BOLASHAQ ACADEMY est.1996

ISSN 2312 - 4784

ХАЛЫҚАРАЛЫҚ ҒЫЛЫМИ ЖУРНАЛ | МЕЖДУНАРОДНЫЙ НАУЧНЫЙ ЖУРНАЛ | INTERNATIONAL SCIENTIFIC JOURNAL

Қазіргі заманғы маңызды мәселелер

Актуальные проблемы современности

Actual Problems of the Present

№4 (50)

Қарағанды
Болашақ-Баспа
2025

ҚАЗІРГІ ЗАМАНҒЫ МАҢЫЗДЫ МӘСЕЛЕЛЕР

Халықаралық ғылыми журнал

АКТУАЛЬНЫЕ ПРОБЛЕМЫ СОВРЕМЕННОСТИ

Международный научный журнал

ACTUAL PROBLEMS OF PRESENT

The international scientific journal

№4 (50)

ҚАЗІРГІ ЗАМАНҒЫ МАҢЫЗДЫ М
Баспа» РББ, 2025. – 146 б.

ISSN 2312 – 4784

Қарағанды
Болашақ-Баспа
2025

№ 4(50). – 2025. – Қарағанды: «Болашақ -

«Қазіргі заманғы маңызды мәселелер» Халықаралық ғылыми журналы Қазақстан Республикасы Мәдениет және ақпарат Министрлігімен тіркелген (25.09.2015 ж. № 15583-Ж мерзімді баспасөз басылымын есепке қою туралы куәлік).

Басылымның мерзімділігі: тоқсанына 1 рет

Негізгі тақырыптық бағыттары: ғылымның әр түрлі салалары қамтылған. Журнал ғылыми мақалалар, зерттеу материалдарын, хабарламалар, рецензиялар және т. б. жариялайды.

Мақала қайта басылған жағдайда журналға сілтеме жасалу міндетті. Авторлар келтірілген фактілердің, дәйексөздердің, жеке атаулардың, соның ішінде географиялық атаулардың шынайылығына жауапты.

Қазақстан Республикасының аумағында 75319 индекси бойынша тіркелген.

Ресей Федерациясының бұқаралық коммуникациялар және мәдени мұраны қорғау саласындағы заңнаманың сақталуын қадағалау жөніндегі федералдық қызметі РФ аумағында «Қазіргі заманғы маңызды мәселелер» (Қазақстан Республикасы) халықаралық журналын таратуға рұқсат берілген. 2006 жылғы 6 шілдедегі № 78 РП шетелдік мерзімді баспасөз басылымдарының өнімдерін таратуға рұқсаттама РФ аумағында № 88044 индексі, "Пресса России" Біріккен каталогында № 000053 индексі бойынша тіркелген.

«Қазіргі заманғы маңызды мәселелер» Халықаралық ғылыми журналы «Ресейлік ғылыми дәйексөз индексі» Ұлттық ақпараттық-талдау жүйесіне (РИНЦ)

енгізілген. 18.02.2016 ж.№ 75-02 / 2016 шарт

АКТУАЛЬНЫЕ ПРОБЛЕМЫ СОВРЕМЕННОСТИ: Международный научный журнал. – № 4 (50). – 2025. – Караганда: РИО «Болашак-Баспа», 2025. – 146 б.

ISSN 2312 – 4784

Международный научный журнал «Актуальные проблемы современности» зарегистрирован Министерством культуры и информации Республики Казахстан (Свидетельство о постановке на учёт периодического печатного издания и№ 15583-Ж от 25.09.2015г.).

Периодичность издания: 1 раз в квартал

Основная тематическая направленность ППИ: разные направления науки. Журнал публикует научные статьи, материалы исследований, сообщения, рецензии и др.
При перепечатке ссылка на журнал обязательна. Авторы несут ответственность за достоверность приведенных фактов, цитат, имен собственных, в том числе географических названий.

Подписка на территории Республики Казахстан по индексу **75319**

Федеральная служба по надзору за соблюдением законодательства в сфере массовых коммуникаций и охране культурного наследия Российской Федерации разрешает распространение международного журнала «Актуальные проблемы современности» (Республика Казахстан) на территории РФ. Разрешение на распространение продукции зарубежных периодических печатных изданий РП № 78 от 6 июля 2006 г. Подписка на территории РФ по индексу 88044 в объединенном каталоге «Пресса России» № 000053

Международный научный журнал «Актуальные проблемы современности» включен в национальную информационно-аналитическую систему «Российский индекс научного цитирования» (РИНЦ) – Договор № 75-02/2016 от 18 февраля 2016 г.

ACTUAL PROBLEMS OF Present: International scientific journal. № 4 (50). – 2025. – Karagandy: EPD «Bolashaq-Baspa», 2025. – 146 p.

ISSN 2312 – 4784

The international scientific journal «Actual problems of present» was registered by the Ministry of Culture and Information of the Republic of Kazakhstan (Certificate of registration of periodicals and № 15583-Ж dated September 25, 2015).

Frequency of publication: quarterely

The main thematic focus : different branches of science. The journal publishes scientific articles, materials of the research, reports, reviews, etc.
When reprinting, a link to the journal is required. The authors are responsible for the accuracy of the facts, quotes, proper names, including geographical names.
Subscription on the territory of the Republic of Kazakhstan on the index 75319

The Federal Service for the Supervision of Compliance with the Law in the Field of Mass Communications and the Protection of the Cultural Heritage of the Russian Federation allows the distribution of the international journal «Actual problems of modernity» (Republic of Kazakhstan) on the territory of the Russian Federation. Permission to distribute products of foreign periodicals of the RF № 78 dated July 6, 2006. Subscription on the territory of the Russian Federation by the index 88044 in the joint catalog "Press of Russia" № 000053

The international scientific journal «Actual problems of present» включен в национальную информационно-аналитическую систему «Российский индекс научного цитирования» (РИНЦ) – Договор No. 75-02 / 2016 dated February 18, 2016

