IRSTI 20.15.05 https://doi.org/10.26577/jpcsit4220264

Zhansaya Segizbayeva[1*], Marek Miłosz[2]

[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2]Lublin University of Technology, Lublin, Poland
*e-mail: segizbayeva.zhansaya@gmail.com

COMPUTATIONAL MODEL OF KAZAKH VOWEL–CONSONANT HARMONY

Abstract. This study presents a computational model of morphological generation based on the agglutinative nature of the Kazakh language and vowel–consonant harmony. In this work, we model the Kazakh vowel–consonant harmony system, which governs allomorphic selection of suffixes based on both vowel and consonant properties. Using a rule based approach, the model generates verb and noun forms across various grammatical categories (tense, aspect, case, possession). Experimental results demonstrate the model's high efficiency and full compliance of the generated word forms with vowel harmony rules. Specifically, thousands of combinations were created for derivative words (DTL), possession (POSS1, POSS2), and case (CASE) categories. This research makes a significant contribution to the development of natural language processing (NLP) morphological analysis and automatic translation systems for the Kazakh language.

Keywords: kazakh language, computational linguistics, morphological generation, vowel–consonant harmony, harmony law, agglutinative languages, and natural language processing.

1. Introduction

The Kazakh language is an agglutinative language belonging to a diverse language family, characterized by a rich morphological structure and a complex sound system. The main sound system is the law of harmony, ensuring vowel harmony within a single syllable. This rule underlies word formation, controlling the correct connection of suffixes to the root [1], [2]. The rapid development of digital technologies has increased the importance of studying Kazakh natural language processing (NLP) performance. However, these features of the Kazakh language do not facilitate the ready use of natural language processing models, that is, special computers are needed to understand the language's structure [3], [4]. Computational modeling of syllable and sound harmony is key for developing practical tools such as morphological analysis, filters, automatic translations, and text editors. Early studies relied on rule-based approaches, including finite state automata and two-level morphology [1]. While these models describe morphological regularities, creating them requires much work. Recent data indicate that neural systems and statistical methods in hybrid models are well-developed. Such models can effectively handle deviations from the language norm, but require a large database to achieve high-quality results. Still, creating generative models that preserve the regularities of syllable andsound harmony in Kazakh is an urgent problem.

The study aims to develop a computational model for Kazakh morphological rules and regularities of syllable and sound harmony and experimentally assess its effectiveness.

The study's objectives are: 1) Develop an algorithm for forming Kazakh word class forms (nouns, verbs, pronouns, prepositions, moods). 2) Create rules considering syllable types, sound harmony, and inflectional differences, including solid-soft and labial variations. 3) Conduct generative-numerical analysis for various grammatical categories. 4) Identify the model's strengths and weaknesses based on results.

Vowel–consonant harmony in the Kazakh language is a complex phonological process that regulates the harmony of vowels, and sometimes consonants, in a word. This is a feature characteristic of Turkic languages, where all vowels in a word usually harmonize with the first vowel of the root. Such harmony ensures the phonetic integrity of the language and plays an important role in the agglutinative structure (i.e., the system of conveying grammatical meaning through affixes). McCollum (2018) examined the gradual and local nature of consonant harmony in Turkic languages and roposed a variance analysis utilizing the "Maximum Entropy Harmonic Grammar" [5].

There are two main types of harmony patterns in the Kazakh language: palatal harmony (solid-soft harmony) and labial harmony (lip harmony). According to the law of palatal harmony, all vowels in a word must be solid (/'a', 'o', ' ū', 'y', 'u'/) or thin (/ä', ' ö', ' ü', 'e', 'i'/). There is also lip sync, but it is considered secondary and regulates the correspondence of the lip features of vowels.

Zhusupov and Saparova (2024) divide syllable and sound sync in Turkic languages into four types: singarmolinguohard, singarmolinguosoft, singarmolinguolabiohard, and singarmolinguolabiosoft, which indicates that it is a universal prosodic feature [6]. They conclude that the phonostylistics of the Kazakh language is mainly singarmolinguo, and labial syncharmonism is weak.

