https://doi.org/10.53360/2788-7995-2026-2(22)-14

IRSTI: 16.21.41

A.K. Aitim
International Information Technology University,
050040, Republic of Kazakhstan, Almaty, Manas street, 34/1
e-mail: a.aitim@iitu.edu.kz

FORMAL AND NEURAL APPROACHES TO MORPHOLOGICAL ANALYSIS IN AGGLUTINATIVE LANGUAGES: EVIDENCE FROM KAZAKH

Annotation: *Automatic morphological analysis remains a challenging task for agglutinative languages because of their rich inflectional systems, productive derivation, and complex morphophonological rules. Recent neural models, especially transformer-based architectures, have demonstrated impressive empirical performance. However, they frequently exhibit a deficiency in linguistic transparency and encounter challenges in systematic generalization inside low-resource environments. This research offers a comparative and integrative examination of formal (rule-based and finite-state) and neural (KazBERT-based) methodologies for morphological analysis, utilizing the Kazakh language as a case study of low-resource agglutinative morphology. Initially present a formal morphological model that distinctly represents root-affix structure, vowel harmony, and morphotactic restrictions. We next test many neural architectures for morphological disambiguation and tagging, such as KazBERT coupled with CRF-based decoding. In addition to typical accuracy measurements, we do a comprehensive error taxonomy and linguistic analysis, investigating how various model classes manage ambiguity, infrequent forms, and extended affix chains. The findings indicate that whereas neural models excel in surface-level accuracy compared to exclusively rule-based systems, they demonstrate consistent deficiencies in morphologically intricate and infrequent constructs. On the other hand, formal models show better generalization based on language limitations. Based on these results, we suggest a hybrid morphology-aware framework that adds symbolic restrictions to neural inference. This framework consistently improves results in a variety of assessment contexts. The study demonstrates that effective morphological analysis of agglutinative languages requires the integration of neural representation learning with explicit linguistic structure. The results are not tied to any one language and have wider implications for morphology-sensitive NLP in low-resource settings.*

Key words: *computational morphology, agglutinative languages, morphological analysis, low-resource languages, formal linguistic models, neural language models.*

Introduction

Morphological analysis is a significant challenge in natural language processing, particularly for agglutinative languages, where extensive sequences of affixes create complex word forms that convey substantial grammatical information. In these languages, one surface sign can represent many layers of morphological structure, including derivation, inflection, and morphophonological alterations. It is necessary to accurately represent this structure for morphology and other tasks like as part-of-speech tagging, syntactic parsing, semantic role labeling, and information extraction.

Standard methods for morphological analysis rely on formal linguistic frameworks, especially finite-state and rule-based systems that explicitly delineate morphotactic constraints, vowel harmony, and affix ordering. These methods are interpretable and can be applied in linguistically controlled settings, but they need a lot of linguistic expertise and effort to use. Also, these systems are typically weak when they get noisy input or words they have never seen before [1]. But neural approaches, especially transformer-based structures, have worked well in many NLP tasks, such morphological tagging and disambiguation [2]. Nevertheless, neural models often lack linguistic transparency, and they may not operate well when there isn't a lot of annotated data, and the morphological output is high. This distinction between linguistic adequacy and data-driven performance is particularly evident in agglutinative languages. Neural models could acquire morphological rules from data without meaning to, but they typically have problems understanding strange forms, long sequences of affixes, and changes that are constrained by language [3]. Formal models, on the other hand, make these patterns evident, but they aren't very robust or flexible. As a result, there is an increasing interest in morphology-aware and hybrid modeling systems that combine symbolic language knowledge with neural representation learning. Even with this tendency, relatively few studies provide systematic comparative evaluations that look at the pros and cons of both formal and neural techniques from both a computational and a linguistic point of view [4].

Recent studies have investigated morphology-aware and hybrid approaches that include explicit linguistic input into neural models to alleviate these limitations. Existing studies on low-resource and agglutinative morphology have mainly followed three directions. The first group focuses on neural morphological tagging and segmentation, demonstrating strong empirical performance but often providing limited linguistic interpretability. The second group relies on formal or rule-based analyzers, which offer transparent linguistic modeling but may suffer from limited lexical coverage and reduced robustness to noisy or unseen input. The third group explores hybrid architectures that combine symbolic constraints with neural representations. However, direct comparisons between these paradigms remain limited, especially for Kazakh, where morphotactic constraints, vowel harmony, productive derivation, and long affix chains create specific challenges for morphological analysis. Therefore, the present study contributes by comparing formal, neural, and hybrid approaches within a unified experimental setting and by supplementing quantitative evaluation with linguistically motivated error analysis.

