From Cyrillic to Latin: A Step Toward Phonemic Precision in Kazakh

Sadyk Didar – Kyzdarkhan Rysbergen – Schochenmaier Eugen – Bazarbayeva Zeinep

DOI: 10.18355/XL.2026.19.02.18

Abstract
This study reassesses Kazakhstan's 2021 Latin script design against the 1940 Cyrillic through a phoneme-to-grapheme correspondence (PGC) lens. We formalize an inventory-level estimator (PGC-Core) and evaluate three scenarios: Minimalist (Core phonemes), Standard (Core + Integrated loan segments), and an Extended set reported for coverage (Marginals documented but excluded from the mean). The Latin design exhibits consistently higher correspondence, reflecting a nine-vowel analysis, clearer treatment of glides, and a consistent diacritic system. We review earlier Latinization drafts and resolve contested cases (e.g., /j/ as a phoneme with context-conditioned mapping; loan affricates via a fixed digraph). Methodologically, the paper couples a transparent scoring scheme (A1–A3) with qualitative analysis of vowel polysemy, graphemic multifunctionality, and the management of dual-script coexistence. Status flags (Core/Integrated/Marginal) and a single formula specify what is averaged and what is excluded, ensuring replicability. A comparative section situates Kazakhstan among other reforms – especially in the Turkic sphere – distilling policy levers that correlate with smoother adoption. Beyond linguistic fit, script choice carries digital consequences. We outline practical measures—time-boxed coexistence, clear style guidance, funded teacher re-tooling, conversion pipelines, and unified ICT standards for input, hyphenation, OCR/ASR, and cross-script retrieval—to translate orthographic clarity into lower learning and production costs. Overall, the evidence supports Latin as the more faithful and operationally scalable orthography for contemporary Kazakh.
Key words: Phoneme-to-Grapheme Correspondence (PGC), IPA, Kazakh language, orthography, phonemic transparency, phonotactics

Introduction

The Kazakh language, a member of the Kipchak branch of the Turkic language family, underwent multiple script transitions over the past century, reflecting sociopolitical shifts and linguistic reforms. Historically, Kazakh was written in the Arabic script, which, despite some adaptation to Turkic phonology, lacked sufficient vowel representation to accurately convey Kazakh phonemes. In 1929, the Soviet Union formally introduced a unified Latin-based script known as *Yañalif* (Jaŋa Älifba, "New Alphabet") as part of its campaign to Latinize the scripts of non-Slavic languages. Although this policy is often described as a modernization initiative aimed at improving literacy and reducing Arabic influence in the Turkic-speaking regions of Central Asia, its scope was in fact broader. The Latinization process was applied not only to Muslim Turkic populations but also to non-Muslim and non-Turkic groups, including Yakut, Khakas, Altay, Buryat, Itelmen, and Ostyak, many of whom had little to no prior exposure to Arabic-script literacy (Baldauf, 1993). Furthermore, plans for Latin-based orthographies had been circulating in Azerbaijani intellectual circles since the late 19th century and were implemented in print as early as 1922 (Amirjanova, 2016: 14-15), making the Soviet reform part of a longer continuum of script modernization.

In the Kazakh context, the local variant of Yañalif was sometimes referred to as *Çaŋälip* (Çaңәліп), though the core principles remained consistent with the broader Soviet Latinization framework. The Kazakh adaptation included phoneme-grapheme

modifications specific to Kazakh phonology, but the umbrella term Yañalif continues to be used in scholarly literature to describe this period of Latin-script adoption across the Turkic-speaking republics (Amirjanova, 2012: 25). However, this Latin-based system was short-lived; in 1940, the Cyrillic alphabet was imposed, aligning Kazakh orthography with Russian linguistic norms. The adoption of the Cyrillic script introduced several phonemic inconsistencies into the Kazakh writing system. These inconsistencies increased orthographic complexity, making literacy acquisition more challenging and creating discrepancies between written and spoken Kazakh.

In 2017, Kazakhstan announced a national transition from Cyrillic to a Latin-based script, aligning with modernization efforts and broader linguistic reforms among Turkic-speaking nations. While the transition to the Latin script in Kazakhstan has followed a multi-phase implementation process, several linguistic and typographic challenges have emerged along the way. As Aqtay (2020: 23) notes, the initial version introduced in 2017 employed apostrophes to mark specific phonemes (e.g., a', o', i'), drawing inspiration from Baıtursynuly's use of the *dáıekshe*. However, in practice, this design led to severe phonetic ambiguity, digital incompatibility, and disruption of Kazakh morphophonemic regularity. For instance, forms like *jıyn* were visually and phonologically confusing when compared to their morphologically transparent equivalents such as *jyıyn*.

In response, the 2018 revision eliminated apostrophes in favor of diacritics, ostensibly improving digital usability. Nevertheless, the revised version retained deeper structural issues inherited from the Cyrillic system. Specifically, vowel-approximant sequences like *iy*, *uw*, and *jy* remained inconsistently represented. This created pseudo-diphthongs and morphophonemic confusion, with examples like *tıyn* (intended as *tıyn*) and *jüdek* (instead of *jiidek*). These instances exemplify how transliteration-based reforms failed to fully realign graphemic choices with Kazakh phonology. Aqtay further highlights that this incomplete reform process allowed Cyrillic-based vowel misrepresentations – especially for phonemes traditionally rendered by ‹и› and ‹у› - to persist under new Latin guises. In such cases, the resulting orthography obscures both pronunciation and morpheme boundaries, undermining linguistic transparency. Her findings strongly support the need for phonological reanalysis rather than typographic substitution, reinforcing the central thesis of this article: that successful script reform must prioritize phonemic clarity, systematic graphemic logic, and cognitive accessibility.

These orthographic challenges underscore the need for a systematic metric to evaluate the effectiveness of Kazakhstan's script reform. A qualitative critique alone is insufficient; what is needed is a quantifiable measure that reflects how well each script maps onto the actual sound system of the Kazakh language. This brings us to the concept of *Phoneme-to-Grapheme Correspondence* (PGC), a key indicator of orthographic transparency and phonological fidelity, which measures how accurately a writing system represents the phonemic inventory of a language. Effective PGC ensures phonemic predictability of spelling rules, facilitates literacy acquisition, and maintains linguistic integrity. This study examines whether the 2021 Latin script improves PGC in comparison to Cyrillic by systematically analyzing phoneme representation, quantifying PGC scores, and comparing how well each script preserves the phonetic integrity of the Kazakh language.

To approach this issue comprehensively, the article is structured in two interconnected parts. The first part offers a descriptive overview of the historical and phonological challenges surrounding Kazakh orthography, focusing on structural inconsistencies and the influence of Cyrillic conventions. The second part adopts an experimental methodology using the PGC metric. This dual structure allows the study to combine qualitative linguistic analysis with empirical validation, linking historical critique with measurable reform outcomes.

Background and theoretical framework

This section addresses core phonological and orthographic challenges in the Kazakh language, focusing particularly on the complex relationship between sound systems and their representation in writing. Special attention is paid to problematic phonemes, positional variation, and the ongoing Latinization process.

Phonemic Inventory and Variation

Kazakh phonology comprises a system of 9 vowels - <а>, <ә>, <ы>, <і>, <ұ>, <ү>, <о>, <ө>, <е> - and 19 consonants - <п>, <б>, <м>, <т>, <д>, <н>, <ш>, <ж>, <з>, <с>, <қ>, <к>, <г>, <ғ>, <й>, <у>, <ң>, <р>, and <л>. Additionally, marginal phonemes such as [h] and [ф], though restricted to interjections, play a functional role in lexical differentiation and are thus considered phonemically relevant. For instance, in interjections *like ah, yh, міфә,* and *сүф*, the contrast between [h] and [к] or [ф] and [п] illustrates minimal-pair opposition, indicating their phonemic status (Bazarbayeva, 2012: 45). Today, [h] and [f] occur widely in loans.

The essence of a phoneme lies in its function: to distinguish words, identify grammatical forms, and act as an invariant mental representation. In spoken language, phonemes are realized as a range of phonetically conditioned allophones or variants. In writing, these phonemes are mapped onto graphemes. For example, the phoneme <к> is realized through a wide range of variants: [к], [ғ], [г], [х], depending on position and neighboring sounds. Yet in writing, it is represented uniformly by the grapheme <к>.

Phonological Challenges and Theoretical Models

Phonological ambiguity in Kazakh arises primarily from the interplay of weak and strong positions. The Moscow Phonological School distinguishes between invariant phonemes (those realized in strong positions) and their variants. Variants are phonetically realized forms that may overlap with the strong form of another phoneme, whereas variations are marginal realizations that deviate from the paradigmatic axis altogether. This typology also informs the analysis of phonographic challenges in Kazakh orthography, where Jüsipuly (Jusipuly, 1998) highlighted the problematic status of high vowels <ы>, <і>, <ұ>, <ү>, and <у>, along with the orthographic rendering of diphthongal clusters such as -ый and -ій. These issues arise not only from phonetic reduction but also from mutual sound influences in rapid or unstressed contexts, leading to forms that deviate from their canonical graphemic expectations.

Our analysis builds on, but diverges from, prior proposals (McCollum & Chen, 2021: 276) by adopting an inventory-level metric and a conservative nine-vowel treatment. This gap poses problems when describing morphophonemic transitions such as *tij* + *edi* → *tije-di*, where the vowel /e/ is clearly realized and contrastive (Xin, 2017). If we were to adhere strictly to the phoneme set proposed by McCollum and Chen, such forms would require awkward and phonetically implausible transcriptions like [tij̯-æd̪ɪ] or [tij̯-e̯d̪ɪ] – both of which misrepresent the actual spoken realization. Our analysis departs from this model by explicitly recognizing /e/ as a core vowel phoneme in Kazakh, especially in suffixation and compound contexts. We argue that accurate phonemic inventories must account for such vowel behavior in morphologically productive environments, and that Latin-based orthographic design should preserve this distinction to ensure phonological transparency.

