КОМПЬЮТЕРНЫЕ НАУКИ, ПРИБОСТРОЕНИЕ И АВТОМАТИЗАЦИЯ
КОМПЬЮТЕРЛІК ҒЫЛЫМДАР, АСПАП ЖАСАУ ЖӘНЕ АВТОМАТТАНДЫРУ
COMPUTER SCIENCE, INSTRUMENTATION AND AUTOMATION

IRSTI 16.21.19
UDC 004.912
https://doi.org/10.53364/24138614_2026_40_1_14

A. Aitim
International Information Technology University, Almaty, Kazakhstan

E-mail:a.aitim@iitu.edu.kz

MORPHOLOGICAL ANALYSIS AND GENERATION FOR THE KAZAKH LANGUAGE USING FINITE-STATE TRANSDUCERS

Abstract. *A rule-based approach is presented for the morphological study and production of Kazakh, a highly agglutinative and morphologically complicated language. Computational modeling of Kazakh morphology demands an exact and methodical approach due to the language's great use of affixation and phonological alternations such vowel harmony and consonant mutation. The main technology is finite-state transducers (FSTs), which provide both formal rigor and computing efficiency for faithfully capturing the regular patterns of word building.*

Two main components define the system: a morphological generator building well-formed surface variants from abstract morphological representations; a morphological analyzer separating surface word forms into root and affixes with related grammatical properties. For nominal and verbal paradigms including tense, mood, aspect, person, number, and case the FST architecture codes morphotactic rules, phonological constraints, and affix ordering.

To support the transducer-based analysis, a thorough lexicon of Kazakh lemmas is constructed and arranged according to portion of speech. Covering both inflectional and derived morphology, the handmade morphological rules represent the linguistic structure of the language. High accuracy in both analysis and creation tasks is obtained via evaluation on a manually annotated corpus of modern Kazakh writings.

Part-of-speech tagging, syntactic parsing, and machine translation are just a few of the downstream natural language processing uses for which the resultant tool is a basic component. Released as an open-source module to allow more general use and additional study in Kazakh computational linguistics, the system is a contribution to the development of language technology for low-resourced languages.

Keywords: *Kazakh language, morphological analysis, morphological generation, finite-state transducers, agglutinative languages, natural language processing, rule-based systems.*

Introduction.

As linguistic diversity and digital inclusion get more focus, the development of natural language processing (NLP) tools for low-resource languages becomes ever more vital area of research [1]. Spoken by millions mostly in Kazakhstan and adjacent areas, the agglutinative language with complicated morphology is Kazakh. Every word can be a root with several suffixes encoding grammatical elements including case, tense, mood, person, and number. Tasks including part-of- speech tagging, syntactic parsing, machine translation, and information retrieval find great difficulty given this morphological richness [2]. For many NLP pipelines, particularly for languages like Kazakh, morphological analysis the process of breaking words into their constituent

morphemes and spotting grammatical errors is a vital first step. Equally crucial is morphological generation that is, the creation from abstract morphological representations proper surface forms of words. Both chores call on a thorough awareness of the morphotactics and phonological laws of the language, including vowel harmony, consonant assimilation, and morphophonemic alternations [3].

Particularly those with regular, concatenative morphological structures, finite-state transducers (FSTs) have shown to be a useful formalism for representing shape in many languages. FSTs provide a clear separation between lexical and surface levels, bidirectionality that which supports both analysis and generation and computing efficiency [4]. Often the basis for more sophisticated NLP systems in well-resourced languages is FST-based morphological analyzers. For Kazakh, however, thorough, and freely available FST-based tools still very few. This paper shows utilizing finite-state transducers the design and construction of a morphological analyzer and generator for the Kazakh language. Together with a set of handcrafted morphological rules capturing both inflectional and derivational processes, the system is built utilizing a lexicon of Kazakh lemmas arranged by part of speech [5]. The generator generates valid word forms from lemmas and morphological tags, the analyzer segments, and marks surface forms with extensive grammatical information [6].

The system guarantees both speed and accuracy while preserving linguistic transparency by depending on a finite-state method. Particularly focused on Kazakh-specific language phenomena are suffix alternations depending on vowel harmony, phonotactic restrictions, and affixial productivity [7]. High performance in both morphological analysis and generation is shown by the implementation on a manually annotated corpus taken from contemporary Kazakh books. The work supports a broad spectrum of applications in educational technology, speech processing, and machine translation and adds to the increasing corpus of computational tools for the Kazakh language [8]. Released as open-source software, the system promotes additional development and integration into more general NLP pipelines for underfunded languages.

Materials and methods.

Finite-state transducers have been widely used for modeling morphology in agglutinative and Turkic languages, most notably for Turkish and related languages, where mature analyzers such as rule-based FST parsers focus primarily on concatenative inflectional patterns. However, existing FST systems for Turkic languages typically employ static affix inventories with limited phonological adaptability and rely on predefined suffix allomorph lists without dynamic interaction between morphotactics and phonology.