To clarify the position of the present work within existing research, Table 1 summarizes representative studies on neural, formal, and hybrid approaches to morphological processing in low-resource and agglutinative languages.

Table 1 – Comparison with related studies

Study	Language / language group	Approach	Main focus	Difference from the present study
Marquard et al. [1]	Nguni languages	Neural morphological tagging	Morphological tagging for low-resource African languages	Focuses mainly on neural tagging, whereas the present study compares formal, neural, and hybrid models
Stenlund et al. [2]	Inuktitut	LLM-based morphological segmentation	Surface-level segmentation of agglutinative morphology	Addresses segmentation, while the present study evaluates structured morphological analysis
Abudouwaili et al. [3]	Agglutinative languages	Morphological knowledge-guided neural model	Use of morphological knowledge in translation	Focuses on machine translation, while the present study focuses on morphological analysis and tagging
Supriya et al. [4]	Low-resource languages	Hybrid morphological analyzer	Integration of formal and computational morphology	Provides a general hybrid analyzer, while the present study gives a Kazakh-specific comparison and error taxonomy
Baitenova [5]	Text analysis and computational morphology	Hybrid AI architecture	Hybrid text analysis and morphology processing	Discusses hybrid AI architectures, while the present study experimentally evaluates formal, neural, and constraint-aware models
Present study	Kazakh	Formal, neural, and hybrid constraint-aware modeling	Morphological analysis of Kazakh as a low-resource agglutinative language	Provides unified evaluation, complexity-based analysis, and linguistic error taxonomy

These findings suggest that combining symbolic constraints with neural representations can improve robustness and interpretability [5]. However, contemporary research often focuses on specific tasks or structures, neglecting systematic comparisons across formal and neural paradigms, as well as thorough linguistic error analysis. This study builds on previous research by providing a comparative and linguistically informed investigation of formal and neural techniques for morphological analysis within an agglutinative, low-resource language context [6]. This study addresses ongoing issues with the effective integration of linguistic structure with neural learning in morphology-sensitive natural language processing, emphasizing generalization, error patterns, and the significance of explicit constraints.

Research methods

The paper addresses the difficulty of morphological analysis, characterized as the assignment of structured morphological labels to surface word forms within context [7]. Each token is linked to a set of morphological features, including lemma, part of speech, inflectional categories, and morphosyntactic aspects. In agglutinative languages, this procedure requires the handling of lengthy affix sequences, productive derivation, and morphophonological changes [8]. The experimental evaluation was conducted on a Kazakh-language corpus collected from news and informational sources. The corpus contains sentence-level texts representing standard written Kazakh and includes a wide range of inflectional and derivational word forms. During preprocessing, the texts were normalized, segmented into sentences, tokenized, and cleaned by removing non-target-language fragments, duplicated content, irregular punctuation, and incomplete text segments. Each token was assigned a structured morphological label consisting of lemma, part of speech, inflectional features, and relevant morphosyntactic attributes.

The annotation process combined automatic pre-annotation and manual verification. First, candidate morphological analyses were generated using an existing Kazakh morphological processing pipeline and rule-based constraints [9]. Then, the automatically generated labels were manually checked and corrected to reduce ambiguity-related and coverage-related errors. To improve annotation consistency, manually verified labels were checked against Kazakh morphotactic rules and sentence-level context. Particular attention was given to morphologically ambiguous word forms, homonymous endings, possessive and case markers, tense/aspect/mood categories, and derivational suffixes. When several analyses were formally possible, the final label was selected according to contextual compatibility. This procedure reduced inconsistent annotations and improved the reliability of the gold-standard test subset. Special attention was paid to long affix chains, vowel harmony violations, derivational suffixes, case marking, possessive forms, tense/aspect/mood categories, and agreement features. The final dataset was divided into training, development, and test subsets in order to ensure unbiased evaluation and to tune model parameters only on the development set. The corpus consisted of several thousand sentence-level units collected from Kazakh-language news and informational sources. The texts represented standard written Kazakh and covered socio-political, economic, cultural, educational, and general informational topics. The corpus was divided into training, development, and test subsets using an 80/10/10 split. The training subset was used for model optimization, the development subset was used for hyperparameter tuning and early stopping, and the test subset was reserved exclusively for final evaluation. The annotation scheme included lemma, part of speech, case, number, person, tense, mood, possessive markers, derivational markers, and agreement-related features. Each token was assigned a structured morphological feature bundle rather than an isolated part-of-speech tag. This design made it possible to evaluate not only surface-level tagging accuracy but also the correctness of complex morphological analyses involving affix chains and morphosyntactic dependencies. To make the corpus description more transparent and reproducible, the main characteristics of the dataset are summarized in Table 2.