Consider examples such as *ақ үй, көк үй, түнгі*: in coarticulated contexts, phonemes <қ>, <к>, <н> shift to [ғ], [г], and [ң], respectively – e.g., [ағүй], [түңгү]. Despite their acoustic divergence, these variants revert to their canonical forms in isolation, confirming their phonemic unity. Uali illustrates cases of variation in sequences like *ақша, бақша*, and *шұқшии*, where the <қ> sound softens toward [х] or

becomes fricative – not as a new phoneme but as a positional variation. Similarly, disharmonic forms such as *күнәларың* and *ләззаттандым* contain a transitional vowel between <а> and <ә> ([а′]), underscoring the gradient nature of vowel realization. Following Saussurean phonology (Joseph, 2022: 203), it is auditory perception, not mere articulation, that defines phonemic relevance in such contexts.

The Leningrad Phonological Model (LPM) further supports this interpretive flexibility. In morphologically complex words like *жұмұшшұ* ('worker'), variation in <ұ> and [ш] is normalized as phonemic constancy. The segmentation *жұмұш + шұ* indicates that even perceived surface irregularities preserve consistent phonological identities.

The Problem of <и> and <у>: Phonological and Orthographic Ambiguity

A central orthographic issue in Kazakh is the representation of <и> and <у>, both of which are compound phonemes rather than pure monophthongs. <и> is typically composed of either *ый* or *ій*, and <у> of *ұу* or *үу*. Their treatment in the Latin script has sparked intense debate over a conflict between phonetic transparency and orthographic economy. Eldes Omarov observed that the <и> phoneme, when analyzed phonetically, always contains a component of <ы> or <і> (Omarov, 1928), which is why forms like біп-бийт are pronounced [bip-bijt], not [bip-bit]. Hence, <и> is best classified as a diphthongoid or closed diphthong. Similarly, <у> encodes both a vowel and a glide, as in *жазу, келу, жабу*.

Surveys indicate a generational split in orthographic preference: younger and middle-aged speakers favor single-letter representations (u, i), while older speakers prefer compound forms (*ұу, үу, ый, ій*) that align more closely with phonetic structure. The Latin script reform ultimately adopted single-letter forms for both <и> and <у>, citing user familiarity, digital compatibility, and orthographic compactness. This choice aligns with spoken usage. For instance, the words *my* and *би* are better transcribed phonemically as *tuu* and *biy*, with the final [j] representing a palatal glide. These are not indivisible syllabic units as in English or German, but rather span two syllables: *сиыр → сы-йыр, киім → кі-йім*.

It is also important to clarify that the Latin letter ‹j› in the 2021 alphabet represents the voiced postalveolar fricative /ʒ/, corresponding to Cyrillic ‹ж›. This has led to occasional confusion among readers accustomed to other Latin orthographies (e.g., English, German, or Polish), where ‹j› often denotes the palatal glide /j/. We analyze/j/ as a phoneme (consonantal glide). Glide-like realizations within vowel sequences are treated as phonetic implementation; they do not negate phonemic status when contrast and graphemic distinctness obtain.

Hyperphonemes in Kazakh Orthography. Kazakh orthography exhibits several "hyperphonemes" – letter combinations that signify unstable phonological units in weak positions. These include <а-ә>, <у-w>, <п-б>, and <н-ң>.

– <а-ә> Hyperphoneme: Although some linguists (Zhubanov, 1999: 45) questioned the independent status of <ә>, it appears in over 500 initial-syllable words, demonstrating contrastive function. Nevertheless, in second syllables, <ә> may arise from positional variation of <а>, forming a hyperphonemic pair rather than distinct units.

– <у-w> Hyperphoneme: The dual function of <у> - as a vowel in final position (*му, оқу*) and as a consonant in medial or initial position (тау, қауын) – has led to proposals of using <w> in Latin orthography. However, this creates unnatural clusters in native words and inconsistency in loanwords. A dual-system approach – using one spelling for native and another for borrowed words – is cognitively burdensome and counterproductive.

– <п-б> Hyperphoneme: Legacy issues from Arabic and Persian scripts, which lacked a native <п> character, led to instability between <п> and <б>. Today, pairs like *пейіл/бейіл, палуан/балуан* show no semantic contrast, suggesting a hyperphonemic relation.

– <н-ң> Hyperphoneme: The <н>/<ң> opposition is often blurred in roots adjacent to velar consonants: *жоңқа, дүңген, шеңгел*. While <ң> is clearly phonemic in strong positions (*лаң, жаңбыр*), its representation in weak or root-internal positions may reflect historical assimilation rather than true phonemic status.

Zero Opposition and Vowel Reduction. Kazakh exhibits zero opposition involving high vowels <ы> and <і>, where the presence or absence of a vowel is contrastive. These vowels often undergo reduction in unstressed syllables, functioning similarly to Russian <о> and <е>. According to statistical phonetic data, <ы> and <і> exhibit multiple variants – up to four each – depending on prosodic and syllabic position, further complicating their orthographic treatment.

Comparative script reforms: case vignettes (successes and challenges)

To place the Kazakh reform within a broader comparative horizon, we distill six brief vignettes. The first three (Turkey, Azerbaijan, Uzbekistan) anchor the Turkic experience; the remaining non-Turkic contrasts (Vietnam, Mongolia, Serbia) illuminate alternative trajectories – rapid consolidation under mass-education mobilization, extended dual-script regimes, and stabilized digraphia.

Turkey (1928): rapid consolidation via mass schooling: Turkey's reform replaced the Arabic-based Ottoman script with a Latin alphabet via Law No. 1353 ("Türk Harflerinin Kabul ve Tatbiki Hakkında Kanun"), adopted in 1928. Implementation combined compulsory timelines and an unprecedented mass literacy mobilization through the *Millet Mektepleri* (National Schools), opened in January 1929 to teach the new script to out-of-school adults. The social experience – rather than merely legal fiat – proved decisive for rapid uptake (Yilmaz, 2011).

Azerbaijan (1991–2001): deadline-driven enforcement: Following independence, Azerbaijan legislated a Latin script in the early 1990s and completed the transition in August 2001 by presidential decree, mandating Latin across official documents, media, outdoor signage, and publishing; later executive orders consolidated mass re-issuing of works in Latin. This deadline-driven strategy produced swift convergence in public domains, though it relied on centralized enforcement (Landau & Kellner-Heinkele, 2012).

Uzbekistan (1993–present): prolonged coexistence and redesign: Uzbekistan introduced a Latin alphabet by Law No. 931-XII (2 Sept 1993), but the reform has passed through multiple redesigns and extended dual-script coexistence. Official drafts since 2018 repeatedly revisited the treatment of digraphs vs. diacritics (e.g., proposals to replace Sh/Ch, O‘/G‘ with Ş/Ç and diacritic letters), culminating in further tweaks proposed in 2023. In 2020, a presidential decree set a 2020–2030 language-policy horizon, tasked a working group with a full transition plan, and effectively extended the 2023 target. The Uzbek experience demonstrates the cumulative transaction costs of prolonged coexistence and repeated re-specification (keyboards, typography, textbooks, search/indexing) (Landau & Kellner-Heinkele, 2012).

Vietnam's success through vernacular print expansion: Romanized *quốc ngữ* - initiated in early missionary practice – consolidated in the early 20th century as vernacular schooling and print ecologies expanded. Marr's intellectual history tracks how *quốc ngữ* underpinned a surge in vernacular periodicals, primers, and political vocabulary in the 1920s–30s, enabling literacy growth and modern bureaucratic practice; later state policy entrenched its status as the exclusive script. The lesson is

that print-school synergies, not only decree, but also drive durable adoption (Pham, 2022).

Mongolia's reintroduction of traditional script alongside Cyrillic: Policy since 2020 has promoted re-adoption of Mongol bichig in official domains alongside Cyrillic; official communications indicate dual scripts in public office from 2025, with symbolic and geopolitical dimensions shaping timelines. This is a paradigmatic case of state-managed digraphia, where cultural heritage aims and administrative capacity must be balanced to avoid permanent duplication (Makhanov, 2025).

Serbia's stable digraphia: Article 10 of the Serbian Constitution (2006) stipulates the Cyrillic script for official state use, while Latin is widely used across media, commerce, and digital platforms. Recent scholarship classifies Serbia as a coherent case of active digraphia, in which two scripts circulate across the same language community under asymmetric prestige conditions. The Serbian example shows how durable dual-script equilibria can stabilize without formal timelines to convergence – an outcome the Kazakh roadmap explicitly seeks to avoid (Jung & Kim, 2023).

Rapid consolidation (Turkey) involves time-boxing plus mass training; deadline-driven mandates (Azerbaijan) can deliver swift public-domain convergence; extended redesign and coexistence (Uzbekistan) raises duplication costs; print-school alignment (Vietnam) embeds reform socially; heritage-driven co-officiality (Mongolia) requires resourcing to prevent permanent duplication; stabilized digraphia (Serbia) cautions against open-ended windows.

Phonemic Polysemy in Kazakh: Examples from Vowel Systems

The micro-level analysis of the Kazakh sound system has shown which phoneme variants and variations are represented by the 9 vowel letters and the 22 consonant letters of the modern Kazakh alphabet, along with the compound letters *и, я, ю, у*. Although the analysis includes both vowels and consonants, only the vowel system (Kuderinova, 2013: 57) is charted in Table 1 for reasons of space.