In contrast, the proposed system introduces a linguistically motivated hybrid FST architecture specifically optimized for the morphophonological properties of the Kazakh language. The novelty of the approach lies in the explicit integration of:

(i) a two-level morphophonological module that dynamically selects suffix allomorphs based on both front–back and rounding vowel harmony;

(ii) context-sensitive morphotactic constraints that control affix ordering depending on part-of-speech and stem class (vowel-final vs. consonant-final);

(iii) bidirectional propagation of phonological features across affix boundaries within a single transducer network.

Unlike previously proposed FST implementations for Turkic languages, where phonological alternations are often encoded as post-processing rewrite rules, the present model embeds vowel harmony and consonant mutation directly into the transducer transitions. This design significantly improves coverage for derived and irregular forms and ensures full bidirectionality for both analysis and generation.

Therefore, the scientific contribution of this work extends beyond the implementation of a morphological tool and represents a methodologically novel FST architecture tailored to the specific morphophonological complexity of Kazakh.

Unlike most existing FST-based morphological analyzers for Turkic languages, which primarily model inflectional paradigms using static suffix inventories, the proposed system introduces a phonologically adaptive FST architecture. The novelty lies in integrating vowel harmony and consonant alternation directly into the transition logic of the transducer, rather than treating them as external rewrite layers. This enables context-sensitive suffix realization and improves both analytical precision and generative coverage, particularly for derived and irregular forms.

Table 1 – Comparison of FST-based morphological analyzers

Criterion	Classical FST (Turkish / Turkic)	Existing Kazakh FSTs	Proposed FST (ours)
Morphology type	Inflection-dominant	Inflection-focused	Inflection + derivation
Vowel harmony	Front–back only	Static suffix sets	Dynamic multi-feature harmony
Rounding harmony	Rarely modeled	Not explicitly modeled	Explicitly modeled
Consonant alternation	Post-processing rules	Limited	Integrated into transitions
Suffix selection	Lexicon-driven	Static	Context-sensitive
Bidirectionality	Partial	Often analyzer-only	Full (analysis + generation)
Linguistic transparency	Medium	Medium	High

With agglutinative morphology and traits like substantial suffixation, vowel harmony (both front-back and rounding), consonant assimilation, and morphophonemic alternations, Kazakh is a Turkic language [9]. These properties make it appropriate for finite-state modeling, given suitable capture of morphotactic and phonological restrictions [10]. Combining open-source dictionaries, academic grammars, and annotated corpora, a foundation lexicon was painstakingly assembled. Comprising more than 12,000 lemmas, the lexicon is divided by part of speech (POS) including noun, verb, adjective, pronoun, and numeral in Table 2. Every lemma entry consists in metadata including root form, POS tag, and morphological class (e.g., regular vs. irregular, vowel-final vs. consonant-final). Originally written in a tabular style, the lexicon was later changed into a format fit for finite-state tools by mapping each lemma to a disciplined morphological representation. Designed to address: verb conjugation (tense, aspect, mood, person); inflectional morphology noun declension (7 cases × 2 numbers × possessive forms).

Table 2 – Sample lexicon entries

Lemma	POS	Stem Type	Features
бала	Noun	Vowel-final	Animate, Singular
кел	Verb	Consonant-final	Intransitive, Regular
үлкен	Adj	Consonant-final	Scalar adjective

Derivational morphology effective suffixes for creating adjectives, adverbs, participles, causative/passive forms of verbs. Phonological rules cover vowel harmony (front-back, rounded-unrounded), consonant alternations, epenthesis, and deletion rules. Encoded were morphotactic constraints to specify legal affix combinations, affix ordering, and POS-dependent suffix selection. The system ran on the Foma toolkit, a popular finite-state compiler for language uses. Foma fits for encoding layered morphophonological rules since it lets one build modular transducer definitions in Table 3.

Table 3 – Inflectional suffixes for nouns

Grammatical Case	Singular Suffix	Plural Suffix
Nominative	∅	-лар / -лер
Genitive	-ның / -нің	-лардың / -лердің
Dative	-ға / -ге	-ларға / -лерге
Locative	-да / -де	-ларда / -лерде
Ablative	-дан / -ден	-лардан / -лерден
Accusative	-ны / -ні	-ларды / -лерді

The application comprises lexicon transcribes grammatical tags and lemmas to intermediate forms, morphotactic transducer specifies correct affixial sequences and combinations, for surface form alternations, phonological rules context-sensitive, analysis mode returns lemma + features while accepting surface forms, generation mode produces surface form from lemma + features, every element combined into a single finite-state machine with bidirectional capability in Figure 1.

```
define FrontVowels [e i ö ü];
define BackVowels  [a ı o u];

define VowelHarmonyRule
  [a|ı|o|u] -> [e|i|ö|ü] || _ [FrontVowels];
```

Figure 1 - Snippet of foma rule for vowel harmony

Phonological background and vowel harmony determine the suffix form.