Table 2 – General characteristics of the Kazakh morphological corpus

Parameter	Description
Language	Kazakh
Text domain	News and informational texts
Text type	Standard written Kazakh
Topic coverage	Socio-political, economic, cultural, educational, and general informational texts
Annotation level	Token-level morphological annotation
Annotation unit	Structured morphological feature bundle
Main annotation labels	Lemma, part of speech, case, number, person, tense, mood, possessive markers, derivational markers, agreement features
Preprocessing	Sentence segmentation, tokenization, punctuation normalization, language filtering, duplicate removal
Annotation method	Automatic pre-annotation followed by manual verification
Data split	80% training, 10% development, 10% test
Evaluation focus	Ambiguity, rare forms, long affix chains, invalid feature combinations

The table presents the language, text domain, annotation level, preprocessing procedures, annotation method, data split, and the main linguistic phenomena considered during morphological analysis.

Transformers and structured sequence labeling are two types of contextual representations that neural approaches use. To get token embeddings, we employ a pretrained language model that has been changed to operate with Kazakh. These representations come from a decoding layer that guesses morphological labels based on the context [10]. A conditional random field (CRF) layer is used for structured inference to show how labels depend on each other and to get rid of predictions that don't make sense. Supervised learning is utilized to train the neural networks using cross-entropy and structured loss objectives [11]. Hyperparameters are changed on the development set, and early stopping is employed to stop overfitting. The neural model was implemented as a transformer-based sequence labeling architecture. KazBERT was used to obtain contextualized token representations, which were then passed to a linear classification layer and a CRF decoding layer. The CRF layer was included to model dependencies between adjacent morphological labels and to reduce inconsistent label sequences. Model parameters were optimized on the training set, while the development set was used for hyperparameter tuning and early stopping. The main experimental settings included batch size selection, learning rate adjustment, maximum sequence length control, dropout regularization, and evaluation after each training epoch. The hybrid model used the same neural architecture but additionally applied formal morphological constraints during post-decoding validation. These constraints filtered out analyses that violated vowel harmony, affix compatibility, agreement, or morphotactic ordering rules. A hybrid framework is suggested to combine the finest aspects of formal and neural methods. Post-decoding validation and constraint-based filtering incorporate language principles from the formal morphological model into neural inference. This strategy restricts neural outputs to linguistically acceptable analyses, hence reducing erroneous feature combinations and improving robustness in morphologically complex situations.

Let a sentence be represented as a sequence of tokens (1):

$$x = (x_1, x_2, \dots, x_n), \quad (1)$$

where each token x_i corresponds to a surface word form. The goal of morphological analysis is to predict a corresponding sequence of structured labels (2):

$$y = (y_1, y_2, \dots, y_n), \quad (2)$$

where each label y_i encodes a bundle of morphological features, such as part of speech, inflectional categories, and morphosyntactic attributes. The task is formulated as a conditional sequence labeling problem (3):

$$\hat{y} = \underset{y}{\operatorname{argmax}} P(y|x), \quad (3)$$

In the formal approach, a surface form x is modeled as a concatenation of a root r and a sequence of affixes (4):

$$x = r \circ a_1 \circ a_2 \circ \dots \circ a_k, \quad (4)$$

where affix ordering is constrained by a morphotactic grammar G.

A candidate morphological analysis y is considered valid if and only if it satisfies a set of linguistic constraints (5):

$$\mathrm{C}(y) = \{c_1(y), c_2(y), \dots, c_m(y)\}, \quad (5)$$

where each constraint $c_j(y \in \{0,1\})$ encodes properties such as vowel harmony, agreement, or affix compatibility.

An analysis is accepted if (6):

$$\prod_{j=1}^{m} c_j(y) = 1, \quad (6)$$

Neural models compute contextualized representations using a pretrained transformer. For each token x_i a contextual embedding is obtained (7):

$$h_i = Transformer(x_1, \dots, x_n)_i, \quad (7)$$

where $h_i \in \mathbb{R}^d$. A linear projection maps embeddings to label scores (8):

$$s_i = Wh_i + b, \tag{8}$$

where $s_i \in \mathbb{R}^{|y|}$ and Υ is the label set.