(**Key.** INIT = word-initial; MED/FINAL = non-initial positions; AFTER-C = after consonant; AFTER-FRONT-SYL / AFTER-LAB-SYL = harmony triggers; 2SYL/3–4SYL-LABHARM = syllable-indexed harmony; LOAN = loan-driven reading; COLLOQ = colloquial speech; **SURF** = phonetic surface realization (not a distinct phoneme); **Ø** = deletion.)

Grapheme	Phoneme	Other readings (context → value)	Status / note
а	/ɑ/ → **[а]** (all positions)	V-SEQ → **Ø** (first of two vowels drops); BETWEEN **ж/ш/й** → **[ə]**; AFTER-FRONT (≈3rd syl) → **[ä]**; 2SYL-AFTER-FRONT → **[ə]**; ‹а-е› sequence → **[а]**	**SURF** variants [ä], [ə]; **Ø**=deletion.
ə	/æ/ → **[ə]** (INIT / 1st syl, rarely 2nd)	Mostly **LOAN** contexts	Loan-heavy distribution noted.
о	/(o~ө)/ → **[о]** (AFTER-C, 1st syl)	**INIT** → **[uо] SURF**; AFTER-FRONT-SYL → **[ө]**	**SURF** on-glide at word start.
ө	/ө/ → **[ө]** (AFTER-C, 1st syl)	**INIT** → **[uө] SURF**	**SURF** on-glide at word start.
е	/e/ → **[е]** (MED/FINAL)	WITH-BACK-COMPOUND / AFTER **с/ш/ж** → **[э]**; **INIT** → **[je] SURF**; **COLLOQ** → **[ə]**; **BEFORE ш** → **[і]**; **LOAN** → **[и]**; BETWEEN-V / PHRASE-EDGE → **Ø**; AFTER-LAB-SYL → **[ө]**; **3–4SYL-LABHARM** → **[е°]**; ‹е-а› between dissimilar syllables → **[ə]** (often **LOAN**)	Highly polysemous grapheme; several **SURF** values; **Ø** possible.
ы	/ɯ/ → **[ы]** (INIT / non-initial)	CLOSED-SYL & LABHARM → **[ɣØ]**; CLOSED-SYL / BETWEEN-V / 1st-SYL → **Ø**; AFTER-LAB-SYL → **[ɣ]**; **3–4SYL-LABHARM** → **[ы°]**; WITH-FRONT-SYL counterpart → **[і]**; ‹і-ы› between non-harmonizing syl → **[ы]**	Reduction & harmony-driven alternations; **Ø** possible.
і	/ɪ/ → **[і]** (FINAL in monosyllables; non-initial sylls)	AFTER-LAB-SYL → **[ɣ]**; CLOSED-SYL & LABHARM → **[ɣØ]**; ADJ-V / FINAL-WITH-POSS / 1st-SYL → **Ø**	Harmony effects toward **[ɣ]**; **Ø** possible.
ұ	/ʊ/ → **[ұ]** (1st two syllables)	CLOSED-SYL & LABHARM → **Ø**	Reduction in closed/harmony contexts.
ү	/y/ → **[ү]** (1st two syllables)	CLOSED-SYL → **Ø**	Reduction in closed syllables.

Figure 1: Phonemic Polysemy of the Kazakh Vowels

Among 9 vowels, the letters е (10), ы (7), і (4), а (6), and о (3) are polysemous (multi-functional). The graphemes ө (2) and ə (1) carry one or two meanings respectively. The symbol [Ø] indicates that the grapheme is replaced by zero (i.e., it disappears) in spoken language. Symbols like [ɣØ], [ұØ] mean that the graphemes [ɣ] and [ұ] undergo reduction in pronunciation. The notations [yө], [yo], [je] represent diphthongal characteristics and suggest that a slight consonantal sound is heard preceding them. Another set of symbols is used to mark transitional sound shades between back–front and labial–non-labial vowel variants. For example, between the variants [а] and [ə], there exists a transitional variation [ä], which is neither [а] nor [ə], but somewhere in between. Similarly, up to the third syllable, a labial syllable may be replaced by a non-labial syllable with a distinct coloring. Therefore, between labial and non-labial vowels, there are variations such as [eº], [î], [û], which carry features of both. The table also highlights the positions where letters *а, е, ы, і* represent hyperphonemes, which are particularly challenging in Kazakh orthography.

The issue at stake is not merely the much-discussed principle of orthographic economy, nor the fear of unnecessarily lengthening the text due to an excess of letters. Rather, it concerns the cognitive overload that such complexity creates – particularly for those who are learning the alphabet and language for the first time. The task of preserving the authentic pronunciation of Kazakh words and accounting for all their phonetic nuances is the domain of practical transcription and orthoepic dictionaries. Orthography, by contrast, has its own specific function: namely, to simplify writing

and to regulate the rules, norms, and standards of spelling. Thus, the new Latin-based Kazakh script must be easy to read and perceive, and each sound must be assigned a corresponding symbol that does not contradict the principles of Kazakh phonology or the rules of standardized spelling.

Materials and Methods

This study seeks to answer the following research question: To what extent does the Latin script improve phoneme-to-grapheme correspondence in Kazakh compared to the Cyrillic script? It is hypothesized that the Latin script provides superior phonemic representation by eliminating Russian-origin letters that do not correspond to native Kazakh phonemes, employing diacritics to maintain sound-level precision, and reducing graphemic inconsistencies. To evaluate this hypothesis, this study introduces three scenarios that differ in their criteria for phoneme inclusion.

Phoneme recognition and loanwords: criteria and controversies

Competing traditions weigh evidence differently. In this paper we group the most commonly invoked diagnostics into five criteria: (i) contrastive distribution (minimal/near-minimal pairs in the native lexicon), (ii) morphophonemic productivity (stable alternations and regular affixation), (iii) phonotactic integration (occurrence in native slots rather than only in foreign clusters), (iv) functional load & frequency in contemporary usage, and (v) sociolinguistic diffusion (spread across age, region, register, and education). Loanword segments often begin as allophonic or cluster artifacts and phonemicize only when enough of these criteria are met.

Contact with Russian and international vocabulary has introduced segments such as /f, v, h, t͡s, ʂ, ʐ, t͡ɕ, xʷ/. Some of these show broad diffusion (e.g., /f/ and /v/ in general vocabulary), while others remain largely restricted to unadapted clusters (e.g., /t͡s/ in recent borrowings). Interjections and proper names alone are not sufficient evidence for phonemic status; we privilege general-lexicon attestations, compatibility with vowel harmony and Kazakh syllable structure, and participation in derivational morphology. For glides, we apply the contrast + graphemic distinctness criterion: /j/ functions phonemically in the onset position. Where orthography employs ‹i› to cue the glide, this is a context-conditioned mapping and is scored accordingly (see Table 3).

To keep theory and measurement aligned, we classify each segment with a status flag and evaluate two complementary indices:

– Core – contrasts in the native lexicon and participates in productive alternations; fully integrated phonotactics.

– Integrated loan – diffused beyond proper names/quotations; appears in ordinary vocabulary and fits regular phonotactics, but morphophonemic evidence may be lighter.

– Marginal loan – largely confined to foreign clusters/quotations; limited diffusion; no clear morphophonemic participation.

Our three inventories now correspond transparently to these statuses:

– Minimalist = Core only (pre-contact native backbone).

– Standard = Core + Integrated loans (segments meeting ≥2 of the criteria above, at least one being contrast or phonotactic integration).

– Extended = Core + Integrated + Marginal loans (to test orthographic coverage in multilingual usage).

Because PGC is intended to quantify phoneme↔grapheme alignment for Kazakh, segments flagged Marginal loan are reported separately and do not depress the Core average. Where Cyrillic contains letters dedicated to non-Kazakh segments that do not pass our integration tests, we annotate them as NA (not applicable to Kazakh core phonology) rather than “0.0” (which would imply failure to represent a Kazakh

phoneme). This preserves interpretability while still documenting coverage in the Extended scenario.

Comparative analysis of phoneme-grapheme consistency across Cyrillic and Latin scripts

This section introduces a comparative analysis based on three distinct phoneme inventories, allowing for a more structured evaluation of phoneme-grapheme consistency (PGC) across Cyrillic and Latin scripts. By comparing phoneme representation in both Cyrillic and Latin scripts across these three scenarios, this research provides a comprehensive evaluation of script efficiency. The findings will contribute to ongoing discussions on orthographic reform in Kazakh and similar Turkic languages, offering insights into whether Latinization enhances systematic phoneme representation and supports Kazakhstan's linguistic modernization goals.

The Kazakh language, a member of the Kipchak branch of the Turkic language family, has a rich phonemic inventory that plays a crucial role in determining orthographic representation. Vowel harmony is a significant feature in Kazakh, requiring vowels within a word to harmonize as either front or back, influencing suffixation and overall phonotactics (Junisbek, 2009: 58). To assess how well each script represents Kazakh phonology, we employ three distinct phoneme inventories: Minimalist (Native Only), Standard (Native + Stable Borrowings), and Extended (Including Russian Influence). Each scenario represents a different linguistic perspective, influencing how we calculate PGC and interpret the suitability of each writing system. A comprehensive examination of the various approaches to phoneme inventory calculation is provided *inter alia* by Badanbekkyzy and Yeshimbetova (Badanbekkyzy & Yeshimbetova, 2014: 95).