They compare this to non-syntactic languages, such as Russian, and show that, in addition to vowels, consonant harmony also plays an important role in the morphophonemic structure of the Kazakh language. This includes phenomena such as progressive and regressive assimilation, in which consonants adapt to neighboring sounds and adopt their articulatory properties. Such consonant consonance patterns impose strict morphophonemic constraints on the combination of roots and suffixes in addition to vowel consonance [7].

The choice of a particular affix form is often determined by these phonological rules, i.e., it corresponds to the sounds preceding the affixes to which it is attached. If the root contains thick and thin vowels, then the vowel in the final syllable determines the type of consonant. These patterns are often irregular and do not fully obey the law of consonance.

The complexity of the Kazakh language's syllabic and phonetic harmony—its agglutinative nature, the coexistence of several types of harmony, and the processes of mutual assimilation of vowels and consonants—poses significant challenges for computational linguistics. Therefore, any effective computational model for the Kazakh language must not only accurately reflect these complex phonological patterns, but also take into account the difficulties arising from the writing system.

The first and most fundamental approaches to computationally modeling vowel–consonant harmony in Kazakh are based on rule-based systems. Boundary state modifiers (BSTs) and two-level morphology are often used. Such models are especially useful for agglutinative languages such as Kazakh, since morphological processes can be systematically and accurately described by boundary rules and state transitions.

Tukeyev (2015) used finite-state automata-based models to formally describe the morphological structure of the Kazakh language [8]. The study examined the problems of morphological analysis of words and checking the completeness of conjugations using an automatic approach, and proposed effective ways to formalize morphological processes characteristic of agglutinative languages.

The work of Kessikbayeva and Chichekli (2014) [9] is based on two-level morphology implemented using Xerox boundary state tools. This analyzer describes in detail the patterns of suffix change and suggests rules for using different suffix forms according to compatibility. Similarly, Kairaqbay and Zaurbekov (2013) proposed a boundary state-based approach to analyzing the noun paradigm in Kazakh [10]. Their model takes into account the morphophonemic constraints (vowel-consonant and voiceless patterns) specific to the Kazakh language.

Due to the limitations of rule-based systems, data-driven and hybrid approaches are currently being actively developed. These models combine the accuracy of rule-based methods with the flexibility offered by statistical and machine learning techniques.

Zhetkenbai et al. (2025) proposed a hybrid computational model for formatting and predicting the morphological inflection of Kazakh nouns based on the new Latin alphabet [11]. This model integrates rule-based grammatical formatting with a Bayesian regularized backpropagation neural network (BR-BPNN). The BR-BPNN model showed high accuracy (91.5%) and specificity (89.4%).

Stoyer et al. (2023) conducted a study to describe vowel harmony from an information theory perspective [12]. They suggested an approach based on neural recurrent language models (LSTMs).

Overall, the literature emphasizes the critical role of rule-based approaches (FST, two-level morphology) in accurately modeling synchronization for Kazakh morphological production, alongside the growing importance of hybrid and neural models (BR-BPNN, Transformer, LLM) in enhancing flexibility and accuracy. The current work continues this tradition by presenting a precise computational model for syllable and sound harmony generation.

2. Materials and methods

The proposed computational model uses a rule-based approach to implement the morphological formation of the Kazakh language. This approach is based on the regularity of morphological processes in agglutinative languages and formally implements the generative aspect of the concept of finite-state transducers (FST) [1] [3]. The primary objective of the model is to generate the correct word form for a given root and set of grammatical categories, strictly adhering to the laws of syllable and sound harmony.

The architecture of the model is based on the interaction of the following main components:

- Phonetic-lexical resources: Phonetic groups based on the classification of vowels and consonants and the intralinguistic features of the root, for example, the thickness/thickness of the root.

- Morphotactic rules: A set of rules that determine the order and permissible combinations of roots and suffixes.

- Syllable and sound harmony control logic: Conditional logical operators that control the choice of a solid/soft or labial/non-labial version of a suffix. This logic depends on the characteristics of the vowel in the final syllable of the root.

