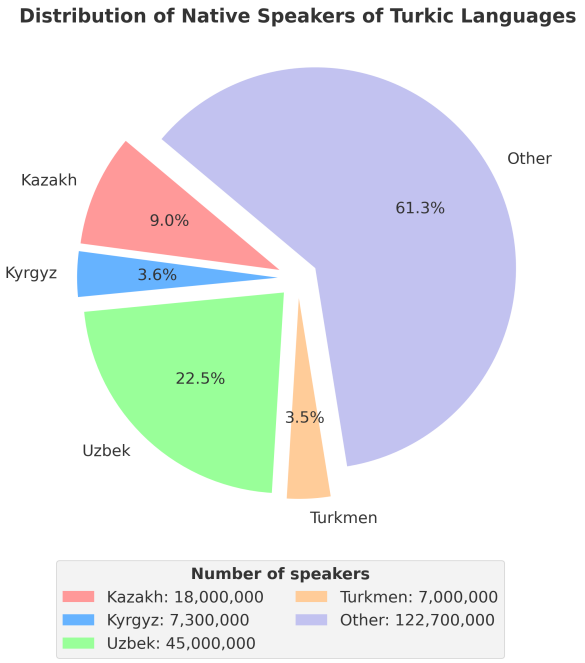

Figure 1: Distributions of numbers of Kazakh, Uzbek, Kyrgyz, and Turkmen native speakers among all Turkic language speakers. Numbers in the legend are approximates. Source: https://en.wikipedia.org/w/index.php?title=Languages_of_Asia&oldid=1230214231

larities and challenges in their processing. While in depth linguistic comparison of each language with their richer counterpart, Turkish (Çöltekin et al., 2022), lies beyond the scope of the paper, a basic summary is provided in Table 1.

Turkic languages are primarily described as morphologically rich. Over the past years, several studies revolving around the specifics of processing their complex morphological structures and morpho-syntactic features have been conducted (Ciddi, 2013). These studies showed that the highly agglutinative nature of Turkic languages is specifically problematic in machine translation (Alkim and Çebi, 2019; Mirzakhalov et al., 2021) and named entity recognition (Küçük et al., 2017) tasks. Additionally, it is not exactly clear how these peculiarities will be reflected in the transfer learning applications.

3.2 Similarities and Differences

Many grammatical features that impact natural language processing - such as word order and verb tense systems - are common across all the languages considered. Word order similarity is particularly crucial because of its impact on multilingual learning (Dufter and Schütze, 2020). On the other hand, shared verb tense systems can facilitate cross-lingual data enrichment (Asgari and Schütze, 2017), important for transfer learning. Nevertheless, minor differences among the languages do exist, including a heavier reliance on vowel harmony in Kazakh, Uzbek, Kyrgyz, and Turkmen. This results in additional noun cases and plural suffixes, with the exact form depending on the vowel of the preceding syllable.

Analyzing the linguistic differences in Table 1, such as variations in the number of pronouns or vowel sounds, one can infer that Kazakh and Kyrgyz are typologically closer to each other than Kazakh is to Uzbek, or Uzbek is to Kyrgyz; similarly, Turkmen is grammatically closest to Turkish. This suggests not only potential differences in transfer learning efficacy from Turkish, but also the possibility of successful transfer learning within the Central Asian group itself - likely making transfer between certain pairs, such as Kazakh and Kyrgyz, more effective than between others.

Another significant difference among these languages is the script used and its inconsistent application. For example, while Uzbek is written in Latin script, Kazakh, despite efforts by the Kazakhstani government to adopt Latin script, is still primarily written in Cyrillic. Such script inconsistencies could limit the effectiveness of transfer learning (Gheini and May, 2019; Amrhein and Sennrich, 2020), suggesting that additional steps in

data cleaning or preprocessing may be necessary to optimize model performance.

Overall, these observations offer a promising opportunity to leverage linguistic similarities for transfer learning both from Turkish and Turkic Central Asian languages themselves.

4 Datasets Availability

Open-source language resources enable researchers to scale and reuse data, reducing overall time spent on corpora collection and evaluation. However, data access varies from low to almost non-existent for the languages considered in the paper. A full summary of all the reviewed datasets and their primary features can be found in the Appendix B.

4.1 Sources of Data and Stakeholders

While developing a comprehensive dataset classification system would lie beyond the scope of the current paper, it is important to identify potential data collection sources and main stakeholders of the process. While a significant effort has been made by the researchers within the Central Asian countries themselves (e.g., most notable research on the Kazakh language is contributed by the Institute of Smart and Intelligent Systems (ISSAI) of Nazarbayev University), outside effort is driven primarily by the studies of NLP applications for Chinese minority languages (Du et al., 2023) or Turkic languages in general (Mirzakhalov et al., 2021). Additionally, while the researchers focus on collection of handcrafted high-quality datasets, other, potentially lower-quality data, is available via multilingual datasets crawled from the web, such as Common Crawl (CC)[1] or OSCAR (Ortiz Suárez et al., 2020). One more potential data source is the machine-translated content from high-resource languages; however, assessing the exact quantity of such data poses to be challenging, so the survey does not focus on it. Thus, there is a greater variety of data sources and stakeholders than might initially appear.

4.2 Kazakh Language Datasets

Unannotated Datasets and Corpora Collections. One of the largest available dataset of unannoted text articles in Kazakh is the Kazakh Language Corpus (Makhambetov et al., 2013), containing a broad range of written text on a variety of topics. Additionally, annotated sub-corpora with linguistics and other language features are available within the same dataset. Another large collection of texts is the Kazakh National Language Corpus[2] with over 22 thousand documents available.

Linguistic Features Datasets. Among the four surveyed languages, Kazakh has the most diverse and extensive datasets available. Besides collections of morphologically and syntactically-annotated data like Almaty Corpus of Kazakh Language[3] and UD treebank (Tyers and Washington, 2015; Makazhanov et al., 2015), there exist also more task-specific text corpora.