In the Minimalist scenario, the phoneme inventory includes 28 phonemes (19 consonants and 9 vowels), representing only those sounds that are historically native to Kazakh. This scenario eliminates all Russian-origin phonemes and considers only phonemes that existed before major foreign linguistic influences (Ozcelik, 2023: 2). This approach is most suitable for historical phonology studies, as it evaluates how well each script aligns with the core phonological system of Kazakh without external interference. By restricting the inventory to native phonemes, we can determine which script best represents the fundamental structure of the language.

The Standard scenario expands the inventory to 32 phonemes (23 consonants and 9 vowels) by including stable loanwords and commonly integrated sounds, such as /f/ (ф), /v/ (в), and /h/ (х), which have been fully assimilated into modern Kazakh phonology (Bazarbayeva et al., 2023: 1256). This scenario reflects the practical linguistic reality of contemporary Kazakh and is the most balanced for studies related to education and literacy, as it provides an accurate model of the phonemes that native speakers regularly use. By incorporating widely accepted borrowed phonemes, we ensure that the PGC analysis is relevant to modern Kazakh orthography and script reform discussions.

The Extended scenario includes 34 phonemes (25 consonants and still 9 vowels) by adding marginal Russian-origin segments (e.g., loan vowel /**ɛ**/ ‹Э›) and documenting Russian letter conventions (‹И, У›) where they affect loanword spelling. This scenario is particularly valuable for language policy research, script reform discussions, and studies on Kazakh in multilingual settings, as it considers how well each script accommodates the full range of phonemes encountered in modern Kazakh usage. While Cyrillic was originally adapted to accommodate these sounds, Latin-based adaptations attempt to maintain a structured and predictable system, making this an important test case for phoneme-grapheme consistency. Certain Kazakh phonemes exhibit allophonic variations depending on their phonetic environment. For example, the phoneme /k/ may be realized as a voiceless velar plosive [k] or as a

voiceless uvular plosive [q], complicating the establishment of a consistent orthographic standard (Dotton & Wagner, 2018: 16).

We adopt a nine-vowel analysis for Kazakh: /ɑ, æ, e, i, ɪ, ɯ, u, y, ɵ/. The central /ə ~ ɤ/ is treated as a positional/analysis-dependent realization and is not counted as a distinct phoneme; the loan vowel /ɛ/ (‹Э›) is Marginal and excluded from inventory counts. Accordingly, /f, v, h, t͡ʃ/ are Integrated loan segments; /t͡s, ʂ/ are Marginal (restricted to foreign clusters).

We classify segments as core (native contrast + productive alternations), integrated loans (diffused beyond proper names/quotations, regular phonotactics), and marginal loans (restricted to loan clusters). The Minimalist/Standard/Extended inventories correspond to these three statuses.

PGC scoring scale and formula

Phoneme-to-Grapheme Correspondence (PGC) refers to the relationship between the sounds of a language (phonemes) and their written representations (graphemes). Understanding this relationship is crucial for evaluating the efficiency of an orthographic system. A highly efficient orthography maintains a transparent one-to-one relationship between phonemes and graphemes, minimizing ambiguity and promoting literacy acquisition (Johanson, 2021). Languages with shallow orthographies, such as Finnish and Turkish, exhibit a near one-to-one correspondence between phonemes and graphemes, facilitating easier reading and writing acquisition. In contrast, languages with deep orthographies, like English and French, have complex and inconsistent phoneme-grapheme mappings, posing challenges for learners (Bazarbayeva & Chukayeva, 2023: 8).

The Kazakh language, which has undergone multiple script reforms, provides a pertinent case for assessing PGC. The forthcoming transition from Cyrillic to a Latin-based script in 2031 aims to enhance phonemic transparency and align the orthography more closely with the language's phonological structure.

The formula used in this study builds upon established methodologies in phonemic studies. While there is no universal metric for phoneme-to-grapheme correspondence, researchers have developed structured approaches for assessing orthographic efficiency. The work of Venezky (1970) introduced systematic analyses of phoneme-to-grapheme regularity in English, laying the foundation for quantitative assessments. Similarly, Seymour, Aro, and Erskine (Seymour et al., 2003: 143) explored phonemic predictability of spelling rules across European languages, providing comparative methodologies applicable to our study. Coulmas (Coulmas, 2003) further examined writing system efficiency and the predictability of phoneme-to-grapheme mapping in various scripts.

Building upon these frameworks, our study adapts a structured approach to assess the phonemic accuracy of Kazakh orthography in all three scenarios. The PGC calculation considers:

– How consistently each phoneme is represented by a single grapheme.

– Whether there are multiple graphemes for a single phoneme (reducing transparency).

– The extent to which borrowed phonemes disrupt phoneme-grapheme consistency.

The results of this analysis will allow us to compare Cyrillic and Latin-based orthographies across three structured phoneme inventories, providing insights into the efficiency of Kazakh script reform. The next section introduces the PGC formula and its scoring methodology, which will be applied separately to each phonemic scenario. We ground the 1.0→0.0 scale in five orthographic-efficiency axioms: (A1) one-to-one grapheme↔phoneme mapping merits 1.0; (A2) a consistent digraph for a single

phoneme with no rival values scores 0.75; (A3) context-conditioned or ambiguous digraphs score 0.50; (A4) competing spellings for the same phoneme score 0.25; (A5) the absence of any dedicated representation scores 0.0. We do not penalize the mere presence of a single diacritic if mapping remains one-to-one ('diacritic neutrality'). Assign NA (not applicable) for letters that encode non-Kazakh-core segments will be used only in foreign clusters; NA is reported in PGC-All but excluded from PGC-Core. These axioms align with transparency, learnability, and searchability, and they are applied uniformly across scenarios. Our headline metric is PGC-Core, which averages Core + Integrated segments only. Extended is reported for coverage, with Marginal segments marked NA, it is numerically identical to Standard and is interpreted as a transparency check rather than a separate result.

The development of the Phoneme-to-Grapheme Correspondence (PGC) formula and its associated scoring system was carried out in several stages. First, we reviewed and adapted established approaches from prior phonemic and orthographic studies. In particular, the analytical frameworks proposed by Venezky (1970) for assessing phoneme-to-grapheme regularity in English were examined as a baseline model. These methodologies provided systematic procedures for identifying consistent and inconsistent mappings between spoken and written forms.

Next, we incorporated comparative elements from the cross-linguistic analysis of Seymour, Aro, and Erskine (2003), which investigated spelling rule predictability across 13 European orthographies. This comparative framework guided our selection of parameters relevant to measuring orthographic transparency and predictability. Additional conceptual grounding was taken from Coulmas (2003), whose examination of writing system efficiency informed the weighting of variables in our scoring system.

Our proposed formula for PGC calculation follows a structured approach derived from these studies, adapted to assess the phonemic accuracy of Kazakh orthography:

$$PGC_{\{script\}} = \frac{\sum S_i}{N}$$

Where S_i represents the PGC score for each phoneme based on its accuracy in the script, and N is the total number of phonemes included in the given scenario (Minimalist/Standard/Extended), excluding items marked NA. A higher PGC score indicates better phoneme representation. Each phoneme-grapheme mapping is scored according to the following criteria in Figure 2:

Axiom (ID)	Short label	Definition / rule	Typical example	Score
A1	One-to-one mapping	A single grapheme uniquely represents a single phoneme; no rival values.	Cyrillic ‹Ч› for /t͡ʃ/ (unique, unambiguous)	1.0
A2	Consistent Digraph	A fixed two-letter sequence always represents that phoneme; no rival readings.	Latin ‹Ch› for /t͡ʃ/ when not used for other values	0.75
A3	Context-conditioned / ambiguous digraph	The digraph's value varies by context **or** competes with another mapping.	‹Sh› sometimes /ʃ/ but also used across morpheme boundary; or ‹Ch› ~ ‹Ç› co-exist	0.50
A4	Competing spellings	Multiple spellings circulate for the same phoneme within the same scenario.	/t͡ʃ/ alternates between ‹Ch› and ‹Ç› in practice	0.25
A5	No dedicated representation	No stable, recommended way to write the phoneme in the scenario.	/t͡s/ only in unadapted Russian clusters; no Kazakh form	0.0

Figure 2: Scoring Criteria (A1-A5)

Results

In applying the PGC scoring system to Kazakh, we assess a structured set of core phonemes across three phonemic scenarios (Minimalist, Standard, and Extended) in both the Cyrillic and 2021 Latin alphabets. By systematically evaluating phoneme-to-grapheme consistency under different phoneme inventories, we can measure how effectively each script represents Kazakh sounds in different linguistic contexts. The table below presents a sample evaluation of phonemes in Cyrillic and Latin scripts, considering three phonemic scenarios:

Status	Phoneme (IPA)	Cyrillic (letter)	Latin (letter)	Score – Cyrillic	Score – Latin
Core	/ɑ/	А	a	1.0	1.0
Core	/æ/	Ә	ä	1.0	1.0
Core	/e/	Е	e	0.5	1.0
Core	/ɪ/	І	i	1.0	1.0
Core	/i/	И	i	0.5	1.0
Core	/u/	У	ū	0.5	1.0
Core	/y/	Ү	ü	0.5	1.0
Core	/ɵ/	Ө	ö	1.0	1.0
Core	/ɯ/	Ы	y	0.5	1.0
Core	/ʊ/	Ұ	u	1.0	1.0
Marginal loan	/ɛ/	Э	e	NA	NA
Core	/b/	Б	b	1.0	1.0
Core	/d/	Д	d	1.0	1.0
Integrated loan	/f/	Ф	f	1.0	1.0
Core	/g/	Г	g	1.0	1.0
Core	/ɣ/	Ғ	ğ	1.0	1.0
Integrated loan	/h/	Х	h	1.0	1.0
Core	/ʒ/	Ж	j	1.0	1.0
Core	/k/	К	k	1.0	1.0
Core	/l/	Л	l	1.0	1.0
Core	/m/	М	m	1.0	1.0
Core	/n/	Н	n	1.0	1.0
Core	/ŋ/	Ң	ñ	1.0	1.0
Core	/p/	П	p	1.0	1.0
Core	/q/	Қ	q	1.0	1.0
Core	/r/	Р	r	1.0	1.0
Core	/s/	С	s	1.0	1.0
Core	/ʃ/	Ш	ş	1.0	1.0
Marginal loan	/ʂ/	Щ	şş	NA	NA
Core	/t/	Т	t	1.0	1.0
Integrated loan	/tʃ/	Ч	ch	1.0	0.75
Marginal loan	/ʦ/	Ц	ts	NA	NA
Integrated loan	/v/	В	v	1.0	1.0
Core	/z/	З	z	1.0	1.0
Core	/w/	У	u	0.5	0.5