This structure allows for computationally accurate and efficient modeling of the morphophonemic processes of the Kazakh language.

Below are the subsets of letters that play an important role in the rules of vowel harmony of words with nominal roots [13].

vsolid = [‘a’, ‘o’, ‘ ū’, ‘y’, ‘u’]; vsoft = [‘ ä’, ‘ ö’, ‘ ü’, ‘i’, ‘e’, ‘u’, ‘i’]; cdeaf = [‘p’, ‘k’, ‘q’, ‘t’, ‘s’, ‘ ş’, ‘h’]; cvoiced = [‘b’, ‘v’, ‘ g’, ‘ ğ’, ‘d’, ‘j’, ‘z’]; cv1 = [‘b’, ‘v’, ‘g’, ‘ ğ’, ‘d’]; cv2 = [‘j’, ‘z’]; cvoiced1 = [‘b’, ‘v’, ‘g’, ‘d’]; csonor = [‘l’, ‘m’, ‘n’, ‘ ñ’, ‘r’, ‘ı’, ‘u’]; cs1 = [‘m’, ‘n’, ‘ñ’]; cs2 = [‘l’, ‘r’, ‘ı’, ‘u’]; cs3 = [‘l’, ‘m’, ‘n’, ‘ñ’]; cs4 = [‘r’, ‘ı’, ‘u’].

Table 1
Rules for forming words with plural endings

LAR/LER rules	TAR/TER rules
if stem[-1] in (vsolid or vsoft or cs4):	if stem[-1] in (cdeaf or cv1):
if stem[-1] in vsolid and ending[1] in vsolid:	if stem[-2] in vsolid and ending[1] in vsolid:
if ending[:3] == ‘lar’:	if ending[:3] == ‘tar’:
endingtl = ending	endingtl = ending
#print(‘11.a solid’)	#print(‘11.a solid’)
if stem[-1] in vsoft and ending[1] in vsoft:	if stem[-2] in vsoft and ending[1] in vsoft:
if ending[:3] == ‘ler’:	if ending[:3] == ‘ter’:
endingtl = ending	endingtl = ending

Table 1. A brief explanation of the rules for the plural endings (-lar/–ler, –tar/–ter). This table provides phonetic rules that automate the correct selection of plural endings. The rules are based on the final sound of the word and the vowel harmony [16].

– LAR / –LER

If the word ends in a vowel or a voiced/voiced consonant: The last sound is thick → lar.

The last sound is thin → ler. The code checks the consonance of the ending using ending[:3].

– TAR / –TER

If the word ends in a hard consonant: The previous sound is thick → tar. The previous sound is thin → ter. This rule is fixed by the conditions ending[:3] == ‘tar’/’ter’. The algorithm automatically selects the correct form of the plural ending, taking into account the law of consonance in the Kazakh language.

These 2 tables provide consonant rules that determine the correct version of the possessive conjunction when the final sound of the word is a vowel. The algorithm makes a choice taking into account the thickness or thinness of the final and preceding sounds of the root, as well as the presence of the last letter “u”. The algorithm checks the correspondence of the possessive conjunction (1st person, 2 nd person singular/plural) to the thickness or thinness of the consonant and the final sound of the root, and automatically selects the correct conjunction.

This rule automatically generates classified forms of the present tense depending on whether the vowel in the final syllable of the verb stem is thick or thin.

Table 2
Rules for forming words with possessive endings

Rules for possessive endings (the last letter is a vowel)

if stem[-1] in vsolid and ending[1] in vsolid and stem[-2] in vsolid and stem[-1] == ‘u’:
if ending == ‘m’ or ending == ‘ń’ or ending[:3] in (‘ńyz’, ‘myz’) or (ending[:2] == ‘sy’):
endingtl = ending
if stem[-1] in vsolid and ending[1] in vsolid:
if ending == ‘m’ or ending == ‘ń’ or ending[:3] in (‘ńyz’, ‘myz’) or (ending[:2] == ‘sy’):
endingtl = ending
if (stem[-1] in vsoft and ending[1] in vsoft and stem[-2] in vsoft and stem[-1] == ‘u’):
if ending == ‘m’ or ending == ‘ń’ or ending[:3] in (‘ńiz’, ‘miz’) or (ending[:2] == ‘si’):
endingtl = ending
if (stem[-1] in vsoft and ending[1] in vsoft and stem[-2] in vsoft and stem[-1] == ‘u’):
if ending == ‘m’ or ending == ‘ń’ or ending[:3] in (‘ńiz’, ‘miz’) or (ending[:2] == ‘si’):
endingtl = ending