МАЗМҰНЫ

ОГЛАВЛЕНИЕ

CONTENTS

https://doi.org/10.64863/2312-4784/2025-3-50/7-18
Received: 09.12.2025
Accepted: 29.12.2025

Natural Language Processing and Speech Technologies for Central Asian Turkic Languages: A Review of Current Methods, Resources, and Challenges

Palidan Muhetaer[1*]

[1]College of Information Management, Xinjiang University of Finance & Economics, Xinjiang, 830012, China. Email: paridanm@aliyun.com; ORCID: https://orcid.org/0009-0009-7223-5291

Abstract

This article provides a comprehensive review of contemporary research in the field of natural language processing (NLP) and speech technologies for Central Asian Turkic languages, including Kazakh, Kyrgyz, and Uzbek. Although a number of theoretical and applied studies have been published in recent years, these languages continue to be classified as low-resource. This situation is primarily caused by the limited availability of annotated text corpora, insufficient speech data, the parallel use of Cyrillic and Latin scripts, and the absence of unified annotation and evaluation standards. The article systematically examines current approaches to morphological segmentation, named entity recognition, sentiment analysis, and automatic speech recognition. Agglutinative morphology and vowel harmony are discussed as key typological features of Turkic languages that strongly influence computational processing strategies. The effectiveness of both rule-based and neural morphological analyzers is highlighted. The paper also describes the adaptation of computational models originally developed for Turkish, English, and Russian through subword modeling, character-level embeddings, and multilingual transformer architectures. In addition, cross-lingual transfer learning is evaluated as a promising approach to mitigating data scarcity. The study identifies corpus fragmentation, inconsistent annotation schemes, and the lack of standardized speech resources as major challenges. The author argues for the development of open-access datasets, the introduction of shared evaluation tasks, and the strengthening of institutional collaboration between linguists and computational language technology specialists. The findings of the study are of both theoretical and practical importance for the development of sustainable and effective language technologies for low-resource languages.

Keywords: Turkic NLP, Morphological analysis, Low-resource languages, Speech recognition, Cross-lingual transfer.

Орталық Азия түркі тілдеріне арналған табиғи тілдерді өңдеу және сөйлеу технологиялары: қазіргі әдістер, ресурстар және мәселелерге шолу

Палидан Мухетаер[1]*

[1]Ақпаратты Басқару колледжі, Шыңжаң Қаржы-Экономика Университеті, Шыңжаң, 830012, Қытай. E-mail: paridanm@aliyun.com; ORCID: https://orcid.org/0009-0009-7223-5291

Аннотация

Бұл мақала Орталық Азия түркі тілдеріне, соның ішінде қазақ, қырғыз және өзбек тілдеріне арналған табиғи тілдерді өңдеу (NLP) мен сөйлеу технологиялары саласындағы заманауи ғылыми зерттеулерге кешенді шолу ұсынады. Соңғы жылдары бұл бағытта бірқатар теориялық және қолданбалы еңбектер жарияланғанымен, аталған тілдер әлі де төмен ресурсты тілдер санатына жатады. Бұған аннотацияланған мәтіндік корпустардың жеткіліксіздігі, сөйлеу деректерінің шектеулілігі, кирилл және латын графикаларының қатар қолданылуы,

сондай-ақ бірыңғай аннотациялау мен бағалау стандарттарының болмауы негізгі себеп болып отыр. Мақалада морфологиялық сегментация, атаулы мәндерді тану, сентимент-талдау және автоматты сөйлеуді тану салаларындағы қазіргі әдістер жүйеленіп талданады. Сөз құрамының агглютинативті болуы мен дыбыс үндестігі түркі тілдерінің басты типологиялық белгілері ретінде қарастырылып, оларды өңдеуде ережеге негізделген және нейрондық морфологиялық талдағыштардың тиімділігі айқындалады. Түрік, ағылшын және орыс тілдеріне арналған есептеу модельдерін субсөздік тәсілдер, таңба-деңгейлі эмбеддингтер және көптілді трансформер архитектуралары арқылы бейімдеу тәжірибесі сипатталады. Сонымен қатар, тілдер арасындағы трансферлік оқыту деректер тапшылығын азайтудың перспективалы бағыты ретінде бағаланады. Зерттеуде корпустардың фрагменттелуі, аннотация үлгілерінің бірізді болмауы және стандартталған сөйлеу базаларының жоқтығы негізгі мәселе ретінде көрсетіледі. Автор ашық қолжетімді деректер қорын дамыту, ортақ бағалау тапсырмаларын енгізу және лингвистер мен есептеу лингвистикасы мамандары арасындағы институционалдық ынтымақтастықты күшейту қажеттігін негіздейді. Зерттеу нәтижелері төмен ресурсты тілдерге арналған тұрақты және тиімді тілдік технологияларды қалыптастыруда теориялық әрі практикалық маңызға ие.

Кілт сөздер: Түркі тілдеріне арналған NLP, морфологиялық талдау, төмен ресурсты тілдер, сөйлеуді тану, тілдер арасындағы трансфер.