Task-specific Datasets. For example, the KazNERD corpus, spanning over 100,000 sentences with 25 entity classes for the task of Named Entity Recognition (Yeshpanov et al., 2022), the KazQAD corpus for open-domain question answering with over 6000 questions (Yeshpanov et al., 2024a), the KazParC parallel corpus of Kazakh, English, Russian, and Turkish for machine translation (Yeshpanov et al., 2024b), and the KazSAnDRA dataset, containing over 180 thousand reviews to be used for the task of sentiment analysis (Yeshpanov and Varol, 2024), are all publicly available.

Multimodal Datasets. There are also numerous collections of multimodal data for the language. Most notable and largest language dataset up until now is the Kazakh Speech Corpus 2 (KSC2), collected by Mussakhojayeva et al. (2022b), which contains over 1,200 hours of transcribed audios, and which also integrates research work on previously collected datasets, such as Kazakh Speech Corpus (KSC), compiled by Khassanov et al. (2021). Another prominent resource is the KazakhTTS2 dataset (Mussakhojayeva et al., 2022a), which extended its previous version, KazakhTTS (Mussakhojayeva et al., 2021), to more than 270 hours of recorded audio. Drawing from this work on speech, a speech commands dataset was also recently collected (Kuzdeuov et al., 2023). Additionally, more multimodal data is available in Kazakh for even more peculiar applications; for example, Mukushev et al. (2022) have gathered over 40,000 video samples of 50 signers of Kazakh-Russian Sign Language, which was also an expansion of a previous project by the same authors in 2020. Further work on the topic includes recent development of the KazEmoTTS dataset (Abilbekov et al., 2024) that collects more than 54 thousands

[1] https://www.commoncrawl.org

[2] https://qazcorpora.kz

[3] http://web-corpora.net/KazakhCorpus/search/?interface_language=en

audio-text pairs in over 5 emotional states. Several datasets were also collected for the purpose of handwritten text recognition, some of them combining Kazakh and Russian languages (Abdallah et al., 2020; Nurseitov et al., 2021), and some containing text inclusively in Kazakh, for example, Kazakh Offline Handwritten Text Dataset (Toiganbayeva et al., 2022), with over 3000 exam papers available. This variety of multimodal data can potentially cover a wide range of the data needs, enabling efficient data reuse and eliminating the necessity to use precious resources for additional data collection. For example, the audios from the speech corpus can be transcribed and used as text for the tasks of text classification, text generation, information retrieval, and others. Thus, these datasets have a great potential of making a strong contribution in development of other NLP tasks.

4.3 Uzbek Language Datasets

Uzbek ranks second in terms of data availability among Central Asian languages; however, data that exclusively covers this language remains relatively scarce comparing even to the Kazakh language alone.

Linguistic Features Datasets. A distinguishing feature of most Uzbek datasets is their focus on purely linguistic tasks; for example, several datasets, such as UzWordnet (Agostini et al., 2021) and SimRelUz (Salaev et al., 2022) capture prominent semantic features of the language; others, such as a dataset collected by Sharipov et al. (2023), that includes a variety of POS tags and syntactic features of this Central Asian language.

Task-specific Datasets. There is substantial data for certain tasks, such as sentiment analysis. For example, there exist a sentiment analysis dataset collected by Kuriyozov et al. (2019) and one based on restaurant reviews by Matlatipov et al. (2022). There also exists data for text classification, mainly scraped from the news sources in Uzbek, one collected by Rabbimov and Kobilov (2020) and another one by Kuriyozov et al. (2023).

Other Datasets. Some additional datasets, including multimodal and/or general unannotated data also exist. For example, one of the multimodal datasets is the open-source Uzbek Speech Corpus (Musaev et al., 2021). On the other hand, there is the Uzbek corpus[4], which includes a large collection of educational, scientific, official, and artistic texts together with a morphological database and dictionaries, and an Uzbek Community corpus[5], collected by Leipzig University and containing over 660 thousand sentences of community data only.

4.4 Kyrgyz Language Datasets

Unannotated Datasets and Corpora Collections. One of the largest open-source datasets for Kyrgyz is the Manas-UdS corpus[6] of over 84 literary texts in 5 genres, marked with lemmas and parts of speech (Kasieva et al., 2020). For most other datasets, linking them directly to publications is not possible; however, some of them are available on Github. One of those is the KyrgyzNews dataset[7] with over 250 thousand scraped news. Leipzig University has also compiled a corpus of over 3 million sentences from publicly available web sources.[8]

Task-specific datasets. The few available task-specific datasets also can only be found on Github. One such example, for the task of NER, is the NER dataset[9] that is currently under development by the researchers from Kyrgyz State Technical University (KSTU). Another example is the hand-written letters dataset[10] (Kyrgyz MNIST equivalent) also available on Github.

4.5 Turkmen Language Datasets

The data landscape of Turkmen language is even more scarce. Besides the Leipzig University's corpora[11] of over 270 thousand sentences of web-scraped data, other resources are practically non-existent. With only a few dictionaries and poetry and literature collections[12] collected by enthusiasts and available in an open-source fashion on Github, Turkmen language falls largely behind every other Turkic language of Central Asia in terms of data availability.

