

Anglicisms in Uzbek Online Media Discourse: A Diachronic Corpus-Based Study of Lexical Borrowing, 2010–2024

Gulnora Rashidova^{1*}, Martin Novák²

¹*Department of Uzbek Linguistics, Uzbekistan*

²*Faculty of Humanities, Bohemian University of Prague, Czech Republic*

ARTICLE INFO

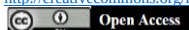
Keywords:

language contact; lexical borrowing; anglicisms; Uzbek language; corpus linguistics; media discourse

Vol.01, No.02, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CCBY4.0).

<http://creativecommons.org/licenses/by/4.0/>



ABSTRACT

Contact between English and the languages of Central Asia has intensified sharply since the 2000s, yet quantitative descriptions of its lexical consequences for Uzbek remain scarce. This study documents the growth, distribution, and integration of English lexical borrowings (anglicisms) in Uzbek online news discourse through a purpose-built diachronic corpus of 2.4 million tokens sampled biennially from six major Uzbek-language news portals between 2010 and 2024, stratified across four topical domains (technology, economy, culture, sport). Anglicisms were identified through a semi-automatic pipeline combining seed-list matching, orthographic pattern detection, and full manual validation by two annotators ($\kappa = 0.88$), and were analysed for normalised frequency, domain distribution, and formal adaptation strategy. The overall anglicism rate more than trebled across the period, from 18.4 to 56.7 items per 10,000 tokens, with statistically significant increases at every biennial step (log-likelihood, $p < 0.001$) and the steepest domain growth in technology (31.2 to 84.3 per 10,000). Orthographic assimilation to Uzbek Latin script was the dominant integration strategy (62.4% of tokens), followed by unassimilated transfer (24.1%), hybrid formations and loan translations (8.7%), and morphologically integrated forms carrying Uzbek agglutinative suffixation (4.8%). For eight of the fifteen most frequent anglicisms, an established Uzbek or earlier Russian-mediated equivalent lost corpus share to the English item over the period. The findings position Uzbek media discourse in an active phase of contact-induced lexical change in which English has displaced Russian as the primary donor, and demonstrate that borrowing is proceeding with rapid grammatical integration rather than code-mixing at the margins.

Introduction

Lexical borrowing is the most visible linguistic consequence of cultural and economic contact, and the global expansion of English has made anglicisms a central object of contact linguistics across dozens of recipient languages [1–4]. For the Turkic languages of post-Soviet Central Asia, the phenomenon carries particular sociolinguistic interest: through the twentieth century, Russian served as the region’s donor of modern terminology and its gateway for internationalisms, a role institutionalised by Soviet language policy [5, 6]. The independence period reorganised this contact ecology—Uzbekistan’s shift to a Latin-based script, the expansion of English education, and above all the arrival of a globalised internet created conditions for direct English–Uzbek contact bypassing Russian mediation [6, 7].

Online news media constitute a strategic observation site for this process. Media language is innovation-friendly yet publicly normed, more sensitive than legal or academic registers but more edited than social media, and it is where new terminology reaches, and is ratified for, the general reading public [8, 9]. Impressionistic commentary on the “flood” of English into Uzbek media abounds in local metalinguistic debate, but systematic quantification—how much, in which domains, at what pace, and with what degree of structural integration—has been lacking, leaving both descriptive linguistics and language-policy discussion without an empirical baseline.

This study provides that baseline through a diachronic corpus investigation guided by three research questions. RQ1: How has the frequency of anglicisms in Uzbek online news changed between 2010 and 2024, overall and across topical domains? RQ2: Which formal adaptation strategies characterise the integration of these borrowings into Uzbek orthography and morphology? RQ3: Is there evidence of competition in which anglicisms displace established native or Russian-mediated equivalents? The article reviews the theoretical and regional literature (Section 2), details corpus construction and the identification pipeline (Section 3), reports quantitative results (Section 4), and discusses their linguistic and sociolinguistic significance (Sections 5–6).

2. Literature Review

2.1. Lexical borrowing in contact linguistics

Foundational work established the core descriptive apparatus: Haugen’s typology of loanwords, loanblends, and loanshifts [1], Weinreich’s account of interference in contact [2], and subsequent syntheses situating borrowing on a continuum of contact intensity [3, 10]. Cross-linguistic projects have shown that lexical borrowability follows semantic-field hierarchies, with technology and trade vocabularies most permeable [4]. Anglicism research in European languages provides the methodological template for the present study—corpus-based frequency profiling combined with formal integration typologies [11, 12]—while media-oriented studies demonstrate the register’s role as the principal diffusion channel for contact vocabulary [8, 9].