Figure 3: PGC Scoring System for Kazakh Phonemes

Each segment in Figure 3 carries a Status flag (Core / Integrated loan / Marginal loan). We report two summary indices: PGC-Core (averaging only Core + Integrated loans) and PGC-All (including Marginal loans; where a mapping applies only to non-Kazakh clusters we annotate NA rather than 0.0). This separation prevents marginal foreign clusters from distorting the Kazakh core signal while still documenting Extended-scenario coverage.

It is important to clarify that the Kazakh phoneme /ɪ/ (the near-close near-front unrounded vowel) is consistently represented in Cyrillic by the letter ‹і›. Earlier

ambiguities regarding the role of ‹и› in representing /ɪ/ were reassessed, and the current analysis confirms that ‹и› corresponds instead to /i/ or diphthongal realizations such as /ij/. This reinforces the consistency of ‹і› for /ɪ› in standard Cyrillic orthography and justifies its PGC score of 1.0. In the Latin script, the phoneme is represented by the dotted ‹i›, offering an equally systematic and transparent one-to-one mapping. These consistent correspondences across both scripts reflect a high degree of phoneme–grapheme alignment in the representation of /ɪ/.

The symbol /ʂ/ was included to account for rare cases involving Russian loanwords where ‹щ› occurs. However, we acknowledge that /ʂ/ is not part of the core phonemic inventory of Kazakh and should not be considered a native phoneme. Its inclusion serves to reflect occasional orthographic practices rather than phonological necessity. Accordingly, its PGC score was reduced to reflect its marginal status.

Example of the phoneme /t͡ʃ/: /t͡ʃ/ is mapped to ‹Ч› in Cyrillic (A1 = 1.0) and to ‹ch› in Latin (A2 = 0.75). We tag /t͡ʃ/ as an Integrated loan where it has diffused into the general vocabulary. The Latin digraph is consistent but not one-to-one; therefore, Latin incurs the A2 penalty while Cyrillic scores 1.0.

PGC Calculations Across Three Scenarios

Based on these PGC scores, we derive the average PGC for each script under different phoneme inventories. For the Minimalist scenario (N = 28), the Cyrillic script achieves in Figure 4.

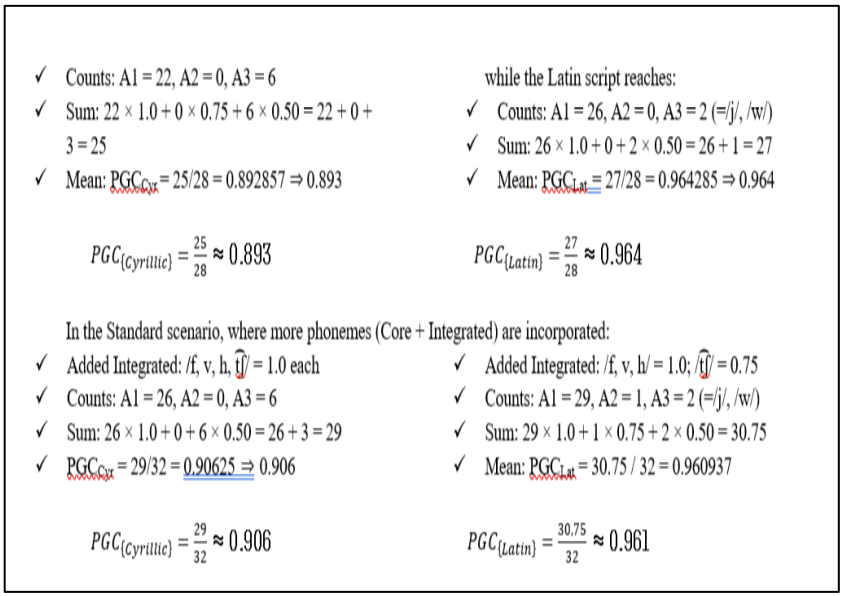


Figure 4: Step-by-step calculation of PGC scores for Kazakh Cyrillic and Latin scripts under the Minimalist and Standard phoneme-inventory scenarios.

Finally, the Extended scenario (which documents marginal borrowed segments) is numerically identical to Standard, because marginal items are marked NA and excluded from the mean.

Discussion

The findings indicate that the 2021 Latin script provides a more transparent phoneme-to-grapheme mapping across all three phonemic scenarios, reducing mismatches and absent phonemes compared to Cyrillic. The Minimalist scenario shows the highest PGC for both scripts, as it excludes borrowed phonemes that introduce irregularities. The Standard scenario, which includes stable loan phonemes, maintains a significant advantage for Latin over Cyrillic, reinforcing the efficiency of Latin orthography for modern Kazakh.

By eliminating Russian phonemes that introduced inconsistencies, the Latin script aligns more closely with the natural phonetics of Kazakh. The use of diacritics in the Latin script enables more precise phonemic representation without introducing unnecessary additional letters. Additionally, the transition to the Latin script supports international transliteration standards, improving the script's adaptability for digital and academic use (Kara, 2022: 99). This study demonstrates that across all three phonemic scenarios, the Latin script consistently offers a more efficient phoneme-to-grapheme mapping system than Cyrillic, reinforcing its suitability for Kazakh orthographic reform.

The larger Latin–Cyrillic difference in the Minimalist scenario follows from Kazakh phonetics. Kazakh's core lexicon is organized by tongue-root–based vowel harmony with nine phonemic vowels and no phonemic palatalization of consonants. Several Cyrillic vowel letters inherited from Russian (notably ‹Е, И, У, Ү, Ы›) encode bundles that mix jotation (onset [j]/[w] on-glides), contextual fronting/rounding, and even vowel deletion in specific environments; these are sensible for Slavic palatalization, but they miscue the underlying Kazakh vowel categories in isolation. By contrast, the Latin design uses single, monofunctional symbols that align with Kazakh's phonemic vowels and treats /j/ and /w/ as independent glides, matching their phonetic realization in onsets. In the Standard scenario, the segments added (/f v h t͡ʃ/) are harmony-neutral consonants that do not interact with Kazakh vowel geometry; Cyrillic already has dedicated letters for them, and Latin's only concession is the conventional ‹ch› digraph. Their inclusion adds unnecessary graphemic complexity, reducing phonemic predictability. The Latin-based *ch* adaptation aligns more closely with phonetic structures in other Turkic languages, enhancing phoneme-grapheme transparency. Because these loans sit outside the harmony-driven core, they dilute the impact of the vowel-system mismatch that advantages Latin in the Minimalist set – hence the narrower gap in Standard. Therefore, we express the following "prediction": any reform that further separates glide symbols from vowel letters and minimizes multifunctional vowel graphemes will enlarge the Minimalist advantage, while adopting single-letter affricates would widen the Standard gap in favor of Latin.

The complexity of Cyrillic orthography hindered standardization and complicated literacy acquisition, as observed in studies on Kazakh phonetics (Bekzhanova & Makoelle, 2022: 3). Research indicates that phonetic adaptations from Russian resulted in orthographic uncertainty, as certain phonemes were represented inconsistently, affecting spelling consistency and fluency. These inconsistencies contributed to lower phoneme-grapheme consistency (PGC) scores, as certain phonemes had overlapping or inconsistent representations.

Additionally, linguistic studies suggest that the retention of Russian phonemes was influenced by socio-political factors rather than linguistic necessity. The need to maintain continuity with previous Cyrillic-based education and the historical role of Russian as a lingua franca contributed to the script's persistence. However, these factors resulted in a less efficient orthographic system compared to the Latin-based alternative.

The transition to the 2021 Latin script was designed to eliminate redundant letters and introduce diacritics for precise phoneme representation. Unlike Cyrillic, the

Latin-based orthography incorporates diacritics (ä, ö, ü) to maintain a one-to-one correspondence between letters and sounds, reducing phonemic ambiguity.

By applying the PGC framework to all three phoneme scenarios, the Latin script consistently achieves higher transparency scores, particularly in the Standard scenario, where native and stable loan phonemes are evaluated. The Latin script removes Russian-origin letters, simplifying the alphabet to better match the Kazakh phonemic structure. Research indicates that this reduction in graphemic redundancy enhances literacy acquisition and improves phonetic accuracy.

A major advantage of the Latin script is its alignment with international linguistic standards, facilitating global digital compatibility. Unlike Cyrillic-based Kazakh, which has encoding issues, Latin-based writing is more efficient for data processing, ensuring smooth integration into modern technological interfaces.

A critical argument for Latinization is its positive impact on reading fluency and comprehension. Studies on phonemics suggest that languages with a more consistent phoneme-grapheme relationship exhibit faster word recognition and higher spelling accuracy. This is particularly beneficial for early literacy learners, who require predictable spelling structures and second-language learners, for whom phonemic predictability aids pronunciation.