4.6 Web-Scraped Datasets

All the above-mentioned languages are also represented in multilingual datasets predominantly

[4] https://uzbekcorpus.uz/

[5] https://corpora.uni-leipzig.de/en?corpusId=uzb_community_2017

[6] https://fedora.clarin-d.uni-saarland.de/kyrgyz/index.html

[7] https://github.com/Akyl-AI/Kyrgyz_News_Corpus

[8] https://corpora.uni-leipzig.de/de?corpusId=kir_news_2020

[9] https://github.com/Akyl-AI/KyrgyzNER

[10] https://github.com/Akyl-AI/kyrgyz_MNIST

[11] https://corpora.wortschatz-leipzig.de/en?corpusId=tuk-tm_web_2019

[12] https://github.com/tmLang-NLP/datasets

scraped from the web. In particular, all the languages are available in CC100 (Wenzek et al., 2020), WikiAnn (Pan et al., 2017), and OSCAR (Ortiz Suárez et al., 2020) datasets. Most of other web-scraped datasets present online are either subsets of the bigger multilingual datasets (like CC100 or OSCAR), or are scrapes from newspapers and websites available on the Internet, with many of them not being open-source (for example, the kkWaC dataset[13] with over 139 million Kazakh words). However, as noted by Kreutzer et al. (2022), such datasets should be used with caution, as the quality of low-resource language data in them may be significantly lower. With only a few hundred labeled instances, even a 5-10 percent rate of mislabeled or grammatically incorrect sentences can substantially impact model performance in these languages.

4.7 Multilingual Datasets

Another important source of data in Central Asian languages is handpicked multilingual datasets for Turkic languages. For example, a large corpus was collected by Baisa and Suchomel (2012) for training morphological analyzers and disambiguators. Another example is the xSID (Cross-lingual Slot and Intent Detection) dataset by van der Goot et al. (2021), which includes Turkic languages among other language families. More task-specific multilingual datasets featuring Central Asian languages include the Common Voice dataset (Ardila et al., 2020), the Belebele dataset for machine reading comprehension containing Kazakh, Uzbek, and Kyrgyz (Bandarkar et al., 2023), the MuMiN dataset (Multimodal Fact-Checked Misinformation Dataset) based on the scraped tweets featuring Kazakh language (Nielsen and McConville, 2022), and the AM2iCo dataset for the evaluation of word meaning in context (Liu et al., 2021). Thus, Kazakh, Uzbek, and Kyrgyz make a prominent appearance in both the web domain and Turkic languages related research.

4.8 Parallel Corpora

Additionally, we would like to comment on the availability of parallel corpora for the languages studied. While OPUS provides a comprehensive overview of parallel data available (Tiedemann and Thottingal, 2020) as well as several models (Tiedemann et al., 2023), only a few less resourceful parallel pairs datasets exist, including the already mentioned Kazakh-Russian sign language corpora, Uzbek-Kazakh (Allaberdiev et al., 2024) and Kazakh-Russian (Kozhirbayev and Islamgozhayev, 2023) parallel corpora for MT and ASR.

4.9 Classifying Languages by Data Availability

The resources available for each language are far from abundant. If one were to categorize them according to the classification suggested by Joshi et al. (2020), Kazakh would most likely fit into "The Rising Star" category, with its substantial presence in the web and good variety of multimodal datasets. However, the research on this language is pushed back by a lack of labeled data for 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.

5 Reasons of Data Scarcity

Data scarcity in Central Asian languages stems from the widespread use of Russian, limited internet access, and a lack of AI-focused educational and technological initiatives.

5.1 Russian Language in Central Asia

During the Soviet era, Russian was the dominant language across all republics. National government since then have promoted national languages, but Russian remains influential in science, education, and politics (Fierman, 2012). Russian media and online resources are widely accessible in Central Asia, facilitating intercultural communication but limiting the development of NLP for local languages. Heavy reliance on Russian sources for web-scraped data and media results in limited Kazakh, Uzbek, Kyrgyz, and Turkmen content online, restricting data diversity for these languages.

5.2 Internet Access

Limited Internet access is the second reason for data scarcity in the region. As stated before, while

[13] https://www.sketchengine.eu/kkwac-kazakh-corpus/

a substantial amount of data has been gathered by NLP researchers from the Internet based sources, only 38 and 21 percent of users in Kyrgyzstan and Turkmenistan respectively have access to the global net.[14] While the situation is somewhat better in Uzbekistan (with 55 percent of population being able to access the medium) and significantly better in Kazakhstan (around 79 percent of population), limited connectivity hinders users' abilities to contribute to open-source encyclopedias, blogs, news sites, and more. Additionally, the lack of digitalized resources, such as electronic books, journals, audio transcripts, and video recordings, restricts information sharing within the region.

5.3 Lack of Initiatives

Being a demanding and resource-greedy field, natural language processing requires substantial and long-term financial investments. However, only a few exclusively AI-dedicated initiatives have been launched by the governments: for example, ISSAI (Institute of Smart Systems and Artificial Intelligence) in Kazakhstan or High Technology Park in Kyrgyzstan. However, these institutions are not solely dedicated to NLP research, and cover a wide range of topics in the AI domain, including robotics, IoT, computer vision, and others. Consequently, without a tailored initiative, it is difficult for the researchers to specialize in NLP-related research only.

6 Application of Transfer Learning

The situation when one of the languages in a language family is more resourceful than others is not unique to Turkic languages. This opens the door for potential transfer learning from Turkish or one of the Central Asian languages to the other languages within the same family. Usually, this is more attainable than collecting large datasets from scratch.

Transfer learning has been substantially studied in the domain of machine translation, and the choice of parent language has been highlighted as an important criteria for the technique application (Zoph et al., 2016). Combining transliteration and byte-pair encoding, Nguyen and Chiang (2017) proved that this transfer learning approach might be suitable beyond high-resource to low-resource pairs, extending to pairs within low-resource only, especially the ones belonging to the agglutinative languages. The greater availability of NLP tools for Kazakh made it possible to assess transfer learning potential for some Turkic laguages lying beyond the scope of this paper, namely, Tatar (Valeev et al., 2019). Other studies on transfer learning from Kazakh have been conducted on the task of ASR (Orel et al., 2023). Some of them also aimed at using Russian as a source language, but the efforts proved to be less successful (N. et al., 2020).