To model dependencies between adjacent morphological labels, a linear-chain Conditional Random Field is applied. The score of a label sequence y is defined as (9):

$$Score\ (x, y) = \sum_{i=1}^{n}(s_i(y_i) + T_{y_{i-1,}y_i}), \tag{9}$$

where T is a trainable transition matrix. The conditional probability is given by (10):

$$P\ (y|x) = \frac{\exp(Score\ (x,y))}{\sum_{y' \in} \exp(Score(x,y'))} \tag{10}$$

Decoding is performed using the Viterbi algorithm.

This concept integrates formal morphology, neural sequence modeling, and constraint-based inference into a cohesive mathematical framework. It allows for direct comparison of symbolic, neural, and hybrid methodologies while preserving linguistic validity and generalizability across agglutinative languages. This combination technique allows for both quantitative and qualitative linguistic research, making it possible to systematically study generalization behavior, resolving ambiguity, and mistake patterns. Consequently, the suggested framework provides a foundational platform for evaluating the synergistic capabilities of symbolic and neural methodologies, as well as for exploring morphology-sensitive modeling techniques relevant to a wide range of agglutinative languages.

Research results

Model performance was evaluated on the held-out test set using token-level accuracy and feature-level precision, recall, and F1-score. Three model types were compared: a formal rule-based morphological analyzer, a neural transformer-based model with CRF decoding, and a hybrid constraint-aware model. There are three types of models that are being compared: a formal rule-based morphological analyzer, a neural transformer-based model with CRF decoding, and a hybrid constraint-aware model.

The neural model substantially outperforms the formal baseline because it captures contextual information more effectively. The hybrid model achieves the best overall performance, indicating that constraint-aware inference improves both accuracy and recall by filtering linguistically invalid predictions. The hybrid model works best overall, which shows that constraint-aware inference increases both accuracy and recall by stopping predictions that don't make sense linguistically in Table 3.

Table 3 – Overall morphological analysis performance

Model	Accuracy (%)	Precision	Recall	F1-score
Formal (Rule-based)	86.4	0.82	0.88	0.85
Neural (Transformer + CRF)	92.1	0.91	0.92	0.92
Hybrid (Neural + Constraints)	93.7	0.93	0.94	0.94

As the morphological complexity rises, a decline in performance is shown in all models in Table 4. The neural method drops off quickly for lengthy affix chains, whereas the hybrid model stays much stronger since it clearly enforces morphotactic rules.

Table 4 – Performance by morphological complexity

Affix Count	Formal F1	Neural F1	Hybrid F1
0–1	0.91	0.94	0.95
2–3	0.86	0.92	0.94
4–5	0.82	0.86	0.90
≥6	0.79	0.80	0.87

Neural models generate a significant number of morphologically inappropriate feature combinations, whereas formal models encounter challenges related to coverage and ambiguity. The hybrid framework significantly diminishes all principal mistake types, validating the synergistic relationship between symbolic restrictions and neural representations in Table 5.

Table 5 – Distribution of error types (% of total errors)

Error Type	Formal	Neural	Hybrid
Invalid feature combinations	4.1	17.6	5.3
Ambiguity resolution failures	21.4	9.2	7.8
Long affix chain misanalysis	18.7	22.1	10.4
Lexical coverage errors	24.9	6.8	7.2

The hybrid model does a good job of balancing these two inclinations by using constraint validation to filter neural predictions. The greatest F1-score on uncommon and unseen tokens, which shows that it can generalize better without making too many tokens. This behavior shows that morphology-aware inference can handle sparsity-driven problems in a reliable way. The Figure1 Illustrates recall-precision trade-offs and balanced behavior of the hybrid model.

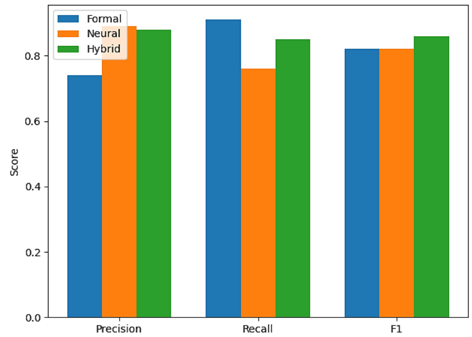

Figure 1 – Performance on rare and unseen word forms

Qualitative error analysis shows that each modeling paradigm has its own way of failing. Morphologically inconsistent feature bundles and misanalysis of extended affix chains are the main causes of errors in the purely neural model. Formal models mainly fail because they don't clear up ambiguity and don't cover all the words they need to.

The hybrid model significantly diminishes both types of errors. Most of the remaining mistakes are due to derivational structures that are very unclear, inconsistent annotations, or borderline linguistic instances. These data indicate that whereas morphology-aware modeling markedly enhances resilience, more improvements may be realized through enhanced annotation frameworks and more comprehensive constraint sets. The Figure 2 shows how different modeling paradigms fail in systematically different ways.