By enhancing script efficiency and reducing phonemic inconsistencies, the Latin script demonstrates clear advantages across all three phonemic scenarios. The results of the PGC evaluation reinforce its role as a more transparent and linguistically consistent orthographic system for Kazakh. However, despite the predicted impact of the Latin script transition suggesting improved graphemic alignment with pronunciation and literacy acquisition, its effectiveness may be limited among the older generation, who are more familiar with the Cyrillic script and often lack full proficiency in the Kazakh language, potentially hindering their adaptation to the new orthographic system.

The 2021 Latin-based Kazakh script enhances phoneme–grapheme consistency by employing diacritics to mark phonemic distinctions with systematic regularity. While the Cyrillic script contains a full set of native letters for Kazakh phonemes – many of which are visually distinct (e.g., ‹н› and ‹г›) – these symbols are structurally heterogeneous, often lacking consistent typographic principles. By contrast, the Latin version relies on diacritical extensions of base Latin letters to represent the same sounds, resulting in a more regular and unified typographic system. This increases structural transparency for users already familiar with Latin orthographies and improves orthographic predictability.

The Latin diacritic system ensures a near one-to-one correspondence between phonemes and graphemes, with examples including:

‹ä› for /æ/, distinct from ‹a› /ɑ/

‹ö› for /ɵ/

‹ü› for /y/

‹ū› for /u/ (back high rounded), distinct from ‹u› = /ʊ/ (near-close back rounded)

‹ñ› for /ŋ/, promoting graphic consistency

‹ğ› for /ɣ/, replacing the more idiosyncratic ‹г›

‹ş› for /ʃ/, avoiding reliance on multigraphs

Although Cyrillic letters like ‹ə ө ү ұ ң ғ ш› are immediately recognizable and have long-established usage, their shapes are not governed by a shared visual logic. In contrast, the Latin equivalents ‹ä ö ü ū ñ ğ ş› are formed through systematic, typographically predictable modifications of base Latin letters using standardized diacritics. This regularity enhances learnability for new users and aligns Kazakh orthography with international norms observed in other Turkic Latin scripts (e.g., Turkish ‹ş›, ‹ğ›, ‹ü›). Moreover, diacritics offer high compatibility with Unicode standards and digital environments, streamlining typing, encoding, and cross-script

interoperability. While some Cyrillic forms may appear more visually distinct, the Latin-based script favors structural coherence and cross-linguistic scalability. As such, it offers a different kind of precision – one grounded in visual logic, typographic economy, and technological adaptability.

Finally, although Cyrillic symbols like ‹ң› and ‹ш› are not digraphs but independent letters with unique forms, the Latin script employs diacritic-based solutions like ‹ñ› and ‹ş› that integrate seamlessly into the broader Latin ecosystem. This structural consistency reduces learning barriers and improves spelling predictability for native speakers, language learners, and digital users alike. From the perspective of phoneme-to-grapheme correspondence, this approach contributes to higher PGC scores by minimizing exceptions and promoting rule-based graphemic representation.

Orthographic and Structural Interference of the Cyrillic Script

The continued use of the Cyrillic script in Kazakh presents not only typographic but also significant phonological challenges. A primary concern is the Russian orthographic influence, particularly the palatalizing effect of Cyrillic letters ‹е› and ‹и›. These graphemes often trigger palatalization of preceding consonants, imitating Russian phonotactic patterns. This leads to phonological distortions such as [t] > [tɕ], [d] > [dʑ], and [ɣ] > [ɟ], particularly among children and new learners. For example, *үйде* (‘at home’) may be mispronounced as [yjdʑe] instead of the native [ɯjˈdɛ].

Beyond palatalization, other distortions stem from systemic Cyrillic inconsistencies. The letter ‹е› is frequently read as /je/, causing names such as *Есенгүл* to be rendered as *Yesengül*, despite the absence of an initial glide in native Kazakh. Similarly, ‹у› represents both [w] and [v], leading to orthographic ambiguity: *Ақтау* may be read as *Aktau*, and even non-Kazakh names like *Vahap* appear as *Uahap*. Additionally, Russian-based stress and vowel reduction patterns interfere with Kazakh prosody. A common example is the name Oljas, misread as Aljas under Russian-influenced stress placement.

These cumulative effects reinforce the argument that Cyrillic does not merely serve as a neutral writing system but functions as a conduit for foreign phonological patterns. The Latin script, in contrast, avoids these confounding cues by offering consistent phoneme-grapheme correspondence, eliminating palatalization triggers, and enhancing the acquisition of native Kazakh pronunciation. The structural coherence and digital adaptability of Latinization also help protect Kazakh's phonological integrity in both spoken and written forms (Rysbergen et al., 2025: 42).

Another crucial advantage of Latinization is its role in improving global transliteration and digital usability by aligning Kazakh with international linguistic standards. The Cyrillic script presents challenges in computational processing, digital encoding, and international communication, as non-Latin scripts require additional input methods and software adaptations (Clement, 2008: 171). The transition to Latin eliminates these issues by enabling seamless transliteration and ensuring smooth integration into global digital systems, including:

Internet search engines (ensuring more effective indexing and results)

Machine translation tools (enhancing accuracy in automated translation)

International publishing platforms (improving access to Kazakh-language texts)

Additionally, many Turkic languages, such as Turkish, Azerbaijani, and Uzbek, have successfully transitioned to Latin-based scripts, fostering greater linguistic and cultural connectivity across the Turkic-speaking world. The adoption of Latin-based writing makes Kazakh more accessible to global readers, researchers, and policymakers, while facilitating cross-linguistic comparisons and strengthening diplomatic and academic exchanges.

By enhancing systematic phoneme representation, eliminating inconsistencies, and improving digital compatibility, the Latin script demonstrates a clear advantage

over Cyrillic across all three phonemic scenarios. The results of the PGC analysis reinforce the Latin script's efficiency as a more linguistically consistent and globally integrated orthographic system for Kazakh.

Beyond phonological accuracy and typographic efficiency, the choice of script also carries significant cultural and geopolitical weight. The continued use of the Cyrillic alphabet in Kazakh maintains a symbolic tie to the Russian linguistic sphere, echoing decades of Soviet influence and policy-driven Russification. While the Cyrillic script has been adapted to represent Kazakh phonemes, its historical and ideological associations contribute to perceptions of linguistic dependency. The shift to a Latin-based script, while not a panacea, represents a broader cultural reorientation – one aligned with Turkic linguistic identity, digital globalization, and symbolic autonomy. As numerous scholars have argued, script change functions not only as an orthographic decision but also as an act of language planning and identity construction in post-Soviet societies (Bazarbayeva&Chukayeva, 2023). From this perspective, the Latinization of Kazakh supports both linguistic modernization and cultural differentiation from the Russian matrix.

Managing dual-script coexistence

A planned period of scriptal digraphia – here understood as the regulated, time-limited coexistence of Cyrillic and Latin – can mitigate transition costs while safeguarding communicative continuity across domains (education, administration, media, and the knowledge sector). The objective is twofold: (i) to minimize transactional frictions in everyday literacy practices and digital infrastructures, and (ii) to prevent long-term path dependence whereby ad hoc workarounds harden into de facto policy. We therefore propose a 5–7-year window with a fixed sunset clause, measurable milestones, and a tightly specified set of technical and editorial controls.

Without clear governance, dual-script regimes tend to produce redundancy (duplicate editions, signage, databases), search and indexing fragmentation, and cohort effects in schooling (assessment inequities between early and late adopters). Editorial drift (hybrid spellings, inconsistent transliteration) and heterogeneous IT readiness further amplify costs. The measures below are designed to reduce these frictions while preserving the integrity of the Kazakh core phonological system that motivates the reform. To operationalize the time-limited digraphia, we propose the following policy- and infrastructure-oriented mitigation checklist:

1.*Canonical transliteration tables in public registries*. Establish a state-maintained, versioned mapping (Cyrillic↔Latin) with round-trip fidelity for toponyms, anthroponyms, and institutional names. Expose the mapping via an open API and embed it in civil registries, cadastres, bibliographic authority files, and e-government forms. Document edge cases (apostrophes, hyphenation, digraph boundaries) to prevent hybrid or non-invertible spellings.

2.*Parallel OCR/ASR models and search bridging.* Maintain paired OCR (scanned text) and ASR (speech-to-text) pipelines for both scripts; publish error-rate benchmarks by domain (press, textbooks, legal). Implement cross-script query expansion in search portals (a query in one script automatically retrieves results in the other via the canonical mapping) and require repositories to index both scripts at ingest.

3.*Strict style guidance for media and publishing.* Issue a national stylebook defining (a) script choice by genre and audience during the window, (b) treatment of loanwords and proper names, (c) typography (diacritics, capitalization, hyphenation), and (d) disallowance of hybrid spellings. Tie public procurement and textbook approvals to compliance with this stylebook to prevent editorial drift.

4.*Wayfinding and signage replacement cadence.* Adopt a staged replacement plan prioritizing high-salience nodes (airports, rail, hospitals, schools) in years 1–3,

regional/municipal signage in years 3–5, and low-traffic assets thereafter. Require dual-script display during the window with standardized layout (Latin primacy or parity clearly specified) and publish a transparent budget and timetable to avoid indefinite postponement.

5.*Assessment and examination accommodations.* For transitional cohorts, permit production in either script while converging test stimuli to Latin by year 3. Provide dual-script item banks, teacher training modules, and marking rubrics that neutralize script as a confound. Archive exemplar answers in both scripts to guide teachers and students.