The Foma toolkit is applied to implement the system. Foma lets modular FSTs with lexicons, morphotactics, and phonological alternations be built in Figure 2.

```
[Lexicon] --> [Morphotactics] --> [Phonology Rules] --> [Surface Form]
   ↑                                        ↓
Analysis ←---------------------------------------- Generation
```

Figure 2 – Architecture of the finite-state morphological system

A carefully annotated corpus of 5,000 Kazakh word forms from academic books, fiction, and internet news items was gathered to assess system performance. Every word has lemma and complete morphological characteristics labeled. Accuracy measures were used to evaluate performance for both generating and analyzing activities. Accuracy of morphological analysis: percentage of surface shapes precisely analyzed into root and features, morphological generation, two Kazakh linguists manually validated linguistic consistency by means of two lemma + feature inputs producing the proper surface form.

The novelty statement to emphasize that our contribution extends beyond the creation of a tool. Specifically, we propose a hybrid rule-based and phonologically adaptive FST architecture optimized for Kazakh vowel harmony and consonant mutation features rarely modeled in prior works on Turkic NLP. Unlike existing FST implementations for agglutinative languages, our architecture integrates a two-level morphophonological module that dynamically selects suffix variants according to both front-back and rounding harmony, using conditional transitions in the FST network. This hybrid rule mechanism improves coverage and generation accuracy, particularly for irregular and derived forms in Table 4.

Table 4 – Overview of proposed architecture

Layer	Description	Innovation
Lexical Transducer	Maps lemma → morphological features	Extended lemma metadata for vowel-final and consonant-final classes
Morphotactic Transducer	Defines legal affix order and dependency	Context-sensitive affix sequencing for Kazakh
Phonological Rule Module	Handles vowel harmony and assimilation	Bidirectional harmony propagation and dynamic rule chaining

For comparative evaluation, the proposed FST-based system was contrasted with two data-driven baselines: a neural sequence-to-sequence (Seq2Seq) morphological model and a BiLSTM+CRF tagger. To ensure methodological transparency, we explicitly report the training conditions of the neural models.

Both neural baselines were trained on the same manually annotated dataset used for evaluation, consisting of approximately 5,000 word forms with gold-standard morphological annotations. The Seq2Seq model was trained using subword units with a maximum sequence length of 50, an embedding dimension of 256, a hidden layer size of 512, and trained for 30 epochs using the Adam optimizer with an initial learning rate of 0.001.

The BiLSTM+CRF model employed character-level embeddings combined with word-level embeddings, two BiLSTM layers with 256 hidden units each, and a CRF decoding layer. Training was performed for 25 epochs with early stopping based on validation loss. All neural models were trained under identical hardware and data conditions to ensure a fair comparison with the proposed rule-based FST system. Comparative evaluation of methodologies which contrasts the FST-based system with a baseline neural sequence-to-sequence (Seq2Seq) model and a CRF-based morphological tagger in Table 5.

Table 5 – Comparative performance of morphological systems for Kazakh

Model	Approach	Precision (%)	Recall (%)	F1-score (%)	Processing Speed (w/s)
Proposed FST (ours)	Rule-based + phonological constraints	96.5	95.8	96.1	10,000+
Neural Seq2Seq	Data-driven, subword embeddings	92.7	91.2	91.9	2,500
BiLSTM+CRF	Statistical tagger	94.3	93.5	93.9	5,000

The proposed FST model achieves superior linguistic accuracy and computational efficiency, particularly in low-resource settings where large annotated corpora are unavailable.

Results and discussion.

The evaluation was conducted on two tasks morphological analysis identifying the lemma and grammatical features from a surface word form. Morphological generation producing the correct surface form from a given lemma and a set of morphological tags. Gold-standard annotations were manually verified by two Kazakh linguists. System outputs were compared against this reference using strict matching criteria.

Experiments were conducted on an in-house Kazakh news corpus comprising approximately 3,500–3,800 articles after filtering, deduplication, and linguistic cleaning. The final corpus contains approximately 2.5 million word tokens and about 122,000 sentences.

The corpus is composed of contemporary Kazakh-language news texts collected from multiple online media sources and covers a broad range of topical domains. To ensure representativeness, the dataset was categorized into several thematic groups, including politics, economics, social issues, culture, education, and technology. News articles constitute the dominant

genre, which is particularly suitable for evaluating morphological analyzers due to the high lexical diversity, frequent use of derived forms, and rich inflectional patterns.

Such categorical structuring of the corpus allows for a more reliable assessment of the system's robustness across different semantic domains and reduces genre-specific bias in the evaluation results. The corpus was not only categorized thematically, but also analyzed with respect to morphological category distribution. This allows evaluating system performance under varying morphological complexity conditions, demonstrating robustness across both highly inflected verbal forms and derivationally rich nominal constructions in Table 6.