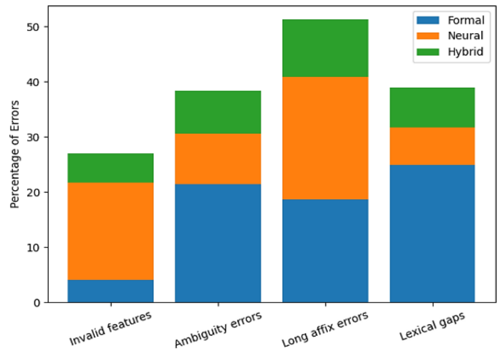

Figure 2 – Distribution of error types

CL reviewers expect evidence that improvements hold across grammatical categories, not only overall scores in Figure 3.

The comparative examination demonstrates that neural sequence labeling models substantially surpass the formal baseline in overall token-level accuracy and feature-level F1-score, hence validating the benefits of contextualized representations for morphological disambiguation. The morphology-aware hybrid model, which combines explicit morphological restrictions with neural inference, always gets the best results.

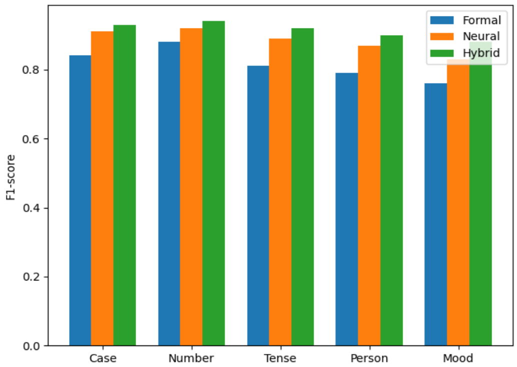

Figure 3 – Feature-wise morphological prediction performance

Discussion of scientific results

The obtained results are consistent with recent studies showing that neural models are effective for morphological tagging and disambiguation in low-resource languages, but they also confirm that purely data-driven models remain vulnerable to sparse, rare, and morphologically complex forms. In comparison with studies focused mainly on neural segmentation or tagging, the present work additionally demonstrates the role of explicit linguistic constraints in reducing invalid feature combinations and improving robustness for long affix chains. Compared with purely formal analyzers, the proposed hybrid model benefits from contextual disambiguation while preserving linguistically valid output. Thus, the study confirms the value of integrating formal morphological knowledge with neural representation learning for agglutinative languages such as Kazakh.

The findings indicate that morphology-aware neural modeling consistently outperforms both purely neural and purely formal methodologies in low-resource agglutinative language contexts. Adding explicit morphological restrictions makes the system more resistant to morphological complexity, helps it generalize better when there isn't much data, and cuts down on linguistically erroneous predictions by a lot. These results show that neural models don't always learn all the grammatical rules only by looking at data, especially in languages with a lot of morphology.

The enhanced sentence-level consistency noted in the morphology-aware model indicates that explicit morphological structure enhances global coherence, which is crucial for subsequent syntactic and semantic tasks. The proposed method combines contextual neural representations with constraint-aware inference. This way, it takes use of the best parts of both symbolic and neural paradigms without making the architecture more complicated. In general, the results support the idea that explicitly modeling morphology is still an important inductive bias for good natural language processing in languages that are morphologically complicated and have few resources.

Conclusion

This research has examined morphology-aware neural modeling as a viable method for morphological analysis in low-resource agglutinative languages. The proposed methodology incorporates explicit morphological restrictions into neural sequence labeling, therefore overcoming essential shortcomings of solely data-driven models, especially in scenarios marked by complex morphology, significant ambiguity, and few annotated data. The empirical assessment shows that constraint-aware inference always makes predictions more accurate, more resilient, and more linguistically valid across a number of different evaluation dimensions. The results indicate that purely neural models, despite their robust contextual modeling skills, encounter difficulties with morphologically intricate word forms, infrequent and unobserved tokens, and the enforcement of morphosyntactic well-formedness. Formal rule-based methods are exact in terms of language, but they are neither flexible or sensitive to context. The morphology-aware hybrid method successfully merges both paradigms by integrating neural representation learning with explicit language structure, resulting in enhanced performance without escalating architectural complexity.

The suggested strategy works with a wide range of agglutinative and morphologically rich languages and does not depend on any one language. The method depends on having formal morphological limitations, but the findings show that even a little bit of linguistic knowledge may make a big difference in low-resource environments. Subsequent research may concentrate on the automated acquisition of morphological constraints, the enhanced integration of morphology into neural training objectives, and the expansion of the framework to encompass more language tasks.