Table 3
Verb (classified forms). Rule of the present tense (Present Simple)

Verb (classified forms). Rules for the present tense (Present Simple)
if stem[-2] in vsolid: for ending in pres_spl_sg_vsolid: forms.append(stem + ending) for ending in pres_spl_pl_vsolid: forms.append(stem + ending) if stem[-2] in vsoft: for ending in pres_spl_sg_vsoft: forms.append(stem + ending) for ending in pres_spl_pl_vsoft: forms.append(stem + ending)

If the previous sound of the root is solid (vsolid):

– The list of suffixes for solid persons (pres_spl_sg_vsolid and pres_spl_pl_vsolid) is added, and all classified forms are generated.

If the previous sound is soft (vsoft):

– The list of suffixes for soft persons (pres_spl_sg_vsoft and pres_spl_pl_vsoft) is removed, and all forms are generated.

This rule automatically selects the pronoun forms –ğan/–gen/–qan/–кen, as well as the forms –atyn/–etyn, –ytyn/–ytyn, based on the final sounds of the root.

Table 4
Participle (PP) rules

Participle (PP). -ĞAN/-GEN/-QAN/-KEN, -ATYN/-ETIN, -ITYN/-ITIN rules

if stem[-2] in vsolid and stem[-1] in csonor and ending[:3] == ‘gan’:
endingtl = ending
if stem[-2] in vsoft and stem[-1] in csonor and ending[:3] == ‘gen’:
endingtl = ending
if stem[-2] in vsolid and stem[-1] in cdeaf and ending[:3] == ‘qan’:
endingtl = ending
if stem[-2] in vsoft and stem[-1] in cdeaf and ending[:3] == ‘ken’:
endingtl = ending
if stem[-2] in vsolid and stem[-1] in csonor and ending[:4] == ‘atyn’:
endingtl = ending
if stem[-2] in vsoft and stem[-1] in csonor and ending[:4] == ‘etin’:
endingtl = ending
if stem[-1] in vsolid and ending[:4] == ‘ıtyn’:
endingtl = ending
if stem[-1] in vsoft and ending[:4] == ‘ıtın’:
endingtl = ending

–GAN / –GEN / –QAN / –KEN
Previous sound thick + last sound soft → –ğan
Previous sound thin + last sound soft → –gen
Previous sound thick + last sound hard → –qan
Previous sound thin + last sound hard → –ken
–ATYN / –ETYN
Last sound soft, root thick → –atyn
Last sound soft, root thin → –etyn
–ITYN / –ITYN
Last sound thick → –ıtyn
Last sound thin → –ıtyn

This rule ensures the automatic correct selection of the main adverb types (–a/–e/–ı, –yp/–ıp/–p, –qaly/–ğaly/–кelı/–gelı) based on the properties of the final sounds of the root.