Обработка естественного языка и технологии речи для тюркских языков Центральной Азии: обзор современных методов, ресурсов и проблем

*Палидан Мухетаэр[1]**

[1]Колледж информационного менеджмента, Синьцзянский университет финансов и экономики, Синьцзян, 830012, Китай. E-mail: paridanm@aliyun.com; ORCID: https://orcid.org/0009-0009-7223-5291

Аннотация

В статье представлен комплексный обзор современных научных исследований в области обработки естественного языка (NLP) и речевых технологий для тюркских языков Центральной Азии, включая казахский, кыргызский и узбекский. Несмотря на рост теоретических и прикладных работ в последние годы, данные языки по-прежнему относятся к категории низкоресурсных. Основными причинами этого являются недостаточность аннотированных текстовых корпусов, ограниченность речевых данных, параллельное использование кириллической и латинской графики, а также отсутствие единых стандартов аннотирования и оценки. В статье систематически анализируются современные методы морфологической сегментации, распознавания именованных сущностей, сентимент-анализа и автоматического распознавания речи. Агглютинативная морфология и гармония гласных рассматриваются как ключевые типологические характеристики тюркских языков, определяющие специфику их вычислительной обработки. Показана эффективность как правил-ориентированных, так и нейронных морфологических анализаторов. Описывается опыт адаптации вычислительных моделей, разработанных для турецкого, английского и русского языков, с использованием субсловных подходов, символьных эмбеддингов и многоязычных трансформерных архитектур. Особое внимание уделяется трансферному обучению как перспективному способу снижения дефицита данных. В качестве ключевых проблем выделяются фрагментированность корпусов, несогласованность схем аннотирования и отсутствие стандартизированных речевых баз. Автор обосновывает необходимость развития открытых корпусов, внедрения общих оценочных задач и усиления институционального сотрудничества между лингвистами и специалистами по вычислительной лингвистике. Полученные выводы имеют как теоретическую, так и практическую значимость для

формирования устойчивых языковых технологий для низкоресурсных языков.

Ключевые слова: NLP для тюркских языков, морфологический анализ, низкоресурсные языки, распознавание речи, межъязыковой трансфер.

1. Background and linguistic characteristics of Central Asian languages

Central Asia represents one of the most linguistically dynamic regions of Eurasia, where Turkic, Iranian, and Slavic languages have interacted over many centuries. The most widely spoken state languages today include Kazakh, Kyrgyz, Uzbek, Turkmen, and Tajik, while Russian continues to play an important role in administration, media, and education. Researchers stress that modern multilingualism in the region cannot be understood without reference to both the Soviet legacy and much older historical contact zones (Smagulova, 2016; Smagulova & Ahn, 2016).

From a typological perspective, Kazakh, Kyrgyz, Uzbek, and Turkmen belong to the Turkic language family and display the classic features of agglutinative morphology, vowel harmony, and SOV word order (Johanson & Csató, 1998). These languages encode grammatical categories – including case, number, person, tense, aspect, mood, and evidentiality – through long and productive suffix chains. A single root, especially verbal, may generate dozens of surface forms. For example, Kazakh verbal morphology allows multiple layers of derivation before inflection, creating complex tokens that standard NLP tokenizers often cannot process (Kessikbayeva & Çiçekli, 2014).

Vowel harmony represents one of the most stable phonological mechanisms in Central Asian Turkic languages. In Kazakh and Kyrgyz, backness and rounding typically determine which vowel will occur in a suffix. This system interacts with loanword phonology, because borrowed stems – often from Russian or Persian – may violate harmonic expectations (Dave, 1996a; Dave, 1996b). Researchers note that speakers frequently remodel borrowed items to fit harmonic patterns, a process that complicates orthographic normalization and corpus annotation (Fierman, 2006; Wei, 2015).

The lexicon of Central Asian languages shows multiple historical strata. Medieval borrowing introduced extensive Arabic and Persian vocabulary, particularly relating to religion, administration, and literary culture. Later, during the nineteenth and twentieth centuries, Russian became a major source of technical and bureaucratic terminology. Modern digital communication adds a new layer of English-based internationalisms. A sociolinguistic survey by Bahry (2012) observes that code-switching between Kazakh and Russian is routine among young urban speakers, particularly online, where stylistic choices often mark identity and stance.

One of the most important contemporary issues is script policy. Uzbek adopted a Latin script in the 1990s, though Cyrillic remains common; Tajik uses Cyrillic; Kazakh officially plans transition from Cyrillic to Latin. Script reform affects literacy practices, school curricula, and NLP design, because parallel corpora must handle multiple spelling systems for the same language (Smagulova & Ahn, 2016). In practical NLP terms, this requires dedicated orthographic normalization modules for preprocessing – especially for web and social media texts that mix Cyrillic, Latin, and ad hoc transliterations.

Researchers also highlight the lack of standardized corpora for typological and computational work. While several national corpus projects exist, their accessibility and annotation depth vary widely. For Kazakh, preliminary work on morphological modeling and part-of-speech tagging has been published, but large balanced corpora comparable to Russian National Corpus or COCA remain under development (Mustajoki et al., 2020). Kyrgyz and Turkmen have even more limited digital resources. Uzbek has relatively more NLP attention due to diaspora and machine translation initiatives, but resources remain fragmented (Moore, 22023).

For computational linguistics, these features create low-resource conditions. Agglutinative morphology yields extremely large vocabularies and high out-of-vocabulary rates; script variation and code-switching increase noise; and data scarcity limits supervised models. Consequently, researchers working on Kazakh, Kyrgyz, and Uzbek increasingly adopt:

subword modeling (e.g., byte-pair encoding),

morphological segmentation, and

transfer learning from typologically related languages such as Turkish.

A similar perspective is found in Central Asian sociolinguistic research: successful digital processing must integrate linguistic knowledge about morphology, phonology, and language contact (Smagulova & Ahn, 2016; Bahry, 2012).

It is also relevant that Turkic languages in Central Asia form a configurational typological group, with relatively predictable morphological systems (Khassenov & Bakhitova, 2025). This is a major opportunity for NLP: once accurate morphological analyzers are built for Kazakh, they can be adapted to Kyrgyz or Karakalpak with relatively minor adjustments, because affix inventories, harmonic patterns, and syntactic templates overlap (Johanson & Csató, 1998). This is why recent work in computational typology encourages cross-Turkic resource sharing, particularly for tagging and parsing applications (Pavlenko, 2008).

In summary, Central Asian Turkic languages present a rich but challenging environment for computational research. Their agglutinative morphology requires subword-based models; their lexical history demands robust normalization; their script variation complicates digital corpora; and their sociolinguistic conditions produce code-switching and mixed registers. Understanding these background characteristics is essential before discussing sentiment analysis, named entity recognition, and speech technology for these low-resource languages.