2.2. The Uzbek contact setting

Uzbek, a Karluk Turkic language with agglutinative morphology, has historically layered Persian-Arabic, then Russian lexical strata [5, 13]. Soviet-era terminology planning routed internationalisms through Russian phonology and Cyrillic orthography, and Russian remained the de facto donor language into the 1990s [5, 6]. Post-independence policy—Latinisation, corpus-planning commissions, and the promotion of Uzbek in official domains—created an environment in which donor-language choice itself became ideologically charged [6, 7, 14]. Existing scholarship on English influence in Uzbek is largely qualitative or pedagogical; corpus-quantified diachronic evidence of the kind available for German, Spanish, or Scandinavian anglicisms [11, 12] has not previously been assembled.

2.3. Research gap

No published study to date tracks anglicism frequency in Uzbek across a multi-year window with domain stratification and integration typology. The present work fills this descriptive gap and, by including items with established pre-existing equivalents, contributes to the theoretical question of whether “luxury” borrowing (adoption despite available native means) or “necessity” borrowing (adoption filling lexical gaps) dominates the current wave [10, 12].

3. Materials and Methods

3.1. Corpus design and composition

A diachronic corpus of 2.4 million tokens was constructed from the archives of six high-circulation Uzbek-language online news portals, sampled in eight biennial subcorpora (2010, 2012, ..., 2024) of approximately 300,000 tokens each. Within each year, articles were stratified into four topical domains—technology, economy, culture, and sport—using the portals’ own section metadata, with equal domain targets per subcorpus (Table 1). Texts in Cyrillic Uzbek were transliterated to the current Latin standard through a rule-based converter with manual spot-checking (error rate < 0.3% on a 5,000-token sample), ensuring orthographic comparability across the period. Boilerplate, advertisements, and direct Russian- or English-language quotations were removed before tokenisation.

Table 1. Corpus composition (tokens per domain and year, thousands).

Subcorpus	Technology	Economy	Culture	Sport	Total
2010	74	76	75	74	299

2014	75	77	74	75	301
2018	76	75	76	74	301
2024	75	76	75	73	299

3.2. Anglicism identification

An anglicism was operationalised as a lexical item of English origin (including internationalisms whose immediate donor form is English, e.g. *startap* < startup) attested in the corpus regardless of script adaptation, excluding proper names, URLs, and single-letter symbols. Identification proceeded in three passes: (i) matching against a 1,150-item seed lexicon compiled from dictionaries of recent borrowings and prior literature; (ii) automatic flagging of orthographic patterns atypical for inherited Uzbek vocabulary (e.g. word-initial consonant clusters, ⟨w⟩, ⟨c⟩ in Latin-script items); and (iii) exhaustive manual validation of all flagged candidates plus a 10% random sample of unflagged text by two trained annotators (inter-annotator agreement $\kappa = 0.88$; disagreements resolved by discussion with a third linguist). Counting was lemma-based: inflected forms carrying Uzbek suffixation (e.g. *smartfonlar* ‘smartphones’, *brendning* ‘of the brand’) were reduced to their borrowed stems.

3.3. Analysis

Frequencies were normalised per 10,000 tokens. Diachronic and domain differences were tested with log-likelihood statistics for corpus frequency comparison [15], with $\alpha = 0.001$ given multiple comparisons. Each validated type was coded for adaptation strategy—(a) orthographically assimilated to Uzbek Latin conventions, (b) unassimilated (donor spelling retained), (c) hybrid or loan-translation formations, (d) morphologically integrated tokens (borrowed stem plus Uzbek affixes)—following the descriptive categories of the anglicism literature [1, 11]. For fifteen high-frequency anglicisms possessing an established earlier equivalent (native, Persian-Arabic stratum, or Russian-mediated), the corpus share of anglicism versus equivalent was computed per subcorpus to assess lexical competition.

4. Results

4.1. Overall diachronic growth (RQ1)

The aggregate anglicism rate rose from 18.4 per 10,000 tokens in 2010 to 56.7 in 2024—a 3.1-fold increase—with every biennial step significant at $p < 0.001$ (Figure 1). Growth was not linear: the curve steepens visibly after 2018, the period coinciding with mass smartphone penetration and the liberalisation-era expansion of the Uzbek internet ecosystem. Type diversity grew in parallel: 214 distinct anglicism lemmas were attested in the 2010 subcorpus against 693 in 2024, indicating that the increase reflects broadening lexical intake, not merely intensified use of a fixed set.