Table 5
Verbal Adverb (VA) rules

(Verbal Adverb, VA). -A/-E/-I, -YP/-IP/-P, -QALY/-ĞALY/-KELI/-GELI

if (stem[-2] in vsolid or stem[-3] in vsolid) and stem[-1] in cdeaf \
 and ending[:1] == 'a':
 endingtl = ending
 if stem[-2] in vsoft and stem[-1] in csonor and ending[:1] == 'e':
 endingtl = ending
 if stem[-1] in vsoft and ending[:1] == 'ı':
 endingtl = ending
 if (stem[-2] in vsolid or stem[-3] in vsolid) and stem[-1] in cdeaf \
 and ending[:2] == 'yp':
 endingtl = ending
 if stem[-2] in vsoft and stem[-1] in csonor and ending[:2] == 'ıp':
 endingtl = ending
 if stem[-1] in vsoft and ending[:1] == 'p':
 endingtl = ending
 if (stem[-2] in vsolid or stem[-3] in vsolid) and stem[-1] in cdeaf \
 and ending[:4] == 'qaly':
 endingtl = ending
 if stem[-2] in vsolid and stem[-1] in cvoiced and ending[:4] == 'galy':
 endingtl = ending
 if stem[-2] in vsoft and stem[-1] in cdeaf and ending[:4] == 'kelı':
 endingtl = ending
 if stem[-2] in vsoft and stem[-1] in csonor and ending[:4] == 'geli':
 endingtl = ending

–A / –E / –I
Root thick + final sound hard → –a
Root thin + final sound soft → –e
Final sound soft → –ı
–YP / –IP / –P
Thick consonant + hard consonant → –yp
Thick consonant + soft consonant → –ıp
Final sound soft → –p
–QALY / –ĞALY / –KELI / –GELI
Thick + hard → –qaly
Thick + soft → –ğaly
Thick + hard → –kelı
Thick + soft → –gelı

The rule analyzes the thick-thin, hard-soft, soft properties of the root and automatically selects the correct ending for the preposition.

These rules provide automatic selection of the correct form of the imperative and conditional mood endings depending on whether the vowel in the final syllable of the verb stem is thick or soft.

Table 6
Rules of harmony of the imperative and conditional moods

Rules of harmony between the imperative and the conditional mood

if stem[-2] in vsolid and ending == 'syn':
endingtl = ending
if stem[-2] in vsoft and ending == 'sın':
endingtl = ending
if stem[-2] in vsolid and ending == 'şy':
endingtl = ending
if stem[-2] in vsoft and ending == 'şı':
endingtl = ending
if stem[-2] in vsolid and ending == 'sa':
endingtl = ending
if stem[-2] in vsoft and ending == 'se':
endingtl = ending

Imperative case **(-syn / -sın)**
If the root is thick → **–syn**
If the root is thin → **–сін**
Imperative case -şy / -şy form
Root thick → –şy
Root thin → –şy

Conditional case **(-sa / -se)**
Root thick → **–sa**
Root thin → **–se**
Similar rules for calculating vowel harmony have been compiled for other parts of speech in the Kazakh language.

3. Results

To assess the derivational potential of the proposed computational model, a quantitative morphological analysis of word forms derived from the same root across the main word classes (noun, verb, preposition, adverb, and pronoun) was conducted to demonstrate the agglutinative-polysynthetic nature of the Kazakh language. This analysis allows us to determine the linguistic metrics of derivational productivity.