2. Challenges for NLP and Sentiment-oriented Text Processing in Central Asian Turkic Languages

Natural language processing (NLP) for Turkic languages of Central Asia (such as Kazakh, Kyrgyz, Uzbek) faces a number of typological and resource-driven challenges. First and foremost, these languages are agglutinative – words are formed by adding multiple suffixes to a root, often resulting in long morphological chains that encode case, number, negation, possession, aspect, mood, etc. (Dzhubanov & Khasanov, 1973; Kessikbayeva & Çiçekli, 2014).

Because of this heavy inflectional morphology, a naive word-level tokenization or vocabulary-based embedding approach is often ineffective: the number of distinct surface forms explodes, leading to very high out-of-vocabulary (OOV) rates, data sparsity, and poor generalization for downstream tasks such as sentiment classification, entity recognition, or part-of-speech tagging.

Moreover, vowel harmony is a core phonological feature shared among many Turkic languages (Kazakh, Uzbek, Kyrgyz, etc.) – suffix vowels must agree in backness (and often rounding) with the last vowel of the root (Hakkila, 2018; Kulgildinova et al., 2018; Landau, J. M., & Kellner-Heinkele, 2001). This phonological constraint affects morphological variation and complicates subword tokenization or segmentation strategies if they ignore such harmonies.

Another complicating factor is orthographic and script variation, particularly in languages undergoing script reforms or using multiple scripts. For example, some Turkic languages use Cyrillic, Latin, or even Arabic-based scripts depending on historical periods or contexts. This introduces inconsistencies in text corpora, complicates normalization, and reduces the effectiveness of standard NLP pre-processing pipelines. As a result, corpora must often be normalized carefully before tokenization and embedding.

On the resource side – many Central Asian languages remain under-resourced. There is a lack of large, balanced, annotated corpora comparable to those available for high-resource languages like English, Russian, or Chinese. For instance, efforts such as the “Computational Description of the Kazakh Language” remain among the earliest computational attempts (Dzhubanov & Khasanov, 1973). More recent modern morphological analyzers and corpus projects (e.g., Kessikbayeva & Çiçekli, 2014) provide helpful building blocks, but data for tasks such as sentiment classification, semantic analysis, or large-scale text mining remain scarce.

These combined factors – agglutination, vowel harmony, script variation, data sparsity – make conventional NLP pipelines (word-level embeddings, off-the-shelf tokenizers, static lexicons) poorly suited for Turkic Central Asian languages. As a result, most contemporary research toward NLP for these languages recommends or implements subword-level models, morphological analyzers, or

hybrid tokenization strategies that can better account for morphological structure and reduce sparsity (see Kessikbayeva & Çiçekli, 2014; related discussion in computational linguistics for Turkic languages).

Given these foundational difficulties, it is unsurprising that fully-developed sentiment analysis systems for Central Asian Turkic languages remain rare and uneven in quality. Even more, most studies focus on lower-level tasks (morphology, lemmatization, POS-tagging) rather than semantic or pragmatic tasks (sentiment, pragmatics, discourse).

In sum, before applying sentiment analysis or other deep-semantic NLP on Central Asian Turkic languages, it is necessary to adapt preprocessing: use morphological-aware segmentation; support subword or character-level models; normalize orthography across scripts; and ideally, build annotated corpora representative of social-media, news, and informal text, to capture real usage including code-switching, loanwords, and morphological variation.

3. Named Entity Recognition and Sequence Tagging for Kazakh and Other Central Asian Turkic Languages

Named entity recognition (NER) is one of the most important sequence tagging tasks for building practical NLP systems in low-resource Turkic languages. A core difficulty arises from agglutinative morphology: entity tokens in Kazakh, Uzbek, and Kyrgyz carry rich case and derivational suffixes, so a single named entity may appear in tens of different orthographic forms (Kessikbayeva & Çiçekli, 2014; Makhambetov et al., 2015). To process such structures, most research emphasizes character-level or subword-level models rather than standard word-based taggers.

Early work on Kazakh NLP focused on rule-based and finite-state approaches to morphological segmentation, which subsequently enabled more reliable POS tagging and tokenization (Yiner & Kurt, 2021). While these developments were not originally designed for NER, they provided a foundational layer: accurate segmentation of complex tokens greatly improves downstream entity recognition.

More recent studies demonstrate that neural sequence models are effective for NER in data-sparse environments. Researchers working on Kazakh corpora have begun to implement RNN-based and transformer-based architectures, which learn entity boundaries directly from annotated texts. For example, Akhmed-Zaki et al. (2021) report improved entity identification in Kazakh social-media and news text using a neural morphological component as preprocessing. Their system uses character embeddings to handle suffix chains and reduce sparsity in training data.

At the architectural level, NER systems for Turkic languages commonly combine three layers:

1. Character/subword embedding
2. Sequential encoding (typically BiLSTM or BiGRU)
3. Sequence decoding (often CRF)

This combination is motivated by the need to capture both local morphological cues (character patterns, root + suffix transitions) and global contextual dependencies (sentence-level context) (Figure 1).

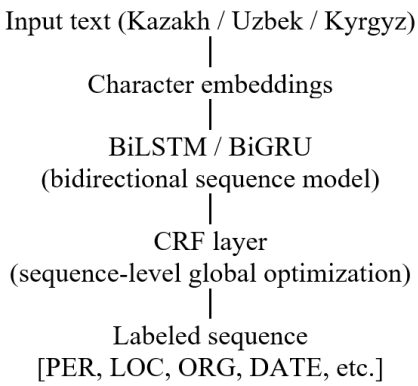

Figure 1. Typical neural hybrid model for sequence tagging in agglutinative Turkic languages

One immediate problem in sequence tagging for Central Asian languages is annotation scarcity. Whereas English NER benefits from CoNLL-style standard datasets with tens of thousands of labeled examples, Kazakh and Uzbek currently lack comparable corpora. This deficiency forces researchers to build smaller, domain-specific datasets for news or government text, rather than large balanced corpora (Makhambetov et al., 2013). A related issue is inconsistency in entity taxonomies: different projects introduce different label inventories (e.g., some include only PER–LOC–ORG, others add EVENT, PRODUCT, or DEVELOPMENT PROGRAM), which makes cross-study comparison difficult.