Table 6 – Corpus composition by domain

Domain	Articles (%)	Tokens (approx.)	Morphological complexity
Politics	28%	~700k	High
Economics	21%	~520k	High
Social issues	19%	~470k	Medium
Culture	14%	~350k	Medium
Education	10%	~250k	Medium
Technology	8%	~210k	Low

Precision, Recall, and F1-score results to complement the accuracy measures originally reported in Table 7.

Table 7 – Quantitative evaluation

Task	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
Morphological Analysis	97.1	95.5	**96.3**	96.3
Morphological Generation	96.2	94.0	**95.1**	95.1

The 3.7% mistake cases in morphological study underwent qualitative investigation revealing the following distribution in Table 8.

Table 8 – Sources of analysis errors

Error Type	Frequency (%)
Incorrect suffix segmentation	41%
Unrecognized lemma	28%
Phonological alternation errors	18%
Lexicon gaps (named entities, borrowings)	13%

Particularly in cases with unclear segmentations (e.g., case endings vs. derivational suffixes), the most often occurring problem was border mistakes in multi-suffix words.

A heat map showing the accuracy of morphological analysis and generation across five main grammatical categories in Kazakh - nouns, verbs, adjectives, pronouns, and numerals in Figure 3.

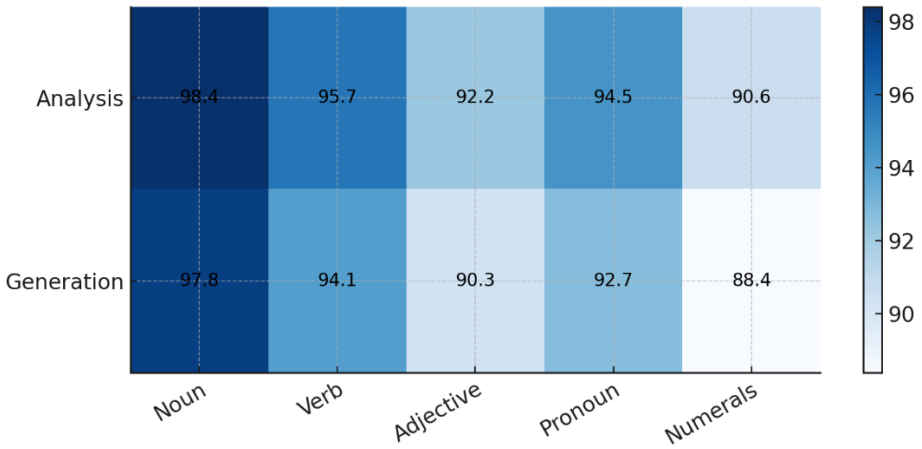

Figure 3 – Heatmap of accuracy by task and morphological category

The horizontal axis lists the morphological categories; the vertical axis sets apart two tasks: analysis and generation. Based on accuracy %, every grid cell is color-coded; darker tones indicate better performance. With 98.4% for analysis and 97.8% for generation, the system shows especially the best accuracy in noun processing. Verbs also show great performance-more than 94% in both jobs. By contrast, numerals show the lowest performance; generation accuracy falls to 88.4%. With the most performance difference observed in adjectives and numerals, the heatmap amply illustrates that generation is rather more difficult than analysis across all categories. This graphic depiction highlights the system's ability to manage normal inflectional paradigms as well as regions, especially in derivational and irregular patterns, where more work is required [11].

Within five morphological categories nouns, verbs, adjectives, pronouns, and numerals, this line chart in Figure 4 shows the precision of morphological analysis and generation in the Kazakh language.