7 Data Augmentation, Transliteration, and R-Drop Regularization

Apart from transfer learning, data augmentation techniques, R-Drop regularization, and transliterations have been substantially researched for enhancing model performance in some Turkic languages. A study on the topic of sentence augmentation using large language models for Kazakh (Bimagambetova et al., 2023) demonstrated that data augmentation works well for already resourceful languages and does so less successfully in the low-resource domain, which might seem somewhat obvious. However, practical applications using other data augmentation techniques, including phrase replacement, proved to significantly improve the BLEU score of certain language pairs, for example, Kazakh-Chinese (Wu and Ma, 2023). Data augmentation with R-Drop regularization has also proved useful for the same Kazakh-Chinese translation task in other studies (Liu et al., 2023).

Other techniques, such as dropout and transliteration, are often applied alongside transfer learning and data augmentation. Furthermore, multilingual models may outperform those using only transfer learning on certain tasks (Nugumanova et al., 2022). Altogether, these various techniques allow researchers to experiment with Central Asian language processing without requiring extensive data collection.

8 Current State of Technologies

8.1 Kazakh Language Technologies

In terms of available technology, Kazakh ranks first, just as it does in data availability.

Linguistic Analysis and Rule-Based Systems. A variety of tools have been developed for linguistic analysis and normalization of texts in Kazakh (Yessenbayev et al., 2020). Early on, rule-based translation systems and morphological analyzers

[14] https://blogs.worldbank.org/en/europeandcentralasia/how-central-asia-can-ensure-it-doesnt-miss-out-digital-future

Task	Kazakh	Uzbek	Kyrgyz	Turkmen
Automatic Speech Recognition	✓	✓	✓	✗
Machine Translation	✓	✓	✓	✓
Named-Entity Recognition	✓	✓	✓	✗
Text Generation	✓	✗	✗	✗
Sentiment Analysis	✓	✓	✗	✗
Text Classification	✓	✓	✓	✗
POS Tagging	✓	✓	✓	✗
Text Summarization	✓	✓	✗	✗
Question Answering	✓	✗	✗	✗

Table 2: Existence of downstream task research per language. Existence is defined as at least one published research paper and/or dataset for the specific language and/or the specific language in conjunction with other closely related languages.

have also been developed (Forcada and Tyers, 2016). This research paved the way for the first advances in the most prominent areas of NLP, like machine translation and ASR, that are well represented in Kazakh language.

Machine Translation. Almost all the datasets mentioned in the previous sections have been released together with the evaluation benchmarks for certain tasks on either already pre-trained models like mBERT (Yeshpanov et al., 2022) or together with completely new systems like Tilmash, which enables a two-way translation between for 4 languages, including Kazakh and Turkish (Yeshpanov et al., 2024b). The latter has proved to be comparable or even better at translating language pairs involving Kazakh than translation technologies developed by Google and Yandex, which dominate the scene of machine translation in the region.

ASR. Another important advancement in processing of Kazakh language lies in the sphere of automatic speech recognition. For that purpose, researchers have been actively leveraging the KazakhTTS and KazakhTTS2 datasets as well as the recently available KazEmoTTS dataset. For example, a Turkic ASR system that employs the KazakhTTS, USC, and Common Voice data and covers, among other Turkic languages, Kazakh, Uzbek, and Kyrgyz, has been developed by the authors of the KazakhTTS dataset (Mussakhojayeva et al., 2022a). The most recent research on exclusively Kazakh language managed to bring down the word error rate to just 7.2 percent (Bekarystankyzy et al., 2023) as well as to summarize difficulties and explore deep learning techniques on end-to-end speech recognition of agglutinative languages (Bekarystankyzy et al., 2024). Additionally, research on speech commands recognition has been conducted (Kuzdeuov et al., 2023).

Fine-tuning and Assessing Existing Models. Regarding the usage and fine-tuning of already existing models, several tasks have been evaluated for Kazakh. For example, on the KazNERD dataset (Yeshpanov et al., 2022), the fine-tuned XLM-RoBERTa demonstrated a micro average precision score of an impressive 97.09 percent, and on the task of sentiment analysis the same model gained an F1 score of 0.87. Additionally, Maxutov et al. (2024) assess the capabilities of 7 LLMs, including GPT-4 and Llama-2, on a variety of tasks, from classification to question answering, concluding that the performance of the models, just as expected. is lower on Kazakh language tasks in comparison to that on English language ones.

Understudied Areas. However, despite the above-mentioned efforts, certain areas of Kazakh NLP remain significantly understudied. One such example is the particularly dynamic field of text generation. Only a few experiments have been conducted regarding the usage of large language models for the benefit of Kazakh language (Tolegen et al., 2023; Maxutov et al., 2024).

8.2 Uzbek Language Technologies

Similar to data availability, Uzbek ranks second in terms of available technology.

ASR. While Uzbek does not enjoy the same variety of machine translation tools as Kazakh, presence of automatic speech recognition technology is somewhat comparable with the most recent work contributed by Musaev et al. (2021), with their model performing at the level of 14.3% word error rate.

Fine-Tuning Existing Models. Uzbek also enjoys a better variety of pre-trained and fine-tuned models, including UzBERT (Mansurov and Mansurov,

2021), capable of outperforming mBERT by leveraging at least 11 times more language specific data; UzRoberta, pre-trained on roughly 2 million news articles (Davronov and Adilova, 2024); BERTBek, a model improving on UzBERT by training on Latin script (Kuriyozov et al., 2024), and a compact and fine-tuned variation of a mT5 model[15]. In general, with Uzbek still catching up on data availability, there is potential for the language to soon enjoy a better variety of NLP technologies.