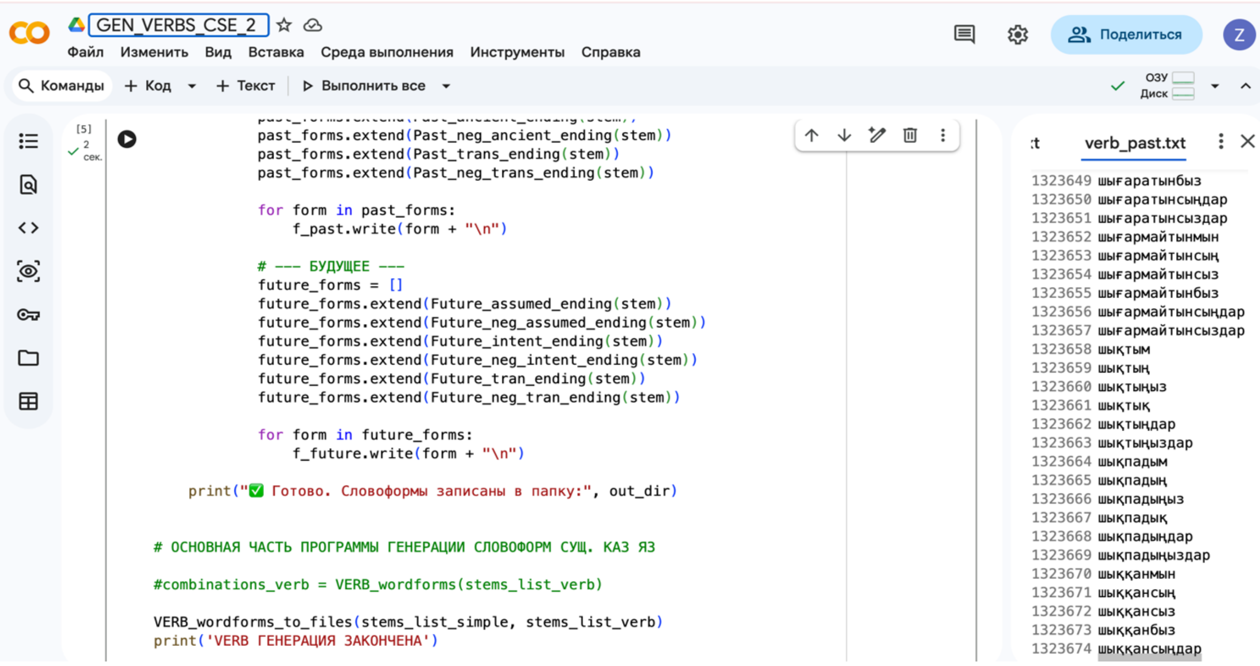
Table 7 shows the quantitative distribution of word forms across the morphological paradigms of the main word classes. All variations of syllable and sound harmony were taken into account in the derivational process.

Table 7
Quantitative indicators of word forms generated from the root word

№	Word classes	(stem)	Type of connections used	Number of generated forms
1	Noun	108	plurality (larp/ler, tar/ter), dependency (-m, -ń, -sy/-sı), accusative (-nyń, -dy, -ğa, -men ...)	21709
2	Verb	108	tense forms (present, past, future), negative, participles	2 307667
3	Voice Verbal Adverb, VA Moods	100	self, otherness, involuntary, common -a/-e/-ı, -yp/-ıp/-p, -ğaly/-kelı	844
4	Participle, PP	100	-ğan/-gen/-qan/-ken, -atyn/-etın, -uşy/-uşı, -ar/-er	17141
5	Word classes	416	-	2 347361

The results of morphological generation for six word classes were calculated using a model based on the given rules, and the results were obtained. Using the rules for each word class, all possible word forms were calculated fully automatically.

Figure 1
Number of verb endings (past tense)



As shown in Figure 1, the results obtained were 21,709 forms for nouns, 2,307,667 forms for verbs, 17,141 forms for pronouns, and 844 forms for prepositions and adverbs. A total of 2,347,3611 forms were obtained for the given word classes. This quantitatively proves the agglutinative nature of the Kazakh language and demonstrates the model's generation ability.

4. Discussion

The proposed rule-based computational model has shown high potential for fully preserving the morphophonemic processes and vowel–consonant harmony patterns of the Kazakh language in a generative paradigm. As described in the methodological formalization, the precise classification of phonetic elements and the conditional logical rules based on this classification allow for high-accuracy computational modeling of palatal vowel–consonant harmony and regressive assimilation of consonants. This approach proves the effectiveness of directly translating linguistic rules into computational algorithms [14].

Advantages of the model:

1. Formal verification and interpretation: The main advantage of the rule-based approach is its high accuracy and interpretability. Each morphological generation process can be fully verified by applying linguistic rules. This property is very important for empirical verification of theoretical conclusions in linguistic research.

2. Data independence efficiency: The fact that the model does not require a large annotated data corpus to work makes it the most effective initial approach for rapidly developing NLP tools for low-resource languages such as Kazakh.

3. Morphological space representation: As shown in the results, it is proven that the model can fully and systematically represent the agglutinative-polysynthetic morphological space of the Kazakh language by generating 2,307,667 verb forms from a lexically given root.