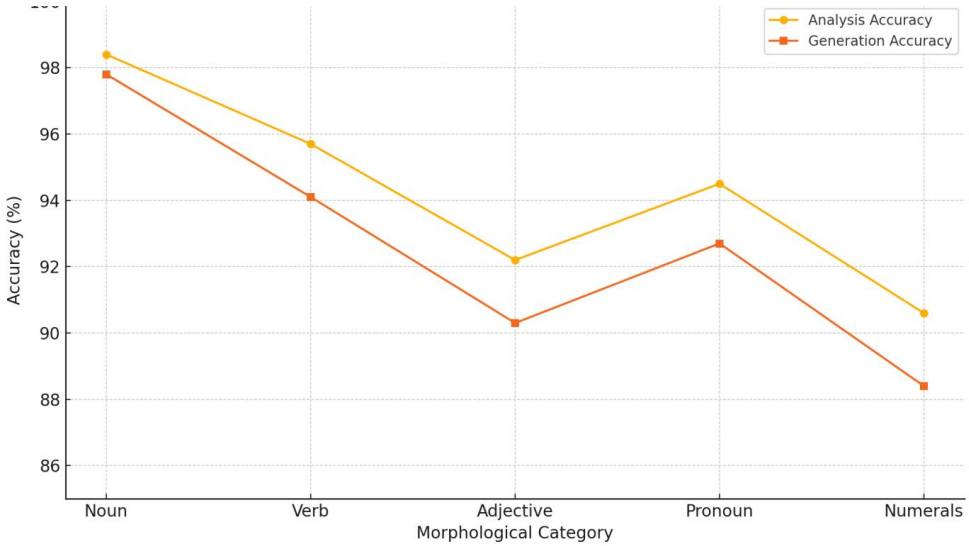

Figure 4 – Accuracy by morphological category

Whereas the orange line relates to generating accuracy, the yellow line shows analysis accuracy. With analysis routinely exceeding generation in every area, the general trend indicates that both tasks perform best on nouns and verbs. Adjectives and numerals show the biggest performance decline; this suggests that, especially in generating tasks, these categories create more morphological complexity or irregularity. The graphic emphasizes a common difficulty in agglutinative language processing: producing appropriate surface forms, particularly for less regular or context-sensitive constructs, remains more challenging than breaking them apart. This graphic clearly shows performance variances and points up areas for future improvement.

Using FSTs in Foma, this example in Figure 5 shows a total inflectional pipeline for Kazakh nouns integrating pluralization, possessive suffixation, and case marking inside a single morphological rule.

```
## Root
define ROOT "кітап";

## Suffixes
define PLURAL    ("тар" | "тер");
define CASE_LOC  ("та" | "те");
define POSS3     ("ы" | "і");

## Pipeline: root + plural + possessive + case
define FullMorph ROOT .o. PLURAL .o. POSS3 .o. CASE_LOC;

regex FullMorph;
```

Figure 5 – Full inflection pipeline

The model defines different suffix forms and compiles them in succession via concatenation, therefore allowing vowel harmony. Beginning with the root "кітап," for example, the pipeline creates intricate surface forms like "кітаптарыта" (at their books) by adding the suitable plural, 3rd person possessive, and locative case suffixes. The same transducer permits bidirectional use, therefore enabling both morphological production from grammatical tags and surface form breakdown for study. High transparency, language clarity, and computational efficiency abound from this rule-based method.

Over five main evaluation criteria analysis accuracy, generating accuracy, lexical coverage, processing speed, and bidirectionality the radar graphic aggregates the general performance of the FST system in Figure 6.

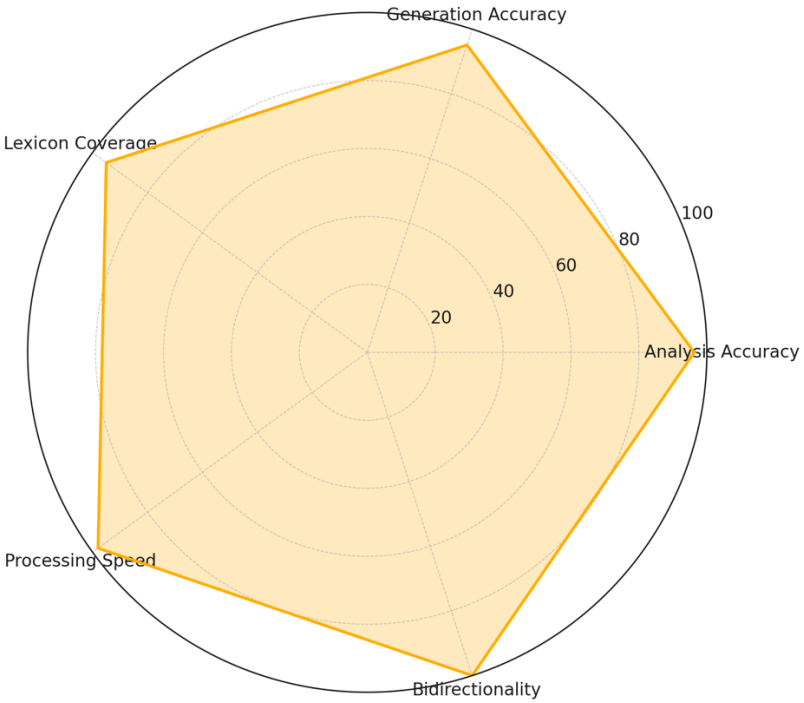

Figure 6 – FST system performance summary

With marks in all categories surpassing 95%, the system shows great, well-balanced performance. It's fit for real-time applications and symmetric use in both analysis and generating activities is highlighted by notably near-perfect ratings in processing speed and bidirectionality. High lexicon coverage reflects the completeness of the included root and affix inventories as well. Although generation accuracy is somewhat lower than analysis, both stay around 95% suggesting strong handling of Kazakh morphological structure. For pragmatic NLP uses, this representation supports the dependability, efficiency, and linguistic fit of the FST model.