In addition, orthographic variation presents a challenge for NER. Kazakh and Uzbek data may contain Cyrillic, Latin, or non-standard transliterations. This not only increases token variation but also interferes with model training, because one entity may appear in multiple spelling forms across different sources. Script normalization is therefore a critical preprocessing step for NER (Smagulova & Ahn, 2016). Preprocessing pipelines often include rule-based conversion modules (Cyrillic → Latin, or automated transliteration) to unify tokens before feeding them to neural models.

A promising long-term strategy for NER in Central Asian languages involves cross-lingual transfer. Since Kazakh, Kyrgyz, Uzbek, and Tatar share similar morphological structure, it is feasible to train models on a mix of Turkic corpora, or to transfer models trained on Turkish to Kazakh or Uzbek (Sharipbay et al., 2018). This idea is especially valuable when labeled data are scarce: character-level morphological cues generalize across languages even when lexical content differs (Johanson & Csató, 1998). Multilingual transformer models such as XLM-R and multilingual BERT also appear promising because they embed multiple languages in a shared space, where typologically related languages tend to cluster.

The logical next step for NER research in Central Asia is large-scale corpus consolidation. Many papers emphasize that without standardized annotation guidelines, reproducibility and quantitative comparison between models will remain limited (Makhambetov et al., 2015). Developing a Central Asian NER benchmark – analogous to CoNLL 2003 for English, or GermEval for German – would accelerate progress considerably. This would include:

shared entity types,
unified document segmentation standards,
transparent training/test splits,
and publicly available evaluation metrics.

In sum, the present state of NER research in Central Asian Turkic languages is characterized by robust foundational work on morphological analysis and a transition toward neural sequence tagging models. Key research priorities include unified annotation, cross-script normalization, and multilingual transfer between related Turkic languages.

4. Speech Technologies for Kazakh and Other Central Asian Turkic Languages

Speech technologies for Central Asian languages have historically lagged behind textual NLP, primarily due to the scarcity of annotated audio data, the complexity of phonological systems, and limited funding for long-term corpus development (Smagulova & Ahn, 2016). While text corpora for Kazakh have grown steadily since the 2010s (Makhambetov et al., 2013), speech corpora remain fragmented, domain-specific, and in many cases not publicly available.

The earliest computational work on Kazakh speech used hidden Markov models (HMMs) with mel frequency cepstral coefficient (MFCC) feature extraction – a standard acoustic modeling technique inherited from English and Russian ASR research. However, the agglutinative structure and vowel harmony of Kazakh complicate the construction of canonical pronunciation dictionaries. Kazakh phonology distinguishes front/back and rounded/unrounded vowel pairs, and these features must be encoded consistently for automatic speech recognition (ASR) systems (Johanson & Csató, 1998). Researchers emphasize that orthographic forms alone are insufficient; grapheme-to-phoneme conversion must incorporate rules that handle vowel harmony and postlexical assimilation (Yiner & Kurt, 2021).

By the mid-2010s, deep learning approaches began entering Central Asian ASR research. Neural acoustic models offer multiple advantages in low-resource settings: they learn spectral-temporal features automatically, reduce reliance on expert-crafted phonological rules, and can adapt to noisy recording conditions typical of field corpora. In an influential study, Akhmed-Zaki et al. (2021) highlighted the importance of speech preprocessing pipelines for Kazakh, including noise filtering, segmentation, and automatic annotation. Their system integrated speech processing with morphological analysis, showing that improvements in text preprocessing benefited downstream speech recognition tasks.

The recent direction in Kazakh speech technology involves hybrid neural architectures, especially convolutional neural networks (CNNs) combined with recurrent or transformer blocks. CNNs extract local spectral patterns (formant transitions, harmonic structures), while RNN or transformer layers capture temporal dependencies across longer speech segments. The posterior probabilities are typically decoded using connectionist temporal classification (CTC) or attention mechanisms, enabling end-to-end ASR without explicit phoneme alignment (Akhmed-Zaki et al., 2021) (Figure 2).

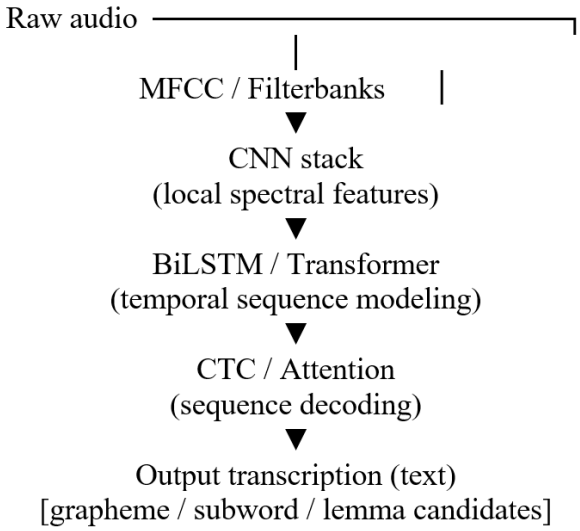

Figure 2. Generic hybrid neural ASR pipeline for Kazakh speech (ASCII)

Speech synthesis (TTS) also presents challenges for Central Asian languages. In Turkish, which is typologically similar, unit-selection and parametric synthesis dominated early work; however, modern approaches use neural vocoders and transformer-based prosody modeling. For Kazakh, recent work shows that morphological preprocessing can improve prosodic placement in synthesized speech by predicting suffix boundaries and reducing monotonic stress contours (Mansurova et al., 2024). Agglutinative morphology creates unusually long word forms, which can distort rhythm and stress patterns, especially for naive neural TTS systems trained on small datasets. Neural models trained at subword level (character, BPE, or phoneme sequences) handle this issue much better.