8.3 Kyrgyz and Turkmen Languages Technology

Unfortunately, the situation for Kyrgyz and Turkmen looks less promising. With little to no work in terms of machine translation or automatic speech recognition, Kyrgyz enjoys little variety of technology for text classification (Alekseev et al., 2023) and NER fine-tuned on WikiAnn data[16]. Turkmen offers even less language-specific technology, since it is has been mainly studied within the scope of comparative studies of Turkic languages. For example, some work on machine translation evaluation was done in the effort to build the infrastructure for Turkic languages, but nowadays this approach seems outdated (Alkim and Çebi, 2019). Overall, neither fine-tuning or pre-training of already available models and architectures have been researched for the majority of basic NLP tasks both in Kyrgyz and Turkmen.

9 Future Work Areas

Kazakh Language. Due to its variety of multimodal data, Kazakh language can potentially expand on the existing work and move onto fine-tuning and developing more advanced models for other downstream tasks, e.g. text generation or question answering. Additionally, transfer learning from Kazakh should be further explored as a potential workaround for the problems of its less resourceful counterparts, with a bigger focus on linguistically closer languages of the Turkic family, like Kyrgyz.

Uzbek Language. In contrast, Uzbek language clearly requires more data to be collected for the creation of the systems like the ones built on top of datasets like KazakhTTS. With already developed pre-trained models like UzBERT, there is a need for research into their further application and comparison with other models available. Generally, leveraging the rich Uzbek linguistic features datasets in combination with further efforts of large-scale data collection paint a bright future for the language.

Kyrgyz and Turkmen Languages. The current situation for Kyrgyz and Turkmen requires significant efforts for data collection and aggregation first. With the amount of data available for both languages, it seems barely useful not only to pre-train models like BERT, but also to research potential applications of statistical algorithms. In parallel with data collection, studies on transfer learning from Kazakh or Turkish might prove beneficial, given the grammatical similarity between the corresponding language pairs.

Additionally, given the incorporation of LLMs in the field, potential of using them for data augmentation or annotation can be another important direction of research in comparison with more expensive methods of human data collection and annotation.

10 Conclusion

Significant improvements in the processing of Turkic Central Asian languages have been achieved in the recent years. However, there still exists an imbalance in the amount of available data and technology, with Kazakh and Uzbek dominating the scene, and Kyrgyz and Turkmen requiring significant efforts in terms of data collection. Besides leveraging already existing datasets or web-sources for curating new data using data augmentation, other potentially successful options for technological advancements include transfer learning from Kazakh to Uzbek, Kyrgyz, and Turkmen. Already existing data in Kazakh might prove useful for the studies of other Central Asian languages, given their close linguistic relatedness.

While significant efforts are yet to emerge and be maintained for Kazakh, Uzbek, Kyrgyz, and Turkmen to reach the level of winners or underdogs (Joshi et al., 2020), a substantial work has already been developed in the most crucial areas of natural language processing, including the automatic speech recognition systems and machine translation. With the potential of transferring this experience to other tasks, such as information retrieval, text generation, and question answering, there is a greater hope for progress towards the state-of-the-art technology for these Central Asian languages.

[15] https://ijdt.uz/index.php/ijdt/article/view/104

[16] https://huggingface.co/murat/kyrgyz_language_NER

11 Limitations

We acknowledge several factors that limit the findings and scope of the presented survey. Firstly, we note that the rapid evolution of NLP technologies and data makes any work aimed at assessing current data sources and research developments outdated fairly quickly. Additionally, some previously published resources might become unavailable with time, which also contributes to the outdatedness. Secondly, we base the resource review process on the information provided by the authors of the relevant papers, which, however, might not reflect the real state of a dataset or a technology. There might exist some discrepancies between the data reported in the papers and those actually available due to access limitations, dataset updates, etc. Future work might address the above-mentioned limitations by establishing an open-source up-to-date list of resources and assessing the datasets quality empirically.

References

Abdelrahman Abdallah, Mohamed Hamada, and Daniyar Nurseitov. 2020. Attention-based fully gated cnn-bgru for russian handwritten text. *Journal of Imaging*, 6(12):141.

Nilufar Z. Abdurakhmonova, Alisher S. Ismailov, and Davlatyor Mengliev. 2022. Developing nlp tool for linguistic analysis of turkic languages. In *2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, pages 1790–1793.

Adal Abilbekov, Saida Mussakhojayeva, Rustem Yeshpanov, and Huseyin Atakan Varol. 2024. Kazemotts: A dataset for kazakh emotional text-to-speech synthesis. *Preprint*, arXiv:2404.01033.

Alessandro Agostini, Timur Usmanov, Ulugbek Khamdamov, Nilufar Abdurakhmonova, and Mukhammadsaid Mamasaidov. 2021. UZWORDNET: A lexical-semantic database for the Uzbek language. In *Proceedings of the 11th Global Wordnet Conference*, pages 8–19, University of South Africa (UNISA). Global Wordnet Association.

Ahmet Afsin Akın and Mehmet Dündar Akın. 2007. Zemberek, an open source nlp framework for turkic languages. *Structure*, 10(2007):1–5.

Anton M. Alekseev, Sergey I. Nikolenko, and Gulnara Kabaeva. 2023. Benchmarking multilabel topic classification in the kyrgyz language. *Preprint*, arXiv:2308.15952.

Emel Alkim and Yalçin Çebi. 2019. Machine translation infrastructure for turkic languages (mt-turk). *Int. Arab J. Inf. Technol.*, 16:380–388.

Bobur Allaberdiev, Gayrat Matlatipov, Elmurod Kuriyozov, and Zafar Rakhmonov. 2024. Parallel texts dataset for uzbek-kazakh machine translation. *Data in Brief*, 53:110194.

Chantal Amrhein and Rico Sennrich. 2020. On Romanization for model transfer between scripts in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2461–2469, Online. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124, Copenhagen, Denmark. Association for Computational Linguistics.

Vít Baisa and Vít Suchomel. 2012. Large corpora for turkic languages and unsupervised morphological analysis. https://api.semanticscholar.org/CorpusID:16550803.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *Preprint*, arXiv:2308.16884.