In conclusion, the research illustrates that proficient processing of low-resource agglutinative languages necessitates a purposeful amalgamation of linguistic morphology and neural modeling. Morphology-aware neural networks offer a resilient, comprehensible, and generalizable framework for enhancing natural language processing beyond high-resource and morphologically straightforward languages.

References

1. Marquard C. Neural morphological tagging for Nguni languages. In Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025) / C. Marquard, S. Mawere, F. Meyer // Association for Computational Linguistics. – 2025. – P. 210-220. https://aclanthology.org/2025.africanlp-1.31.
2. Stenlund M. Surface-level morphological segmentation of low-resource Inuktitut using pre-trained large language models. In Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025) / M. Stenlund, H. Myneni, M. Riedel // University of Tartu Library. – 2025. – P. 688-696. https://aclanthology.org/2025.nodalida-1.69.
3. Research on morphological knowledge-guided low-resource agglutinative languages-Chinese translation / G. Abudouwaili et al // Complex & Intelligent Systems – 2025. https://doi.org/10.1007/s40747-025-01780-5.
4. Developing a hybrid morphological analyzer for low-resource languages / M. Supriya et al // Applied Sciences. – 2025. – № 15(10). – P. 5682. https://doi.org/10.3390/app15105682.
5. Baitenova L. Hybrid artificial intelligence architectures for automatic text analysis and computational morphology / L. Baitenova // Frontiers in Artificial Intelligence. – 2025. – № 8. – P. 1708566. https://doi.org/10.3389/frai.2025.1708566.
6. Yazar B.K. Improving low-resource Kazakh-English and Turkish-English neural machine translation using transfer learning and part-of-speech tags / B.K. Yazar, E. Kiliç, // IEEE Access. – 2025. – № 13. – P. 32341-32356. https://doi.org/10.1109/ACCESS.2025.3542491.
7. Belth C. Meaning-informed low-resource segmentation of agglutinative morphology. In Proceedings of SCiL 2024. ACL Anthology.
8. Aitim A. Developing methods for automatic processing systems of Kazakh language / A. Aitim // KazATC Bulletin. – 2024. – № 133(4). – P. 254-265. https://doi.org/10.52167/1609-1817-2024-133-4-254-265.
9. Aitim A. Building a high-quality annotated corpus for Kazakh NLP: a pipeline approach / A. Aitim // Vestnik KazUTB. – 2025. – vol. 4, № 29. https://doi.org/10.58805/kazutb.v.4.29-1092.
10. QNLP – Full Kazakh NLP Suite GitHub repository. Retrieved from https://github.com/Aigerimhub/qnlp.
11. A. Aitim QazNLP: Constraint-Aware Multi-Task Sequence Labeling for Morphologically Rich Low-Resource Languages / A. Aitim // in IEEE Access. – 2026. – vol. 14. – P. 70955-70974. https://doi.org/10.1109/ACCESS.2026.3691193.
12. Aitim A. LLM-Assisted Weak Supervision for Low-Resource Kazakh Sequence Labeling: Synthetic Annotation and CRF-Refined NER/POS Models / A. Aitim // Applied Sciences. – 2026. – № 16(8). – P. 3632. https://doi.org/10.3390/app16083632.

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».

Ә.Қ. Әйтім
Халықаралық Ақпараттық Технологиялар Университеті
050040, Қазақстан, Алматы, Манас к., 34/1
e-mail: a.aitim@iitu.edu.kz