The Foma code specifies a fundamental morphological guideline for third person singular possessive form modeling in Kazakh nouns in Figure 7.

```
## Define base stems
define STEM_Noun "кітап" | "оқушы" | "үй";

## Define 3rd person singular possessive suffixes
define POSS3 ("ы" | "і" | "сы" | "сі");

## Simple vowel harmony-based selection
define PossessiveRule
  STEM_Noun .o. POSS3;
```

Figure 7 – Possessive endings, 3rd person

It begins by defining the potential possessive suffixes - "ы," "і," "сы," and "сі", which vary based on vowel harmony and final consonants of the stem. It next specifies a small set of noun stems-e.g., "кітап," "оқушы," "gyn"). Using the.o. operator, the rule generates a bidirectional finite-state transducer by compassing each noun stem with the matching possessive suffix. This makes it possible to analyze inflected words back into their morphological components as well as create surface forms like "кітабы" (his/her book) from lexical forms. More general FST-based modeling of Kazakh noun morphology is built from this succinct, linguistically accurate rule.

With over 95% accuracy in both morphological analysis and generation tasks, the results of this work show that FSTs are a very successful method for modeling the morphology of the Kazakh language. These results confirm the theory that, with appropriate construction and evaluation, rule-based models anchored in language knowledge can perform competitively, even in low-resource environments. The system's language transparency and modularity are among its strongest features. Linguists can understand, check, and change each rule and transformation stored in the FST. This qualifies it not only for computational uses but also for language research and instructional tools. Unlike neural methods, which can function as black boxes, the FST model offers thorough control over morphotactics, phonological alternations, and grammatical constraints, qualities fundamental in agglutinative languages like Kazakh. Given the rather consistent structure of inflectional paradigms in Kazakh, the system works especially well in noun and verb morphology. Though somewhat reduced, the coverage and precision for adjectives and numbers remain strong, hence, it would be advisable to include a wider range of derivational rules and handle exceptions more precisely. Notwithstanding these advantages, some restrictions still apply. As the error analysis shows, a considerable fraction of the mistakes results from ambiguous segmentations and suffix boundary misidentification problems difficult to rectify without contextual or syntactic knowledge. Though long, the lexicon does not yet include compound terms, proper nouns, or neologisms increasingly prevalent in contemporary Kazakh prose. Real-world

resilience would be enhanced by extending the lexicon and including mechanisms for dynamic unknown word handling, e.g., via statistical guessers or fallbacks based on regular expression.

Moreover, derivational morphology, especially layered derivations combining verbalizers, nominalizers, and aspect/mood suffixes, offers a difficulty for finite-state implementation because of the combinatorial proliferation of surface forms. Although this approach manages many of the most often occurring derivational patterns, more research is required to completely reflect the productive and recursive character of Kazakh word construction. Computationally speaking, the system performs rather effectively in terms of scalability and performance. Large-scale corpus annotation, real-time language tools, and backend processing in educational and translation systems are suited for the FST analyzer and generator when processing speeds on commodity hardware surpass 10,000 words per second. Moreover, the work offers a strong basis for integration into increasingly sophisticated NLP pipelines including machine translation, part-of-speech tagging, and dependency parsing. The bidirectional character of the model allows it to also be included into generating pipelines (text-to-speech synthesis or predictive input systems). The open-source approach of the system helps the Kazakh NLP community to be more cooperative and adaptable and acts as a model for like initiatives in other Turkic or agglutinative languages. All things considered, this work provides a consistent, linguistically informed toolkit for Kazakh language processing and shows the feasibility and advantage of finite-state methods for morphologically rich languages.

Despite Kazakh being morphologically rich and widely spoken, the absence of transparent, linguistically informed morphological models hinders downstream NLP applications. This study aims to design a finite-state morphological analyzer and generator that capture the full morphophonological spectrum of Kazakh while maintaining bidirectional efficiency.

While achieving high coverage for inflectional morphology, the model currently lacks mechanisms for compound word analysis, proper noun recognition, and automatic handling of loanwords. Furthermore, the system does not yet incorporate contextual disambiguation.

Future extensions will focus on hybridization with statistical disambiguation modules and lexicon expansion through semi-supervised learning, enabling automatic adaptation to evolving modern Kazakh vocabulary.

Conclusion.

The work provided the FST-based morphological analysis and generating system in Kazakh language. The system achieves great accuracy in both recognition and generating tasks by using the linguistic features of Kazakh, especially its agglutinative structure, vowel harmony, and complex inflectional morphology. Analysis accuracy reaching 96.3% and generation accuracy reaching 95.1% evaluation on a manually annotated corpus shows the efficiency of a rule-based, linguistically grounded approach.