Akbayan Bekarystankyzy, Orken Mamyrbayev, and Tolganay Anarbekova. 2024. Integrated end-to-end automatic speech recognition for languages for agglutinative languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*

Akbayan Bekarystankyzy, Orken Zh. Mamyrbayev, DO Oralbekova, and Bagashar Zhumazhanov. 2023. Transfer learning for an integrated low-data automatic speech recognition system. *Bulletin of Almaty University of Energy and Communications*.

Zhamilya Bimagambetova, Dauren Rakhymzhanov, Assel Jaxylykova, and Alexander Pak. 2023. Evaluating large language models for sentence augmentation in low-resource languages: A case study on kazakh. *2023 19th International Asian School-Seminar on Optimization Problems of Complex Systems (OPCS)*, pages 14–18.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2021. An empirical survey of data augmentation for limited data learning in nlp. *Preprint*, arXiv:2106.07499.

Sibel Ciddi. 2013. Processing of turkic languages. Master's thesis.

Rifkat Davronov and Fatima Adilova. 2024. UzRoberta: A pre-trained language model for Uzbek. *AIP Conference Proceedings*, 3004(1):050001.

Wenqiang Du, Yikeremu Maimaitiyiming, Mewlude Nijat, Lantian Li, Askar Hamdulla, and Dong Wang. 2023. Automatic speech recognition for uyghur, kazakh, and kyrgyz: An overview. *Applied Sciences*, 13(1).

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

W. Fierman. 2012. Russian in post-soviet central asia: A comparison with the states of the baltic and south caucasus. *Europe-Asia Studies*, 64:1077 – 1100.

Mikel L. Forcada and Francis M. Tyers. 2016. Apertium: a free/open source platform for machine translation and basic language technology. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia. Baltic Journal of Modern Computing.

Mozhdeh Gheini and Jonathan May. 2019. A universal parent model for low-resource neural machine translation transfer. *Preprint*, arXiv:1909.06516.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Aida Kasieva, Jörg Knappen, Stefan Fischer, and Elke Teich. 2020. A new kyrgyz corpus: sampling, compilation, annotation. Poster at 42. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft.

Yerbolat Khassanov, Saida Mussakhojayeva, Almas Mirzakhmetov, Alen Adiyev, Mukhamet Nurpeiissov, and Huseyin Atakan Varol. 2021. A crowdsourced open-source kazakh speech corpus and initial speech recognition baseline. *Preprint*, arXiv:2009.10334.

Zhanibek Kozhirbayev and Talgat Islamgozhayev. 2023. Cascade speech translation for the kazakh language. *Applied Sciences (Switzerland)*, 13(15). Publisher Copyright: © 2023 by the authors.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Elmurod Kuriyozov, Sanatbek Matlatipov, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2019. Construction and evaluation of sentiment datasets for low-resource languages: The case of uzbek. In *Human Language Technology. Challenges for Computer Science and Linguistics - 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17-19, 2019, Revised Selected Papers*, volume 13212 of *Lecture Notes in Computer Science*, pages 232–243. Springer.

Elmurod Kuriyozov, Ulugbek Salaev, Sanatbek Matlatipov, and Gayrat Matlatipov. 2023. Text classification dataset and analysis for uzbek language. *Preprint*, arXiv:2302.14494.

Elmurod Kuriyozov, David Vilares, and Carlos Gómez-Rodríguez. 2024. BERTbek: A pretrained language model for Uzbek. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 33–44, Torino, Italia. ELRA and ICCL.

Askat Kuzdeuov, Shakhizat Nurgaliyev, Diana Turmakhan, Nurkhan Laiyk, and Huseyin Atakan Varol. 2023. Speech Command Recognition: Text-to-Speech and Speech Corpus Scraping Are All You Need. TechRxiv.

Dogan Küçük, Nursal Arıcı, and Dilek Küçük. 2017. Named entity recognition in turkish: Approaches and issues. In *International Conference on Applications of Natural Language to Data Bases*.

Canglan Liu, Wushouer Silamu, and Yanbing Li. 2023. A chinese–kazakh translation method that combines data augmentation and r-drop regularization. *Applied Sciences*, 13(19).

Qianchu Liu, Edoardo Maria Ponti, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2021. AM2iCo: Evaluating word meaning in context across low-resource languages with adversarial examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7151–7162, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aibek Makazhanov, Aitolkyn Sultangazina, Olzhas Makhambetov, and Zhandos Yessenbayev. 2015. Syntactic annotation of kazakh: Following the universal dependencies guidelines. a report. In *3rd International Conference on Turkic Languages Processing, (TurkLang 2015)*, pages 338–350.

Olzhas Makhambetov, Aibek Makazhanov, Zhandos Yessenbayev, Bakhyt Matkarimov, Islam Sabyrgaliyev, and Anuar Sharafudinov. 2013. Assembling the Kazakh language corpus. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1022–1031, Seattle, Washington, USA. Association for Computational Linguistics.

B. Mansurov and A. Mansurov. 2021. Uzbert: pretraining a bert model for uzbek. *Preprint*, arXiv:2108.09814.

Sanatbek Matlatipov, Hulkar Rahimboeva, Jaloliddin Rajabov, and Elmurod Kuriyozov. 2022. Uzbek sentiment analysis based on local restaurant reviews. *Preprint*, arXiv:2205.15930.

Akylbek Maxutov, Ayan Myrzakhmet, and Pavel Braslavski. 2024. Do LLMs speak Kazakh? a pilot evaluation of seven models. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, pages 81–91, Bangkok, Thailand and Online. Association for Computational Linguistics.