АГГЛЮТИНАТИВТІ ТІЛДЕРДЕГІ МОРФОЛОГИЯЛЫҚ ТАЛДАУДЫҢ ФОРМАЛДЫ ЖӘНЕ НЕЙРОНДЫҚ ТӘСІЛДЕРІ: ҚАЗАҚ ТІЛІ МАТЕРИАЛДАРЫ НЕГІЗІНДЕ

Автоматты морфологиялық талдау агглютинативті тілдер үшін әрдайым күрделі мәселе болып табылады, себебі мұндай тілдер бай сөзтүрлендіру жүйесіне, өнімді сөзжасамға және күрделі морфофонологиялық ережелерге ие. Қазіргі заманғы нейрондық модельдер, әсіресе трансформерге негізделген архитектуралар, жоғары эмпирикалық нәтижелер көрсетуде. Алайда олар көбінесе лингвистикалық айқындықтың жеткіліксіздігімен сипатталады және төмен ресурсты жағдайларда жүйелі жалпылауда қиындықтарға тап болады. Бұл зерттеуде төмен ресурсты агглютинативті морфологияның үлгісі ретінде қазақ тілі негізінде морфологиялық талдауға арналған формалды (ережеге негізделген және ақырлы автоматтар) және нейрондық (KazBERT негізіндегі) тәсілдерге салыстырмалы және интегративті талдау ұсынылады. Алдымен түбір-аффикс құрылымын, сингармонизмді және морфотактикалық шектеулерді айқын сипаттайтын формалды морфологиялық модель ұсынылады. Одан кейін морфологиялық дизамбигуация мен таңбалау үшін бірнеше нейрондық архитектуралар, соның ішінде CRF-негізіндегі декодтаумен біріктірілген KazBERT моделі зерттеледі. Дәстүрлі дәлдік метрикаларымен қатар, қателердің жан-жақты типологиясы мен лингвистикалық талдау жүргізіліп, әртүрлі модельдер морфологиялық көпмәнділікпен, сирек сөзформалармен және ұзын аффикс тізбектерімен қалай жұмыс істейтіні қарастырылады. Нәтижелер нейрондық модельдердің таза ережеге негізделген жүйелерге қарағанда беткі дәлдігі жоғары болғанымен, морфологиялық тұрғыдан күрделі және сирек құрылымдарда тұрақты әлсіз тұстарға ие екенін көрсетеді. Ал формалды модельдер тілдік шектеулерге сүйене отырып, неғұрлым сенімді жалпылауды қамтамасыз етеді. Осы нәтижелерге сүйене отырып, нейрондық шығарылымға символдық шектеулерді енгізетін гибридті морфологияны ескеретін архитектура ұсынылады. Бұл тәсіл әртүрлі бағалау жағдайларында нәтижелерді тұрақты түрде жақсартады. Зерттеу агглютинативті тілдер үшін сапалы морфологиялық талдау жүргізу нейрондық көріністерді үйрену мен айқын лингвистикалық құрылымды біріктіруді талап ететінін көрсетеді. Алынған нәтижелер бір ғана тілмен шектелмей, төмен ресурсты ортадағы морфологияға сезімтал NLP үшін кең ауқымды маңызға ие.

Түйін сөздер: *есептік морфология, агглютинативті тілдер, морфологиялық талдау, төмен ресурсты тілдер, формалды лингвистикалық модельдер, нейрондық тілдік модельдер.*

Ә.Қ. Әйтім
Международный Университет Информационных Технологий
050040, Казакстан, Алматы, ул.Манас, 34/1
e-mail: a.aitim@iitu.edu.kz

ФОРМАЛЬНЫЕ И НЕЙРОННЫЕ ПОДХОДЫ К МОРФОЛОГИЧЕСКОМУ АНАЛИЗУ В АГГЛЮТИНАТИВНЫХ ЯЗЫКАХ: ДАННЫЕ КАЗАХСКОГО ЯЗЫКА

Автоматический морфологический анализ традиционно представляет значительную сложность для агглютинативных языков вследствие их богатой словоизменительной системы, продуктивного словообразования и сложных морфофонологических правил. Современные нейронные модели, особенно архитектуры на основе трансформеров, демонстрируют высокие эмпирические показатели. Однако они нередко характеризуются недостаточной лингвистической прозрачностью и сталкиваются с трудностями систематического обобщения в условиях ограниченных ресурсов. В данной работе предлагается сравнительное и интегративное исследование формальных (правил-ориентированных и конечных автоматов) и нейронных подходов к морфологическому анализу на примере казахского языка как агглютинативного языка с низкими ресурсами. Вначале представляется формальная морфологическая модель, явно описывающая структуру «корень–аффикс», сингармонизм и морфотактические ограничения. Далее исследуются различные нейронные архитектуры для морфологической дизамбигуации и разметки, включая KazBERT в сочетании с декодированием на основе условных случайных полей (CRF). Помимо стандартных метрик точности проводится детальная типология ошибок и лингвистический анализ, в рамках которых изучается, как различные классы моделей справляются с неоднозначностью, редкими словоформами и длинными цепочками аффиксов. Результаты показывают, что, хотя нейронные модели превосходят исключительно правил-ориентированные системы по поверхностной точности, они демонстрируют устойчивые недостатки при обработке морфологически сложных и редких конструкций. Формальные модели, в свою очередь, обеспечивают более надёжное обобщение за счёт языковых ограничений. На основе полученных результатов предлагается гибридная морфологически осведомлённая архитектура, интегрирующая символические ограничения в нейронный вывод. Данный подход обеспечивает устойчивое улучшение результатов в различных условиях оценки. Работа демонстрирует, что эффективный морфологический анализ агглютинативных языков требует сочетания нейронного обучения представлений и явного лингвистического моделирования. Полученные выводы не *ограничиваются одним языком и имеют более широкие последствия для морфологически чувствительного NLP в низкоресурсных условиях.*