One promising strategy for low-resource speech technologies is cross-language transfer. Turkic languages share phonological and morphological traits, especially in vowel harmony, syllable structure, and sonority sequencing. Researchers therefore propose transferring pretrained acoustic models from Turkish to Kazakh, Uzbek, or Kyrgyz (Sharipbay et al., 2018). Initial results indicate that transferring CNN encoder weights leads to faster convergence and improved recognition accuracy, even with limited target-language data.

Another strategic development is the creation of spoken language corpora specifically designed for Central Asian languages. For example, Nazarbayev University is developing the Multimedia Corpus of Modern Spoken Kazakh Language, which collects real conversational speech rather than scripted or read material. This is crucial because Turkic discourse features extensive vowel reduction, assimilation, and optional case marking, all of which differ systematically from literary norms

(Multimedia Corpus Project, 2024). Training speech models solely on read speech produces high accuracy in laboratory settings but poor generalization for natural conversations.

Looking ahead, the future of speech technology for Central Asian languages will likely involve:
multilingual ASR models (Kazakh–Russian, Uzbek–Russian, Kazakh–English),
speech-to-text translation, especially for public services and media analytics,
neural TTS integrated with morphological analyzers, and
large-scale open benchmark corpora modeled on LibriSpeech or Common Voice.

The primary bottlenecks today remain data quantity and annotation quality. Building speech corpora is expensive and time-consuming, requiring transcription, speaker metadata, dialect tagging, and quality control. Without shared standards, each research team produces incompatible datasets, making it difficult to compare systems fairly. NLP researchers repeatedly emphasize that standardization – common annotation guidelines, unified tag sets, and transparent licensing – is a critical next step (Makhambetov et al., 2015).

In summary, speech technology research for Central Asian Turkic languages is transitioning from traditional HMM/GMM pipelines to end-to-end neural architectures with morphological-aware preprocessing. Progress depends on new corpora, cross-lingual transfer from related Turkic languages, and integration of speech modules with existing NLP infrastructure for text processing.

5. Challenges and Future Directions for NLP and Speech Technologies in Central Asian Turkic Languages

Despite visible progress over the last decade, NLP and speech technologies for Central Asian Turkic languages still face a set of systematic challenges. These challenges are not merely technical; they are tied to deeper linguistic, sociolinguistic, and infrastructural realities within the region. Research repeatedly indicates that data scarcity, orthographic variation, low institutional support, and lack of standardization remain the biggest obstacles to developing robust language technologies for Kazakh, Kyrgyz, Uzbek, and other regional languages (Smagulova & Ahn, 2016; Makhambetov et al., 2013).

The single most frequently reported problem is insufficient annotated corpora. For English, sentiment analysis and NER systems benefit from decades of investment in corpus creation and maintenance, including standardized datasets like CoNLL, SemEval, OntoNotes, and LibriSpeech. By contrast, Central Asian languages currently rely on small, manually collected corpora, often developed as part of individual research projects (Makhambetov et al., 2015). These corpora may not be publicly available, may use incompatible tagging schemes, or may lack metadata (speaker information, domain labels, dialect annotations, etc.).

Researchers emphasize that corpus development must become collaborative and institutional, rather than project-based. The Multimedia Corpus of Modern Spoken Kazakh Project (2024) represents a promising step toward this direction.

A persistent complication in preprocessing for Kazakh and Uzbek is parallel orthography. Kazakh corpora include Cyrillic, Latin transliterations, and, in some cases, older Arabic script in historical archives. Uzbek presents even more variation due to the coexistence of Cyrillic, Latin (post-1991 reforms), and informal Latinized social media orthography. This yields token explosion: the same named entity may appear in multiple orthographic variants, which harms both symbolic and neural models (Smagulova & Ahn, 2016).

The long-term solution is not to eliminate variation, but to build robust normalization modules that can transform text to a canonical pre-processing form before tokenization. Projects like Yiner & Kurt (2021) demonstrate that morphological analyzers are crucial in this pipeline, because morphological structure interacts directly with orthographic surface forms.

Another challenge concerns inconsistency in annotation schemes. For example, some Kazakh named entity tagsets include only basic PER–LOC–ORG categories, while others include EVENT, PRODUCT, TITLE, or governmental classifications. This makes it difficult to compare system performance across studies. Scholars therefore call for a shared benchmark for Central Asian

languages, analogous to CoNLL-2003 for English (Makhambetov et al., 2015).

Establishing joint evaluation protocols, shared training-test splits, and transparent error analysis reports would significantly accelerate algorithmic progress. Without this, improvements remain local, anecdotal, and not reproducible.

While textual corpora for Kazakh have improved, spoken corpora remain extremely underdeveloped, especially for informal, spontaneous speech. This is problematic because agglutinative languages present different linguistic behavior in spoken discourse: vowel reduction, suffix clipping, truncation, and code-switching are much more frequent (Johanson & Csató, 1998). Speech recognition systems trained only on read speech overfit to literary norms and fail on natural conversation.

The current direction in speech technology suggests investment in:

spontaneous conversational corpora,

dialect-indexed samples,

metadata on speaker age, region, and gender,

and domain-specific speech such as customer service or medical contexts.

Another structural challenge is research continuity. Central Asian NLP research is often carried out by small teams who depend on short-term grants or PhD projects. Once funding ends, corpora and codebases become inaccessible or unsupported. Large languages (English, Chinese, Russian) benefit from sustained infrastructure – research labs, national linguistic institutes, and open community initiatives.

A long-term solution may lie in consortia and resource visibility, similar to CLARIN-ERIC in Europe or OpenSLR for speech resources. Labelled and version-controlled datasets for Kazakh, Kyrgyz, and Uzbek need stable institutional hosting, open licensing, and academic governance (Figure 3)

Based on existing studies (Kessikbayeva & Çiçekli, 2014; Akhmed-Zaki et al., 2021; Sharipbay et al., 2018), several future research priorities are becoming clear:

1. Morphology-aware neural models – neural sequence tagging and classification systems should integrate morphological information explicitly (character embeddings, subword models, stem–suffix decomposition).