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr., Bekhzodbek Moydinboyev, Esra Onal, Shaxnoza Pulatova, Ahsan Wahab, Orhan Firat, and Sriram Chellappan. 2021. A large-scale study of machine translation in Turkic languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5876–5890, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Medet Mukushev, Aigerim Kydyrbekova, Vadim Kimmelman, and Anara Sandygulova. 2022. Towards large vocabulary kazakh-russian sign language dataset: Krsl-onlineschool. *10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources, sign-lang 2022*.

Muhammadjon Musaev, Saida Mussakhojayeva, Ilyos Khujayorov, Yerbolat Khassanov, Mannon Ochilov, and Huseyin Atakan Varol. 2021. Usc: An open-source uzbek speech corpus and initial speech recognition experiments. *Preprint*, arXiv:2107.14419.

Saida Mussakhojayeva, Aigerim Janaliyeva, Almas Mirzakhmetov, Yerbolat Khassanov, and Huseyin Atakan Varol. 2021. Kazakhtts: An open-source kazakh text-to-speech synthesis dataset. *ArXiv*, abs/2104.08459.

Saida Mussakhojayeva, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022a. KazakhTTS2: Extending the open-source Kazakh TTS corpus with more data, speakers, and topics. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5404–5411, Marseille, France. European Language Resources Association.

Saida Mussakhojayeva, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022b. Ksc2: An industrial-scale open-source kazakh speech corpus. In *Interspeech*.

Amirgaliyev E. N., Kuanyshbay D. N., and Baimuratov O. 2020. Development of automatic speech recognition for kazakh language using transfer learning. *Preprint*, arXiv:2003.04710.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. *Preprint*, arXiv:1708.09803.

Dan Saattrup Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. *Preprint*, arXiv:2202.11684.

Aliya Nugumanova, Yerzhan Baiburin, and Yermek Alimzhanov. 2022. Sentiment analysis of reviews in kazakh with transfer learning techniques. *2022 International Conference on Smart Information Systems and Technologies (SIST)*, pages 1–6.

Daniyar Nurseitov, Kairat Bostanbekov, Daniyar Kurmankhojayev, Anel Alimova, Abdelrahman Abdallah, and Rassul Tolegenov. 2021. Handwritten kazakh and russian (hkr) database for text recognition. *Multimedia Tools Appl.*, 80(21–23):33075–33097.

Daniil Orel, Rustem Yeshpanov, and Huseyin Atakan Varol. 2023. Speech recognition for turkic languages using cross-lingual transfer learning from kazakh. *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 174–182.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling

language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Ilyos Rabbimov and S S Kobilov. 2020. Multi-class text classification of uzbek news articles using machine learning. *Journal of Physics: Conference Series*, 1546.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Ulugbek Salaev, Elmurod Kuriyozov, and Carlos Gómez-Rodríguez. 2022. Simreluz: Similarity and relatedness scores as a semantic evaluation dataset for uzbek language. *Preprint*, arXiv:2205.06072.

Maksud Sharipov, Elmurod Kuriyozov, Ollabergan Yuldashev, and Ogabek Sobirov. 2023. Uzbektagger: The rule-based pos tagger for uzbek language. *ArXiv*, abs/2301.12711.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Nazgul Toiganbayeva, Mahmoud Kasem, Galymzhan Abdimanap, Kairat Bostanbekov, Abdelrahman Abdallah, Anel Alimova, and Daniyar Nurseitov. 2022. Kohtd: Kazakh offline handwritten text dataset. *Signal Processing: Image Communication*, 108:116827.

Gulmira Tolegen, Alymzhan Toleu, Rustam Mussabayev, Bagashar Zhumazhanov, and Gulzat Ziyatbekova. 2023. Generative pre-trained transformer for kazakh text generation tasks. *2023 19th International Asian School-Seminar on Optimization Problems of Complex Systems (OPCS)*, pages 1–5.

Francis M. Tyers and Jonathan N. Washington. 2015. Towards a free/open-source universal-dependency treebank for kazakh. In *3rd International Conference on Turkic Languages Processing, (TurkLang 2015)*, pages 276–289.

Aidar Valeev, Ilshat Gibadullin, Albina Khusainova, and Adil Khan. 2019. Application of low-resource machine translation techniques to russian-tatar language pair. *Preprint*, arXiv:1910.00368.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanovic, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked-language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Mexico City, Mexico. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Hao Wu and Beiqiang Ma. 2023. Kazakh-chinese neural machine translation based on data augmentation. In *Conference on Computer Graphics, Artificial Intelligence, and Data Processing*.

Rustem Yeshpanov, Pavel Efimov, Leonid Boytsov, Ardak Shalkarbayuli, and Pavel Braslavski. 2024a. Kazqad: Kazakh open-domain question answering dataset. *Preprint*, arXiv:2404.04487.

Rustem Yeshpanov, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022. KazNERD: Kazakh named entity recognition dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 417–426, Marseille, France. European Language Resources Association.

Rustem Yeshpanov, Alina Polonskaya, and Huseyin Atakan Varol. 2024b. Kazparc: Kazakh parallel corpus for machine translation. *Preprint*, arXiv:2403.19399.

Rustem Yeshpanov and Huseyin Atakan Varol. 2024. Kazsandra: Kazakh sentiment analysis dataset of reviews and attitudes. *Preprint*, arXiv:2403.19335.

Zhandos Yessenbayev, Zhanibek Kozhirbayev, and Aibek Makazhanov. 2020. Kaznlp: A pipeline for automated processing of texts written in kazakh language. In *Speech and Computer*, pages 657–666, Cham. Springer International Publishing.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Çağrı Çöltekin, A. Seza Doğruöz, and Özlem Çetinoğlu. 2022. Resources for turkish natural language processing: A critical survey. *Language Resources and Evaluation*, 57(1):449–488.

A Search Strategy

A.1 General Approach

Firstly, we queried for papers using generic search terms and engines specified in A.3. Secondly, we focused on exploring popular CL and NLP venues, specific institutes' resources (e.g. ISSAI's website) as well as lists of references of works found in the first step of the process. Finally, we queried for pre-prints and unpublished works and datasets on arXiv, Github, Kaggle, and HuggingFace.