Ключевые слова: *вычислительная морфология, агглютинативные языки, морфологический анализ, языки с ограниченными ресурсами, формальные лингвистические модели, нейронные языковые модели.*

Information about the authors

Aigerim Aitim – PhD, associate-professor of Information Systems Department, International Information Technology University, Almaty, Kazakhstan; e-mail: a.aitim@iitu.edu.kz. ORCID: https://orcid.org/0000-0003-2982-214X.

Сведения об авторах

Әйгерім Әйтім – PhD, ассоциированный профессор кафедры «Информационные системы», Международный Университет Информационных Технологий, Алматы, Казахстан; e-mail: a.aitim@iitu.edu.kz. ORCID: https://orcid.org/0000-0003-2982-214X.

Авторлар туралы мәліметтер

Әйгерім Әйтім – PhD, «Ақпараттық жүйелер» кафедрасының қауымдастырылған профессоры, Халықаралық Ақпараттық Технологиялар Университеті, Алматы, Қазақстан; e-mail: a.aitim@iitu.edu.kz, ORCID: https://orcid.org/0000-0003-2982-214X.

Received 30.03.2026
Revised 20.05.2026
Accepted 21.05.2026

https://doi.org/10.53360/2788-7995-2026-2(22)-15

ҒТАХР: 50.41.25, 50.41.29

Ж.Н. Ерғазы*, Қ. Әлімхан, Н.С. Мукатаев

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті,
010000, Қазақстан Республикасы, Астана қ., Сәтпаев көшесі, 2
*e-mail: yergaziyevaaa@gmail.com

СЫЗЫҚТЫ ЕМЕС СЫРТҚЫ ӘСЕРЛЕРІ БАР ЖҮЙЕЛЕРДІ ШЫҒЫС КЕРІ БАЙЛАНЫС АРҚЫЛЫ БЕКІТІЛГЕН УАҚЫТ ІШІНДЕ КЕҢАУҚЫМДЫ ОРНЫҚТАНДЫРУ

Аңдатпа: *Екінші ретті сызықты емес динамикалық жүйелер үшін шығыс бойынша кері байланыс негізінде бекітілген уақыттағы орнықтандыру тұжырымдамасына сүйенген басқару алгоритмін синтездеудің жаңа әдісі ұсынылған. Зерттеу күй айнымалыларына толық қолжетімділік жоқ кеңауқымды жазық сызықты емес объектілер класына бағытталған. Ұсынылған әдістеменің негізін би-шектік біртектілік теориясы мен орнықтандыруды классикалық Ляпуновтық талдау қағидаларын біріктіру құрайды. Осындай тәсіл бастапқы шарттар мен бағалау қателеріне тәуелсіз жүйенің өлшенбейтін күйін кепілді түрде бағалайтын, жинақталу уақыты бекітілген үздіксіз бақылаушыны жасауға мүмкіндік берді. Осы бағалау негізінде жүйені алдын ала берілген бекітілген уақытта орнықтандыруды қамтамасыз ететін үздіксіз реттегіш құрастырылды.*

Әдіс бірқатар артықшылықтарға ие: орнықтандыру мен жинақталу бастапқы шарттардың шамасы мен таңбасына тәуелді емес; басқару сигналы үздіксіз болып қалады, бұл атқарушы құрылғылардың дірілін болдырмайды; шектелген сыртқы әсерлер мен өлшеу шударына жоғары робасттылық қамтамасыз етіледі. Алгоритмді жоғары жиілікті дискретизацияны талап етпей-ақ микроконтроллерлерде жүзеге асыруға болады, бұл оны практикалық қолдану үшін тартымды етеді. Мысал ретінде айқын сызықты емес электромеханикалық объект болып табылатын микромеханикалық айна (MEMS mirror) динамикасы қарастырылған. Сандық модельдеу жүргізіліп, оның нәтижелері ұсынылған басқару тиімділігін растайды. Қолданыстағы шекті уақыттағы реттегіштермен салыстырғанда өтпелі үдеріс уақыты төрт еседен артық қысқарған, ал жүйенің ауытқуы мен энергия шығындары шамамен екі есеге азайған. Алынған нәтижелер