As mentioned in the literature review, approaches such as rule-based FST models [1], [3] and hybrid BR-BPNN [2] are used to model vowel–consonant harmony in the Kazakh language. The proposed model is similar to FST, but is focused on generation and formalizes rules using simple conditional logic.

Comparative analysis with other computational paradigms:

- With finite automaton transformers (FST): The proposed model, while having the generative capabilities of FST, implements its rules using simple conditional logic operators. This approach reduces the complexity of building an FST and increases the speed of model development.

- With neural and hybrid models: Neural networks (Transformer, BR-BPNN) show flexibility in handlin large data-driven fluctuations and language variations [2], [15]. However, the main theoretical limitation of our model is that it cannot fully capture deviations from the law of voicelessness and secondary, sometimes unstable, manifestations of vowel–consonant harmony in imported words (e.g., “computer”, “institute”). To address these issues, future research should move to a hybrid architecture by incorporating lexical features (e.g., word origins) and statistical checks into the model.

The scientific significance of this study lies in the deep formalization of the morphological syllable of the Kazakh language from a computational perspective. Quantitative analysis of syllable productivity objectively proves the morphological richness of the Kazakh language and can serve as a basis for future NLP research [15], [16].

Practical applications:

- Morphological synthesizers: The model can be used to generate correct word forms in speech systems and machine translation systems at the syllabic stage.

- Educational tools: This will be the basis for creating interactive grammar simulators for Kazakh language learners.

- Spell check tools: A large corpus of syllabic forms will help create dictionaries for checking the correct spelling of words.

5. Conclusion

This study presents a rule-based paradigm for the computational formalization of the morphological syllable of the Kazakh language. The proposed method systematically characterizes the agglutinative nature of the Kazakh language, including vowel harmony, consonant assimilation, and the multilayered structure of word formation and word inflection.

The main scientific novelty of the research is the development of a computational model capable of automatically generating hundreds of new word forms with syntactic and semantic meanings from a single lexical root in accordance with morphological laws. The model fully covers the main morphological categories, such as plural, dependent, verb, pronoun, preposition, imperative, conditional, and present tense forms.

The experimental results show that it is possible to generate morphological paradigms of nouns and verbs in a complete and error-free manner. This not only quantitatively proves the morphological richness of the Kazakh language but also opens up new opportunities for automating linguistic resources.

Thus, the morphological generation model, which combines the structural features of the Kazakh language with computational methods, lays the theoretical foundation for creating intelligent linguistic systems for agglutinative languages and expands the possibilities of integration with machine learning and neural models in the future.

Acknowledgements

The authors would like to acknowledge Ualsher Tukeyev for his valuable academic discussions, methodological feedback, and expert guidance provided during the preparation and revision of this manuscript.

Funding

This research was supported by the Ministry of Science and Higher Education of the Republic of Kazakhstan within the framework of the scientific project “Study of neural models for the formation of speech transcripts and minutes of meetings in Turkic languages” (Grant No. AP23487816).

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, M.M. and Z.S.; Methodology, M.M.; Software, M.M. and Z.S.; Validation, M.M. and Z.S.; Formal Analysis, M.M.; Investigation, M.M. and Z.S.; Resources, Z.S.; Data Curation, Z.S.; Writing – Original Draft Preparation, M.M. and Z.S.; Writing – Review & Editing, M.M.; Visualization, Z.S.

References

1. O. Özchelik, Phonetics, Phonology, Phonotactics, Morphophonology, Prosody. Published in the Encyclopedia of Turkic Languages. Brill.

2. G. B.Shoibekova, S. A.Odanova, B. M.Sultanova, & T. N. Ermekova, “Vowel harmony is the main phonetic law characteristic of Turkic languages.” *International Journal of Environmental & Science Education*, 2016, vol. 11, no. 11, pp. 4617–4630.