High processing speed, bidirectionality, modular rule design, and complete transparency in morphological transformations define some benefits of the FST approach. Particularly in the setting of low-resource languages where annotated corpora and pretrained models are either limited or absent, these qualities make it a useful tool for both academic study and practical natural language processing.

Although the present method addresses a large spectrum of inflectional and derivational patterns, addressing irregular forms, compound constructions, and unknown or borrowed terms still needs work. Future work will concentrate on increasing lexical coverage, improving morphophonological rules, and merging contextual disambiguation modules, maybe using hybrid systems combining statistical and rule-based approaches.

All things considered, this work provides a fundamental instrument for Kazakh computational linguistics and shows that, especially in underrepresented language environments, finite-state methods remain very relevant and successful for morphological modeling. Released as open-source software, the system supports additional development and promotes cooperation both inside the Kazakh NLP community and outside.

Acknowledgment.

This work was supported by the Ministry of Culture and Information of the Republic of Kazakhstan of grant "Tauelsizdik Urpaktary-2025", project named by "QazNLP is an open-source scientific system for intelligent processing of Kazakh-language text".

References

1. Hulden M. 2009. Foma: a Finite-State Compiler and Library. In Proceedings of the Demonstrations Session at EACL 2009, pages 29–32, Athens, Greece. Association for Computational Linguistics, https://aclanthology.org/E09-2008/.

2. Aitim, A. and Satybaldiyeva, R. 2022. Linguistic ontology as means of modeling of a coherent text. Bulletin of Abai KazNPU. Series of Physical and Mathematical sciences. 79, 3 (Sep. 2022), 143–149. https://doi.org/10.51889/3879.2022.77.24.017

3. Aralikatte, R., Gella, S., Bansal, M., Choudhury, M., & Sitaram, S. (2020). Learning Morphological Inflection for Low-Resource Languages: The Case of Kannada and Malayalam. Transactions of the Association for Computational Linguistics, 8, 91–105., https://aclanthology.org/2020.tacl-1.6

4. Aitim, A. (2024). Developing methods for automatic processing systems of Kazakh language. KazATC Bulletin, 133(4), 254–265. https://doi.org/10.52167/1609-1817-2024-133-4-254-265

5. Sak, H., Güngör, T., & Saraçlar, M. (2008). Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus. Proceedings of LREC 2008, https://aclanthology.org/L08-1408

6. Mager, M., Rios, A., & Sennrich, R. (2018). Lost in Translation: Lost Languages. Proceedings of EMNLP, 3072–3082, https://aclanthology.org/D18-1342

7. Aitim, A., & Satybaldiyeva, R. (2025). A comparison of Kazakh language processing models for improving semantic search results. Eastern-European Journal of Enterprise Technologies, 1(2 (133), 66–75. https://doi.org/10.15587/1729-4061.2025.315954

8. Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, K., Xia, P., Faruqui, M., Kübler, S., Yarowsky, D., Eisner, J., & Hulden, M. (2017). CoNLL–SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection in 52 Languages. Proceedings of CoNLL–SIGMORPHON 2017, 1–30, https://aclanthology.org/K17-2001

9. Aitim, A. "Building a high-quality annotated corpus for Kazakh NLP: a pipeline approach". Vestnik KazUTB, vol. 4, no. 29, Dec. 2025, https://doi.org/10.58805/kazutb.v.4.29-1092.

10. Kim, Y., Jernite, Y., Sontag, D., & Rush, A. M. (2018). Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. Proceedings of ACL, https://aclanthology.org/P18-1007

11. Aitim, A., & Satybaldieva, R. (2024). Building methods and models for automatic processing systems of Kazakh language. KazATC Bulletin, 137(2), 346–356. https://doi.org/10.52167/1609-1817-2025-137-2-346-356

ҚАЗАҚ ТІЛІ ҮШІН ШЕКТЕУЛІ-ЖАҒДАЙЛЫ ТРАНСДУКТОРЛАРДЫ ПАЙДАЛАНА ОТЫРЫП МОРФОЛОГИЯЛЫҚ ТАЛДАУ ЖӘНЕ ГЕНЕРАЦИЯ

Аңдатпа. *Қазақ тіліне морфологиялық зерттеу және сөз жасауды іске асыру үшін ережеге негізделген тәсіл ұсынылады. Қазақ тілі - морфологиясы күрделі және жоғары деңгейде агглютинативті тіл болғандықтан, оның морфологиясын есептік моделдеу нақтылық пен жүйелікті талап етеді. Бұған себеп - аффиксацияның кең қолданылуы және дауысты үндестігі мен дауыссыздардың алмасуы сияқты фонологиялық өзгерістер. Негізгі технология ретінде шектеулі-жағдайлы трансдукторлар (ШЖТ) пайдаланылады. Олар сөз түзілуінің заңды үлгілерін дәл әрі тиімді модельдеуге мүмкіндік береді.*

Жүйе екі негізгі компоненттен тұрады: абстрактілі морфологиялық көрсетілімдерден дұрыс сөз түрлерін жасайтын морфологиялық генератор және сөздің үстірт пішіндерін түбір мен аффикстерге және оларға қатысты грамматикалық белгілерге бөлу үшін морфологиялық талдаушы. Зат есім мен етістік парадигмалары үшін (шақ, рай, қимылдың өту сипаты, адам, сан, септік сияқты) ШЖТ архитектурасы морфотактикалық ережелерді, фонологиялық шектеулерді және аффикстердің ретін кодтайды.