A.2 Venues and Repositories

The primary focus of our search was popular CL and NLP venues, including ACL, EACL, CoNLL, EMNLP, and WMT. We also explored non-ACL events and proceedings, including COLING and LREC. For speech-related technology and datasets we researched works published at IEEE venues, including Interspeech, ICASSP, and SLT.
Additionally, we browsed pages of the universities and institutes that have been known for their contributions to the field, including ISSAI, Leipzig University, and Saarland University.
For pre-prints, works in progress, models, and datasets, we explored Github, Kaggle, and HuggingFace as well as the "Computation and Language" section on arXiv.

A.3 Search Engines and Query Details

Search engines used include Google Scholar, Semantic Scholar, and ResearchGate. We used both broad and specific query terms to search for publications on both resources and technologies. In terms of broad queries, we used the format of "[language] [topic]", for example, "Kyrgyz NLP" or "Kazakh speech". For more technology-specific searches we adopted queries of the form of "[language] [technology name/data]", for example, "Kazakh NER" or "Uzbek speech corpus".
The search cutoff date is set to November 1, 2024.

B Datasets

Below we provide statistics on the datasets surveyed in the main body of the paper. Datasets marked with an asterisk ("*") have not been released and/or are not currently available on an open-source basis. For the datasets marked with a dash ("-") in the last column, exact data quantities have not been reported. The amount of data provided is cited according to the datasets' authors report, which might differ from the actual data available.

Language	Dataset	Type/Task	Data Available
Kazakh	KazEmoTTS	sentiment analysis	54.7K audio-text pairs 74H 8.7K unique sentences
	KazSAnDRA	sentiment analysis	180K
	KazakhTTS	text-to-speech	93H
	KazakhTTS2	text-to-speech	271.7H
	KSC2	speech corpus	1,200H 600K utterances
	Kazakh Speech Commands Dataset	speech commands	119 speakers >100K utterances
	KOHTD	hand-written dataset	3K exam papers 140K segmented images 922K symbols
	KazNERD	NER	25 entity classes 112K sentences 136K annotations
	KazQAD	QA	6K questions 12K passage level judgments
	Almaty Corpus (NCKL)	linguistic features	40M word tokens 650K words
	Kazakh Language Corpus*	linguistic features	135M words 400K words
	Kazakh KTB	universal dependencies	300 sentences
	kkwac*	corpora collection	139M words
	Leipzig Corpora	corpora collection	51.4K news 17M web 773K Wikipedia sentences
	Kazakh National Language Corpus	corpora collection	22K docs 23M words
	Common Voice	speech corpus	4H recorded 3H validated
Uzbek	Restaurant Reviews (Matlatipov et al., 2022)	sentiment analysis	4.5K positive 3.1K negative
	Application Reviews (Kuriyozov et al., 2019)	sentiment analysis	2.5K positive 1.8K negative
	Uzbek POS* (Sharipov et al., 2023)	POS	-
	Multi-label Text Classification (Kuriyozov et al., 2023)	classification	512K articles 120M words 15 classes
	Multi-label News Classification* (Rabbimov and Kobilov, 2020)	classification	13K articles
	Uzbek Speech Corpus	speech corpus	105H

Language	Dataset	Type/Task	Data Available
	Common Voice	speech corpus	265H recorded 100H validated
	UzWordNet	WordNet	28K synsets
	SimRelUz	linguistic features	1.4K word pairs
	Uzbek Electronic Corpus	corpora collection	-
	Leipzig Corpora	corpora collection	86K news 663K community 280K newscrawl 263K Wikipedia sentences
Kyrgyz	kloop	corpora collection	16.8K articles
	kkwyc*	corpora collection	19M words
	Manas-UdS	corpora collection	1.2M words
	Leipzig Corpora	corpora collection	251K community 123k newscrawl 1.5K news 3M web 334K Wikipedia sentences
	Kyrgyz MNIST	hand-written symbols	80K images
	UD-Kyrgyz-KTMU	universal dependencies	781 sentences 7.4K tokens
	Kyrgyz news dataset	classification	23K articles 20 classes
	Common Voice	speech corpus	48H recorded 39H validated
Turkmen	Common Voice	speech corpus	7H recorded 3H validated
	Leipzig Corpora	corpora collection	276K web sentences 62K Wikipedia sentences

Table 4: Overview of monolingual datasets and their availability per language.

Dataset	Languages	Type/Task	Data Available
KazParC	KK, RU, EN, TR	parallel corpora	3K exam papers 140K segmented images 922K symbols
Russian-Kazakh Handwritten Database	RU, KZ	handwritten symbols	63K sentences 95% RU, 5% KZ
KRSL*	KK, RU	sign language	890H of videos 325 annotated videos 39K gloss annotations
Uzbek-Kazakh Corpora	UZ, KK	parallel corpora	124K sentences
Large Turkic Language Corpora* (Baisa and Suchomel, 2012)	KK, KY, UZ, TK	web corpora collection morphological segmentation	1.4M KZ 590K KY 320K UZ 200K TR words
Belebele	KK, UZ, KY	reading comprehension	900 questions 488 passages
xSID	KK	syntactic data	-
CC100	KK, KY, UZ	web corpora collection	-
WikiAnn	KK, KY, UZ, TK	NER	-
OSCAR	KK, KY, UZ, TK	web corpora collection	677K KK 144K KY 15K UZ 4.5K TR docs
AM2iCO	KK	lexical alignment	-
ST-kk-ru	KK, RU	speech translation	317H
M2ASR*	KK, KY	speech recognition	-
MuMiN	KK	multimodal fact checking	-

Table 6: Overview of multilingual and/or parallel datasets and their availability per language.