3. U. Tukeyev, “A new computational model for Turkic languages’ morphology and processing.” *Journal of Problem in Computer Science and Information Technologies*, vol 1, no 1, pp. 47-54, 2023. https://doi.org/10.26577/JPCSIT.2023.v1.i1.07

4. Ya. Karabaliev, K. Kolesnikova, “Speech recognition methods in the Kazakh language: analysis of errors and possibilities of improvement.” *Scientific journal of Astana IT University*, vol. 20, pp. 62–74, 2024. doi.org/10.37943/20dzgh8448

5. A. McCollum, “Vowel dispersion and Kazakh labial harmony.” *Phonology*, vol. 35, no 2, pp. 239-276, 2018. doi.org/10.1017/S0952675718000052

6. M. Dzhusupov, N. Saparova, “Comparative study of synharmonic and non-syndharmonic phonostylistics (on the material of Kazakh and Russian languages).” *Journal of Linguistics, Semiotics and Semantics of the Russian State University of Technology,* vol 15, no 3, pp. 895–913, 2024. doi.org/10.22363/2313-2299-2024-15-3-895-913

7. J. Dees, “Heterogeneity in the phonologization period: data from the Kazakh language.” *Proceedings of the Workshop on Turkic and Languages in Contact with Turkic,* vol. 8, pp. 74–88, 2023. doi.org/10.3765/6wbxsg62

8. U. Tukeyev, Automaton models of the morphology analysis and the completeness of the endings of the Kazakh language, Proc. TurkLang-2015, Kazan, Russia. Sep. 17-19, 2015, pp. 91–100.

9. G. Kessikbayeva, I. Cicekli, “Rule based morphological analyzer of Kazakh language.” *InProceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM*, 2014, pp. 46-54.

10. B. Kairakbay, and D. Zaurbekov. “Fsmnlp 2013-proceedings of the 11th international conference on finite state methods and natural language processing.” 2013.

11. *L. Zhetkenbay, A. Sharipbay, B. Razakhova, G. Bekmanova, A. Barlybayev, A. Nazyrova, B. Yergesh,* “Formalization of Morphological Rules for Kazakh Nouns in the New Latin Alphabet,” *Journal of Applied Data Sciences*, 2025. vol. 6, no 3, pp. 1999-2019.

12. J. Steuer, J. List, B. Abdullah, D. Klakow, “An Information-Theoretic Characterization of Vowel Harmony.” *Transactions of the Association for Computational Linguistics*. pp. 96–109, 2023. doi:10.18653/v1/2023.sigtyp-1.10

13. U. Tukeyev. “Computational models of morphology and vowel harmony of the Kazakh language and their use for neural models.” M A T E R I A L S of The International Scientific and Theoretical Conference ARTIFICIAL INTELLIGENCE IN THE SPACE OF CONTEMPORARY ART: PROBLEMS AND PROSPECTS, 11.04.2025. Baku, 2025, pp. 12-19. DOI:10.25045/ASUCAAI.2025.01

14. 14U.Tukeyev, A. Karibayeva, Zh. Zhumanov, (2020). “Morphological Segmentation Method for Turkic Language Neural-Machine Translation.” *Cogent Engineering,* vol. 7, 2020, doi.org/10.1080/23311916.2020.1856500

15. A. Aytim, “Development of methods for automatic processing systems of the Kazakh language.” *KazATK Khabarshsy*, 2024, *vol. 133, no 4, pp.* 254–265, doi.org/10.52167/1609-1817-2024-133-4-254-265

16. S. Altynbek, “The need to reform the Kazakh writing system.” *Educational Science,* vol. 1, no 1, pp. 1–5, 2024. doi:10.61927/igmin148

Information about the authors:

Zhansaya Eldibayevna Segizbayeva – Doctoral Candidate, Al-Farabi Kazakh National University (Almaty, Kazakhstan, e-mail: segizbayeva.zhansaya@gmail.com. ORCID: https://orcid.org/0009-0002-1324-9086).

Marek Milosz – Doctor of Philosophy (English), Professor; Lublin University of Technology (Lublin, Poland, m.milosz@pollub.pl. ORCID: https://orcid.org/0000-0002-5898-815X).

Submission received: 15 January, 2026
Revised: 11 March, 2026
Accepted: 19 March, 2026