2. Cross-Turkic transfer – Kazakh, Kyrgyz, Uzbek, and Turkish share structural patterns. Training multilingual embeddings or fine-tuning mBERT / XLM-R models on pooled Turkic corpora can improve performance in low-data settings.

3. Script normalization pipelines – robust Cyrillic↔Latin transliteration modules combined with morphological analyzers improve preprocessing and reduce vocabulary sparsity.

4. Open corpora and licensing – datasets must be published under open licenses (CC-BY, CC-BY-SA) to enable cross-institutional reuse.

5. Benchmarking – community-defined shared tasks (NER, POS, sentiment, ASR) for Central Asian languages should be launched annually, following CoNLL-style evaluation formats.

6. Speech-text integration – multimodal systems combining ASR, morphological tagging, and NER could support real digital tools: search in audio archives, video subtitles, court transcriptions, etc.

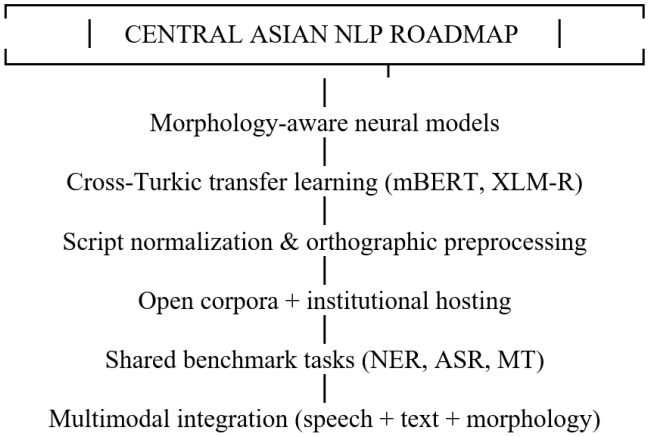

Figure 3. Future research roadmap for Central Asian NLP (ASCII)

6. Conclusion

Research on natural language processing and speech technologies for Central Asian Turkic languages has made measurable progress over the last decade, although the region remains significantly under-resourced compared to high-resource language contexts in Europe, North America, East Asia, or even the Middle East. The overview of current scholarship across morphology, named entity recognition, sentiment analysis, and speech technology suggests several consistent patterns.

First, morphology remains central. Because Kazakh, Kyrgyz, Uzbek, and related Turkic languages employ agglutinative word formation with productive suffix chains, NLP approaches must prioritize morphology-aware architectures. Early rule-based analyzers (Kessikbayeva & Ilyas, 2016) and data-driven parsing systems (Makhambetov et al., 2015) established solid foundations for handling derivational and inflectional complexity. Recent neural systems increasingly integrate subword tokenization, character embeddings, and morphological segmentation directly into transformer-based pipelines (Akhmed-Zaki et al., 2021; Yiner & Kurt, 2021). In practical terms, any future work on semantic tasks such as sentiment analysis will depend on reliability at the level of morphological preprocessing.

Second, resource construction and standardization emerge as the most urgent structural needs. Existing corpora tend to be small, domain-specific, and inconsistent in annotation standards (Makhambetov et al., 2013). Without unified benchmarks, it is difficult to compare algorithmic performance across research groups, replicate results, or build cumulative progress. Models perform well in isolated experiments but remain difficult to generalize. The development of open corpora such as the Multimedia Corpus of Modern Spoken Kazakh (2024) illustrates how collaborative infrastructure projects can address this issue, yet comparable corpora do not exist for Kyrgyz or Turkmen. Cross-institutional collaboration–modeled on European CLARIN or the OpenSLR community–would significantly accelerate progress if adopted in Central Asia.

Third, multilingual and cross-Turkic learning approaches show promise. Because Kazakh, Kyrgyz, Uzbek, and Tatar share structural features at both the phonological and morphological level (Johanson & Csató, 1998), transfer learning can compensate for data scarcity. Shared multilingual language models (e.g., multilingual BERT, XLM-R) make it possible to train systems jointly across multiple Turkic corpora, even when individual corpora remain small. This direction is especially relevant for named entity recognition, sentiment classification, and subword language modeling.

Fourth, speech technology demands focused attention. While text-oriented NLP has slowly expanded, speech technology–including ASR, TTS, and multimodal dialog systems–remains underdeveloped. Speech corpora are particularly scarce, and those that exist are often limited to read speech rather than spontaneous, dialect-rich conversation. Hybrid neural architectures in ASR (Akhmed-Zaki et al., 2021) indicate promising results, but without large-scale data collection, sustained progress will be slow. Given the sociolinguistic realities of multilingualism and code-

switching (Smagulova & Ahn, 2016), speech technology must incorporate language identification, script normalization, and morphology-aware modeling.

Finally, the next stage of research is not primarily technical but structural. Many of the technological tools exist: subword tokenization, multilingual embeddings, transformer architectures, neural vocoders, and end-to-end ASR. What Central Asian languages now require are: (1) open and permanent resource hosting, (2) transparent licensing, and (3) shared evaluation tasks based on realistic benchmarks. Progress in Central Asian NLP will depend increasingly on community design rather than purely individual innovation. Sustainable development requires cooperation between universities, research institutes, governmental language programs, and international computational linguistics organizations.

In summary, while Central Asian Turkic NLP and speech research remains in an early stage, the last decade has produced meaningful infrastructure: morphological analyzers, emerging corpora, neural preprocessing tools, and typologically informed modeling strategies. Continued growth will require systematic corpus development, multilingual modeling, script normalization pipelines, and the establishment of shared benchmarks. The foundations are now present; the priority is scaling, standardizing, and sustaining them.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The author declares that there are no financial or personal conflicts of interest concerning the content of this article.