6.*Data architecture for forms, IDs, and records.* Enforce normalized storage with dual fields (canonical Latin + legacy Cyrillic) and deterministic transliteration at write-time; apply cross-field deduplication in citizen and business registries. Specify Unicode normalization (NFC/NFD) and locale tags to stabilize interchange across operating systems and databases.

7.*Academic and bibliographic practice.* Harmonize citation and cataloging rules: retain original-script titles in references with standardized transliteration in brackets; require repositories and journals to attach both script variants to metadata to ensure interoperability across indices.

8.*Keyboards, fonts, and UI readiness.* Publish national keyboard layouts and font recommendations; certify platform compliance for major OS/browser vendors used in public services. Mandate webfont bundling for e-government sites to prevent fallback glyph issues; disseminate IME (input method editor) guidance for schools and newsrooms.

9.*Monitoring, milestones, and sunset clause.* Define adoption indicators (school textbook share, media output share, signage replacement rates, registry conversion coverage, search parity metrics) with annual public reporting. Link any deadline extension to explicit thresholds rather than open-ended deferral. At sunset, deprecate the legacy script in official domains while maintaining archival access.

For instance, Uzbekistan's Latinization illustrates how prolonged scriptal digraphia accumulates cost and editorial drift. The legal baseline is Law No. 931-XII (2 Sept. 1993) introducing a Latin alphabet, followed by a 1995 revision that regularized several symbols and entrenched *sh/ch* digraphs alongside O‘/G‘ (Latin apostrophe letters). Since 2018, successive public drafts have sought to replace digraphs with diacritics – Ş/Ç for *sh/sh*, and diacritic letters for O‘/G‘ - with iterations in 2018, 2019, and 2021 (public consultations), and a 2023 “four-letters” proposal (O‘→Õ, G‘→Ğ, Ch→C, Sh→Ş). These rounds demonstrate the transaction costs of repeated re-specification for keyboards, typography, education, and search. Policy targets also shifted: after an official roadmap setting 1 January 2023 for full transition, the presidency's 2020 decree established a 2020–2030 language policy horizon and mandated a government working group to deliver the detailed schedule; implementation has since continued under this extended framework. The Uzbek case therefore, supports our mitigation stance: dual-script windows must be time-boxed, style-guided, and infrastructure-funded; otherwise, duplication (parallel textbooks/signage), indexing fragmentation, and heterogeneous editorial practice become persistent rather than transitional features (Narmatova & Abdurakhmanova, 2022: 144).

A well-governed dual-script window complements our phoneme-centric evaluation: by ensuring consistent cross-script mappings and editorial controls, the transition protects the very phonological contrasts that our PGC metrics reward, while containing the socio-technical costs that accompany large-scale orthographic change.

Practical rollout: acceptance, cost, and capacity

Evidence from prior reforms shows that teacher-training volume and speed, publisher conversion pipelines, and public-facing service readiness predict acceptance

more than abstract design debates. Turkey's 1928 reform leveraged a nationwide literacy mobilization through the *Millet Mektepleri* (National Schools) and imposed rapid legal deadlines via Law No. 1353; Uzbekistan's prolonged dual-script phase has multiplied duplication costs, while Azerbaijan relied on decree-driven deadlines to force convergence. These lessons suggest pairing the Kazakh Latin rollout with funded teacher retooling, phased publisher subsidies, and unified digital guidelines.

First, a time-boxed cascade training program (national core trainers → regional leads → school teams) with compulsory certification and open materials reduces regional uptake gaps; Turkey's fast-tracked training architecture offers a clear historical template. Second, publisher conversion grants tied to verifiable milestones (fonts and templates; pilot titles; backlist plans) align incentives with delivery rather than intent; Azerbaijan's switch enforced Latin in official and media domains by decree, illustrating the effectiveness – and risks – of deadline-anchored incentives. Third, digital standards issued at T-0 (keyboard layouts, fonts, hyphenation, OCR/ASR models) constrain platform fragmentation and enable discoverability during coexistence.

Transition costs cluster into (i) one-off investments (teacher training, tooling, high-salience signage) and (ii) recurrent coexistence costs (parallel editions, help-desk load, data QA). Sequencing should prioritize high-visibility, high-throughput domains (primary textbooks, national portals, transport/health signage) in years 1–3, shifting to lower-traffic assets thereafter. Procurement rules can require Latin-first deliverables with dual-script access during the window. Uzbekistan's experience illustrates how coexistence can entrench duplication if not strictly time-boxed.

We propose a compact set of measurable indicators reported annually: (1) share of new titles in Latin; (2) teacher certification coverage; (3) registry conversion to canonical Latin fields; (4) signage replacement at priority sites; (5) search parity (recall@k for cross-script queries (recall@k denotes the share of all relevant items that appear within the top k results (recall@k = $|\text{Rel} \cap \text{Topk}| / |\text{Rel}|$). Here it is computed for cross-script queries (Latin↔Cyrillic) using canonical transliteration/aliasing; higher values indicate better cross-script retrieval.)). In contexts where political targets shift, deadline extensions should be conditional on observable thresholds rather than open-ended. Key risks include editorial drift (hybrid spellings), capacity asymmetries (rural vs. urban), and cost duplication during coexistence. Mandatory style compliance in publicly funded outputs, targeted rural training bursaries, and centralized conversion toolkits (fonts, transliteration APIs, OCR/ASR models) mitigate each risk. Comparative policy syntheses on Central Asia and Turkic reforms consistently stress the need for binding standards and funded capacity alongside orthographic design (Sebba, 2009: 35).

These implementation choices are orthogonal to our phoneme-centric findings, yet determine whether the PGC advantages of the Latin design are realized at scale. Aligning capacity, incentives, and standards helps translate improved phoneme–grapheme mapping into durable literacy practices and digital interoperability.

Long-term socio-cultural effects

Script reform reconfigures both symbolic identity and practical literacies over the course of generations. In contexts such as Kazakhstan and Uzbekistan, alphabet change has been framed not merely as a technical modernization but as a project of nation-branding, geopolitical repositioning, and intergenerational realignment of literacy practices, which makes downstream socio-cultural effects decisive for evaluation. Recent scholarship links Kazakhstan's Latinization explicitly to identity politics and language hierarchies, with attitudes and uptake varying by age, schooling, and urban–rural gradients.

Reform effects manifest through (i) cohort replacement (students schooled exclusively in the new script reshape publishing, exams, media consumption), (ii) market equilibria (publishers, platforms, and search engines converge on the higher-ROI script), and (iii) diaspora divergence (legacy scripts persist longer outside the educational core, stabilizing *scriptal digraphia*). In post-socialist settings, durable dual-script equilibria are well documented; their governance determines whether coexistence remains transitional or becomes a long-term settlement.

In Kazakhstan, studies report that public discourse around Latinization interweaves identity, modernization, and policy narratives; survey and interview work indicate differentiated acceptance and expectations across cohorts and professions. In Uzbekistan, the legal shift (1993; 1995) followed by redesign rounds (2018–2023) and a 2020 decree establishing a 2020–2030 horizon have prolonged dual-script usage, with attendant duplication in publishing, signage, and IT/search – illustrating how governance of the transition window shapes socio-cultural outcomes (Narmatova, S., & Abdurakhmanova, 2022).

In contemporary literacies, Unicode coverage, OCR/ASR accuracy, and cross-script search/indexing mediate prestige and access; without standardized keyboards, fonts, and transliteration pipelines, the same orthographic design can yield very different public experiences and media ecologies. This aligns with broader sociolinguistic accounts that relocate writing systems from static "codes" to infrastructures embedded in institutions and platforms. The Kazakh reform's long-term success will depend less on its letter inventory per se than on cohort-level schooling, market convergence in publishing/media, diaspora interface policies, and digital standards that ensure discoverability and archival continuity across scripts.

Script change reconfigures symbolic identity and practical literacies. Over time, cohorts schooled exclusively in the new script reshape publishing and media consumption, while diaspora communities often maintain legacy scripts longer, producing scriptal diglossia. Digital affordances (Unicode coverage, OCR/ASR quality, web indexing) increasingly mediate prestige and access, so IT standardization is not ancillary but central to socio-cultural outcomes (Sebba, 2009).

Conclusions

Across all three scenarios, the Latin design provides more precise phoneme–grapheme mapping than the current Cyrillic: Minimalist (N=28) Latin 0.964 vs. Cyrillic 0.893 and Standard (N=32) 0.961 vs. 0.906. This ordering is stable under our status tagging (Core / Integrated / Marginal) and scoring axioms, including the conservative treatment of /j/ in Latin as a context-conditioned mapping.

Beyond the aggregate scores, the implication is straightforward: a Latin orthography reduces ambiguity in everyday reading and writing, lowers editorial friction for publishers, and improves searchability and OCR/ASR performance across public services. Realizing these gains hinges on policy design rather than letterforms alone—time-boxed dual-script coexistence with clear style guidance, funded teacher re-training, templated publisher conversion pipelines, and unified digital standards (keyboards, fonts, hyphenation, OCR/ASR, and cross-script recall@k). Progress should be tracked with a compact indicator set (share of new titles in Latin, teacher certification coverage, registry and portal conversion, signage replacement, and cross-script retrieval quality). Our estimates are inventory-level and deliberately conservative; future work should add corpus-weighted robustness and psycholinguistic testing to link orthographic design to comprehension and speed. Taken together, the evidence supports Latin as the more faithful fit to Kazakh phonology and the more operationally scalable choice for a modern information environment.