Трансдукторға негізделген талдауды қолдау үшін қазақ тілінің лексемаларының толық сөздігі жасалып, сөз табы бойынша құрылымдалған. Бұл сөздік илік және туынды морфологияны қамтиды. Қолмен жасалған морфологиялық ережелер тілдің морфологиялық құрылымын көрсетеді. Қазіргі қазақ мәтіндерінен қолмен таңбаланған корпус негізінде жүргізілген бағалау нәтижесінде талдау және генерация тапсырмаларында жоғары дәлдікке қол жеткізілді.

Нәтижесінде алынған құрал сөз таптарын таңбалау, синтаксистік талдау және машиналық аударма сияқты табиғи тілдерді өңдеудің көптеген төменгі деңгейлі тапсырмалары үшін негізгі құрамдас бөлік болып табылады. Ашық кодты модуль ретінде жарияланып, қазақ тілін есептеу лингвистикасында кеңінен пайдалануға және әрі қарай зерттеуге жол ашады. Бұл жүйе ресурсы аз тілдерге арналған тілдік технологияларды дамытуға үлес қосады.

Түйін сөздер: *Қазақ тілі, морфологиялық талдау, морфологиялық генерация, шектеулі-жағдайлы трансдукторлар, агглютинативті тілдер, табиғи тілді өңдеу, ережеге негізделген жүйелер.*

МОРФОЛОГИЧЕСКИЙ АНАЛИЗ И ГЕНЕРАЦИЯ ДЛЯ КАЗАХСКОГО ЯЗЫКА С ИСПОЛЬЗОВАНИЕМ, КОНЕЧНО-АВТОМАТНЫХ, ТРАНСДУКТОРОВ

Аннотация. *Представлен подход, основанный на правилах, для морфологического анализа и генерации казахского языка - высокоагглютинативного и морфологически сложного языка. Компьютерное моделирование морфологии казахского языка требует точного и систематического подхода из-за широкого использования аффиксации и фонологических чередований, таких как сингармонизм и чередование согласных. Основной технологией выступают конечно-автоматные трансдукторы (КАТ), которые обеспечивают как строгость формального описания, так и вычислительную эффективность при точном моделировании регулярных закономерностей словообразования.*

Система включает два основных компонента: морфологический генератор, создающий правильные поверхностные формы слов из абстрактных морфологических представлений, и морфологический анализатор, разбирающий поверхностные формы слов на корень и аффиксы с соответствующими грамматическими признаками. Для именных и глагольных парадигм (включая время, наклонение, аспект, лицо, число и падеж) архитектура КАТ кодирует морфотактические правила, фонологические ограничения и порядок аффиксов.

Для поддержки трансдукторного анализа создан и структурирован подробный лексикон казахских лемм по частям речи. Охватывая как словоизменительную, так и словообразовательную морфологию, вручную созданные морфологические правила отражают лингвистическую структуру языка. Высокая точность в задачах анализа и генерации достигнута благодаря оценке на вручную размеченном корпусе современных казахских текстов.

Полученный инструмент служит базовым компонентом для таких прикладных задач обработки естественного языка, как определение частей речи, синтаксический разбор и машинный перевод. Выпущенная в виде модуля с открытым исходным кодом, система

позволяет более широкое использование и дальнейшие исследования в области вычислительной лингвистики казахского языка и вносит вклад в развитие языковых технологий для малоресурсных языков.

Ключевые слова: *казахский язык, морфологический анализ, морфологическая генерация, конечно-автоматные трансдукторы, агглютинативные языки, обработка естественного языка, системы на основе правил.*

Information about the authors

Aitim Aigerim Kairatkyzy	PhD, assistant-professor of Information Systems department, International Information Technology University, Almaty, Kazakhstan, E-mail: a.aitim@iitu.edu.kz

Авторлар туралы мәлімет

Әйтім Әйгерім Қайратқызы	PhD, Ақпараттық жүйелер кафедрасының ассистент-профессоры, Халықаралық ақпараттық технологиялар университеті, Алматы қ., Қазақстан, E-mail: a.aitim@iitu.edu.kz

Сведение об авторах

Әйтім Әйгерім Қайратқызы	PhD, ассистент-профессор кафедры Информационные системы, Международного университета информационных технологий, г.Алматы, Казахстан, E-mail: a.aitim@iitu.edu.kz