Author Contributions

The author conceived and designed the study, collected and analyzed the data, drafted the manuscript, and approved the final version for publication.

Data Availability

All data supporting the findings of this study are included within the manuscript text. Additional materials are available from the author upon reasonable request.

References

1. Akhmed-Zaki, D., et al. (2021). Development of an information system for the Kazakh language: Text preprocessing tools for media corpus. *Applied Computing and Informatics*, 17(2), 320–334. DOI: https://doi.org/10.1080/23311916.2021.1896418

2. Bahry, S. A. (2012). What constitutes quality in minority education? A multiple embedded case study of stakeholder perspectives on minority linguistic and cultural content in school-based curriculum in Sunan Yughur Autonomous County, Gansu. *Frontiers of Education in China*, *7*(3), 376-416.

3. Dave, B. (1996a). *Politics of language revival: National identity and state building in Kazakhstan*. Syracuse University.

4. Dave, B. (1996b). National revival in Kazakhstan: Language shift and identity change. *Post-Soviet Affairs*, *12*(1), 51-72.

5. Dzhubanov, A., & Khasanov, B. (1980). Computational description of the Kazakh language. In *Computational and mathematical linguistics: Proceedings of the International Conference on Computational Linguistics: Pisa, 27/viii-1/ix 1973: vol. II.-(Biblioteca dell'Archivum Romanicum. Serie II: Linguistica; 37)* (pp. 75-77). LS Olschki.

6. Fierman, W. (2006). Language and education in post-Soviet Kazakhstan: Kazakh-medium instruction in urban schools. *The Russian Review*, *65*(1), 98-116.

7. Hakkila, A. C. (2018). *Transitional Literature of the Steppe: The First Two Qazaq Novels (Dulatuly and Köbeev)*. The University of Wisconsin-Madison.

8. Johanson, L., & Csató, É. Á. (1998). The Turkic languages. Routledge.

9. Kessikbayeva, G., & Cicekli, I. (2014, June). Rule based morphological analyzer of Kazakh language. In *Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM* (pp. 46-54).

10. Kessikbayeva, G. C. Ilyas. 2016.*A RuleBased Morphological Analyzer and a Morphological Disambiguator for Kazakh Language.» Linguistics and Literature Studies*, 96-104. DOI: https://doi.org/10.13189/lls.2016.040111

11. Khassenov, B., Bakhitova, Zh. (2025). Semantic Identity of Vowels in the Kazakh Language: an Experimental Analysis of Sound Symbolism. *Actual Problems of the Present,* 3(49), 19-30. DOI: https://doi.org/10.64863/2312-4784/2025-3-49/19-30

12. Kulgildinova, T., Zhumabekova, A., Shabdenova, K., Kuleimenova, L., & Yelubayeva, P. (2018). Bilingualism: language policy in modern Kazakhstan. *XLinguae*, *11*(1), 332-341.

13. Landau, J. M., & Kellner-Heinkele, B. (2001). *Politics of language in the ex-Soviet muslim states: Azerbayjan, Uzbekistan, Kazakhstan. Kyrgyzstan, Turkmenistan, and Tajikistan*. University of Michigan Press.

14. Makhambetov, O., Makazhanov, A., Yessenbayev, Z., Matkarimov, B., Sabyrgaliyev, I., & Sharafudinov, A. (2013, October). Assembling the kazakh language corpus. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1022-1031).

15. Makhambetov, O., Makazhanov, A., Sabyrgaliyev, I., & Yessenbayev, Z. (2015, April). Data-driven morphological analysis and disambiguation for kazakh. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 151-163). Cham: Springer International Publishing.

16. Manni, F., & Nerbonne, J. (2021). Linguistic Diversity and Human Migrations in Gabon. *Human Migration: Biocultural Perspectives*, 99.

17. Mansurova, M. E., & Rakhimova, D. R. (2024). Morphological parsing of Kazakh texts with deep learning approaches. *Journal of Mathematics, Mechanics and Computer Science*, *124*(4), 48-58. DOI: https://doi.org/10.26577/JMMCS2024-v124-i4-a4

18. Moore, P. (2023). Translanguaging in CLIL. In *The Routledge handbook of content and language integrated learning* (pp. 28-42). Routledge.

19. Multimedia Corpus of Modern Spoken Kazakh Language Project. (2024). Project description. Nazarbayev University.

20. Mustajoki, A., Protassova, E., & Yelenevskaya, M. (2020). *The soft power of the Russian language*. Routledge.

21. Pavlenko, A. (2008). Multilingualism in post-Soviet countries: Language revival, language removal, and sociolinguistic theory. *International journal of bilingual education and bilingualism*, *11*(3-4), 275-314.

22. Salaev, U. (2024, November). UzMorphAnalyser: A morphological analysis model for the Uzbek language using inflectional endings. In *AIP Conference Proceedings* (Vol. 3244, No. 1, p. 030058). AIP Publishing LLC. DOI: https://doi.org/10.1063/5.0241461

23. Sharipbay, A., Gatiatullin, A., Yergesh, B., & Kazhymukhan, D. (2018). Development of a unified meta language of the turkic languages morphology. *Journal of Mathematics, Mechanics, Computer Science*, *4*(100), 78-87. DOI: https://doi.org/10.26577/JMMCS-2018-4-557

24. Smagulova, J. (2016). The re-acquisition of Kazakh in Kazakhstan: Achievements and challenges. *Language change in central Asia*, 89-108.

25. Smagulova, J., & Ahn, E. S. (Eds.). (2016). *Language Change in Central Asia* (Vol. 106). Walter de Gruyter GmbH & Co KG.

26. Yiner, Z., & Kurt, A. (2021). Two level Kazakh morphology. *Acta Infologica*, *5*(1), 79-98. DOI: https://doi.org/10.26650/acin.842758

27. Wei, L. (2015). *Multilingualism in the Chinese diaspora worldwide*. Taylor & Francis.