Funding

This work was supported by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan under Grant BR24993111 "Developing the speech synthesis system based on the orthoepic norm of the Kazakh language"

Bibliographic references

Amirjanova, N. (2012). Qazaqstandagy latyn jazuynyn tarihi tagylymy [The historical lessons of the Latin script in Kazakhstan]. Almaty: Memlekettik Tildi Damytu Instituty.

Amirjanova, N. (2016). Latin grafikasy negizindegi Qazaq alipbiyi (1929-1940). In: Faziljanova, A. (ed.). Jana ulttyq alipbiyi negizinde Qazaq jazuvin reformalav: Teoriyasi men praktikasi. Almaty: Qazaq Tili Baspasy. 14-15.

Aqtay, G. (2020). New Kazakh alphabet based on Latin script. In: Aqtay, G., & Erdem, C. (eds.). Language and society in Kazakhstan: The Kazakh context. Poznan: Adam Mickiewicz University Press. 23-36.

Badanbekkyzy, Z., & Yeshimbetova, Z. B. (2014). Inventory of Phonemes in Kazakh Language. IMPACT: International Journal of Research in Humanities, Arts and Literature 2(5), 95-102. ISSN 2321-8878.

Baldauf, I. (1993). Schriftreform und Schriftwechsel bei den muslimischen Russland- und Sowjetturken (1850-1937): Ein Symptom ideengeschichtlicher und kulturpolitischer Entwicklungen. Bibliotheca Orientalis Hungarica, Vol. XL. Budapest: Akademiai Kiado.

Bazarbayeva, Z. M. (2012). Qazaq fonologiyasynyn negizderi [Basics of Kazakh phonology]. Almaty: Memlekettik tildi damytu instituty. 120 p. ISBN 978-601-7206-38-3.

Bazarbayeva, Z. M., & Chukayeva, T. K. (2023). The phonematic principle of the Kazakh script based on Latin graphics. Bulletin of L.N. Gumilyov Eurasian National University. PHILOLOGY Series 142(1), 8-14. DOI: 10.32523/2616-678X-2023-142-1-8-14. ISSN 2616-678X; eISSN 2663-1288.

Bazarbayeva, Z. M., Zharkynbekova, S. K., Amanbayeva, A. Zh., Zhumabayeva, Z. T., & Karshygayeva, A. A. (2023). The National Corpus of Kazakh Language: Development of Phonetic and Prosodic Markers. Journal of Siberian Federal University. Humanities & Social Sciences 16(8), 1256-1270. ISSN 1997-1370; eISSN 2313-6014. EDN: IVPVAN.

Bekzhanova, Z., & Makoelle, T. M. (2022). Latinization of the Kazakh Alphabet: Implications for Education, Inclusion, and Social Cohesion in Kazakhstan. SAGE Open 12(4), article 21582440221138820. DOI: 10.1177/21582440221138820. ISSN 2158-2440.

Clement, V. (2008). Emblems of independence: script choice in post-Soviet Turkmenistan. International Journal of the Sociology of Language 192, 171-185. DOI: 10.1515/IJSL.2008.042.

Coulmas, F. (2003). Writing Systems: An Introduction to Their Linguistic Analysis. Cambridge: Cambridge University Press. ISBN 9781139164597.

Dotton, Z., & Wagner, J. D. (2018). A Grammar of Kazakh. Durham, NC: Duke University Slavic and East European Language Resource Center. Available at: https://slaviccenters.duke.edu/sites/slaviccenters.duke.edu/files/file-attachments/kazakh-grammar.pdf. Accessed 29 April 2026.

Johanson, L. (2021). Phonemes in Prime Syllables. In: Turkic. Cambridge: Cambridge University Press. 254-291. DOI: 10.1017/9781139016704.014. ISBN 9781009038218.

Joseph, J. E. (2022). Saussure and structural phonology. In: Dresher, B. E., & van der Hulst, H. (eds.). The Oxford History of Phonology. Oxford: Oxford University Press. pp. 203-220. DOI: 10.1093/oso/9780198796800.003.0010. ISBN 9780198796800.
Jung, Y., & Kim, B. (2023). Coexistence of Multiple Writing Systems: Classifying Digraphia in Post-Socialist Countries. Journal of Eurasian Studies 14(2), 139-150. DOI: 10.1177/18793665231188380.
Junisbek, A. (2009). Qazaq fonetikasy [Kazakh phonetics]. Almaty: Arys. ISBN 9965-17-602-7.
Jusipuly, M. (1998). Ahmet Baitursynuly zhane qazirgi qazaq tili fonologiyasy [Ahmet Baitursynuly and the phonology of contemporary Kazakh]. Almaty: Gylym. 216 p.
Kara, A. (2022). Kazakistan'in Latin Alfabesine Gecis Sureci ve Turk Dunyasinda Alfabe Birligi. In: Bedir, A., Soykan, B., & Imamoglu, U. C. D. (eds.). Uluslararasi Prof. Dr. Halil Inalcik Tarih ve Tarihcilik Sempozyumu Bildiriler II. Cilt. Ankara: Turk Tarih Kurumu. 99-110. DOI: 10.37879/9789751749994.2022.7. eISBN 978-975-17-4999-4.
Kuderinova, Q. B. (2013). Qazaq jazuynyn tarihy men teoriyasy [The History and Theory of the Kazakh Writing System]. Almaty: Eltanym Baspasy. ISBN 978-601-7348-60-1.
Landau, J. M., & Kellner-Heinkele, B. (2012). Language Politics in Contemporary Central Asia: National and Ethnic Identity and the Soviet Legacy. London: I. B. Tauris. ISBN 9781848858206.
Makhanov, K. (2025). The last act of de-communization: The reintroduction of the traditional script in Mongolia. Eurasian Research Institute. Accessed 29 April 2026.
McCollum, A. G., & Chen, S. (2021). Kazakh. Journal of the International Phonetic Association 51(2), pp. 276-298. DOI: https://doi.org/10.1017/S0025100319000185.
Narmatova, S., & Abdurakhmanova, M. (2022). Language Policy in Uzbekistan: The adoption of Uzbek language and the Latin alphabet. INContext: Studies in Translation and Interculturalism 2(3), 144-155. DOI: 10.54754/incontext.v2i3.34. ISSN 2799-6190; eISSN 2799-8952.
Omarov, E. O. (1928). O sochetanii zvukov kazakhskogo yazyka [On the combination of sounds in the Kazakh language]. In: God raboty kazakhskogo vysshego pedagogicheskogo instituta. Tashkent.
Ozcelik, O. (2023). Kazakh phonology. In: Johanson, L., Csato, E. A., Karoly, L., & Menz, A. (eds.). Encyclopedia of Turkic Languages and Linguistics Online. Leiden: Brill. Available at: https://oozcelik.pages.iu.edu/papers/Kazakh%20phonology.pdf. Accessed 15 August 2025.
Pham, T. K. L. (2022). Histoire de l'ecriture romanisee du vietnamien (1615-1919). Paris: Les Indes savantes. 320 p. ISBN 978-2-84654-616-4.
Rysbergen, K., Rsaliyeva, N., Pashan, D., Schochenmaier, E., & Arlen, S. (2025). From Foreign to Familiar: Transliteration Challenges in the Latinization of Kazakh Exonyms. Forum for Linguistic Studies 7(6), pp. 42-55. DOI: 10.30564/fls.v7i6.9617.
Sebba, M. (2009). Sociolinguistic approaches to writing systems research. Writing Systems Research 1(1), 35-49. DOI: 10.1093/wsr/wsp002. ISSN 1758-681X.
Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. British Journal of Psychology 94(2), 143-174. DOI: 10.1348/000712603321661859. ISSN 0007-1269.
Venezky, R. L. (1970). The Structure of English Orthography. Berlin, New York: De Gruyter Mouton. DOI: 10.1515/9783110804478. ISBN 978-3-11-080447-8.
Xin, R.-q. (2017). The acoustic analysis of Kazak vowel /e/. Proceedings of the 2017 3rd International Conference on Economics, Social Science, Arts, Education and Management Engineering (ESSAEME 2017). Advances in Social Science, Education and Humanities Research 119, 1928-1932. Atlantis Press. DOI: 10.2991/essaeme-17.2017.392. ISBN 978-94-6252-367-8. ISSN 2352-5398.

Yilmaz, H. (2011). Learning to read (again): The social experiences of Turkey's 1928 alphabet reform. International Journal of Middle East Studies 43(4), 677-697. DOI: 10.1017/S0020743811000900.
Zhubanov, Q. (1999). Qazaq tili zhonindegi zertteuler [Research on the Kazakh language]. Almaty: Gylym. 581 p. ISBN 5-628-02411-2.

Words: 10251
Characters: 75 858 (42 standard pages)

Didar A. Sadyk
Al-Farabi Kazakh National University
050040, Almaty, Al-Farabi Avenue, 71
Kazakhstan
didarsadyk2@gmail.com
ORCID: 0000-0002-8271-7165

Prof. Kyzdarkhan K. Rysbergen
Institute of Linguistics named after A. Baitursynuly
050010, Almaty, Kurmangazy Street, 29
Kazakhstan
ms.rysbergen@mail.ru
ORCID: 0000-0002-4205-9655

Dr. Eugen Schochenmaier
Mondonomo LLC, ICOS, e-Onomastics
D-77933, Lahr/Schwarzwald, Kanadaring 49/1
Germany
dr.schochenmaier@gmail.com
ORCID: 0000-0002-1980-3378

Prof. Zeinep M. Bazarbayeva
Institute of Linguistics named after A. Baitursynuly
050010, Almaty, Kurmangazy Street, 29
Kazakhstan
zeynepmb@mail.ru
ORCID: 0000-0003-1141-1027