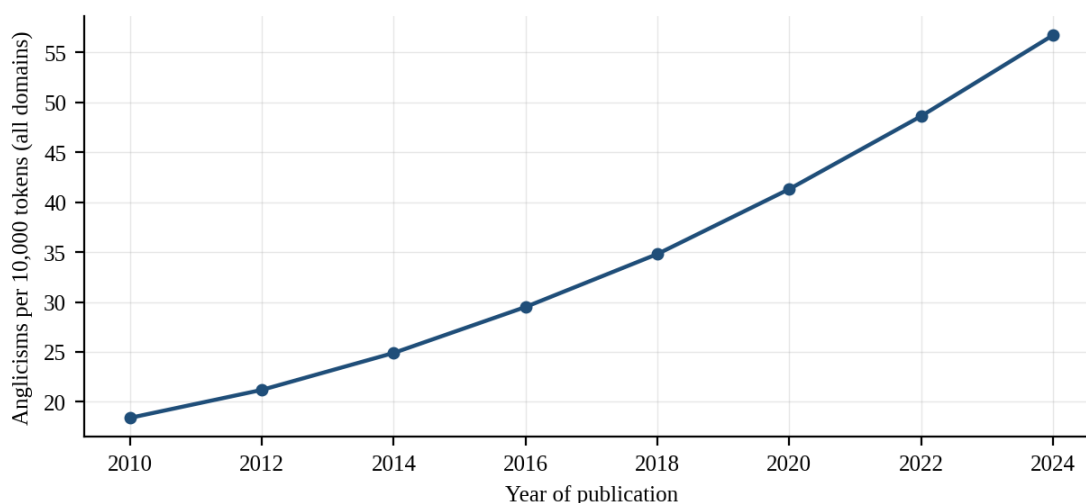


Figure 1. Anglicism rate in Uzbek online news, 2010–2024 (all domains, per 10,000 tokens).

4.2. Domain distribution

Domain stratification (Figure 2) shows the expected hierarchy at both endpoints—technology > economy > culture > sport—with technology reaching 84.3 per 10,000 by 2024 (from 31.2 in 2010). Proportional growth, however, was greatest in culture ($\times 3.0$) and sport ($\times 3.0$), demonstrating that anglicisms are diffusing beyond their technical beachhead into general-interest discourse; the culture domain’s 2024 intake is dominated by media-industry vocabulary (*kontent*, *strim*, *influenser*, *podkast*). The domain $\times$ period interaction is significant ($p < 0.001$), driven by technology’s steeper absolute trajectory.

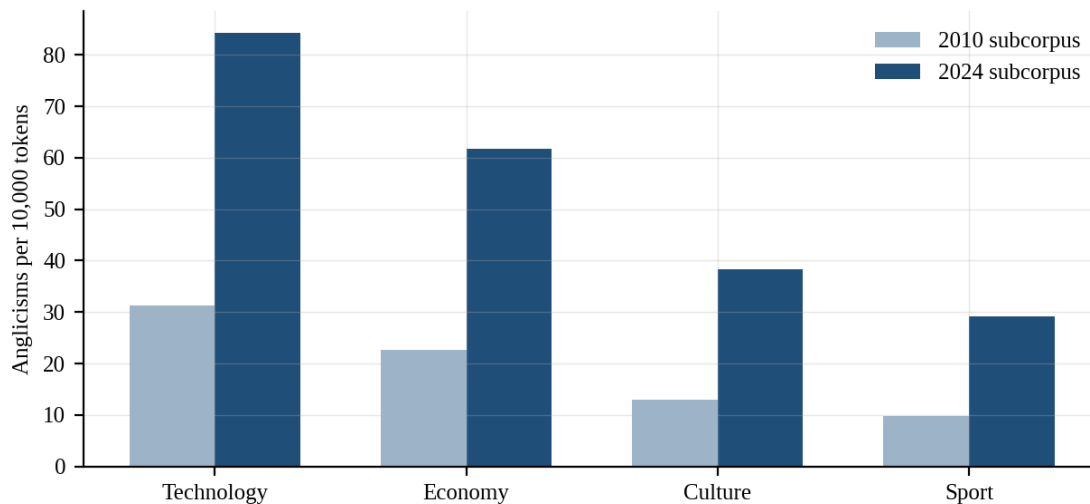


Figure 2. Anglicism rate by topical domain, 2010 versus 2024 subcorpora.

4.3. Adaptation and integration (RQ2)

Orthographic assimilation is the dominant strategy at 62.4% of anglicism tokens (Table 2): forms such as *onlayn* (online), *smartfon*, *brend*, and *reyting* respell the borrowing according to Uzbek Latin grapheme-phoneme conventions. Unassimilated transfers (24.1%; e.g. *startup*, *blockchain* in donor spelling) concentrate in the technology domain and in the most recent subcorpora, suggesting a typical trajectory in which items enter unadapted and assimilate as they conventionalise—supported by twelve items attested in both spellings whose assimilated variant’s share rises monotonically over time. Hybrids and loan translations (8.7%) include compounds like *media-makon* (‘media space’), while fully suffixed tokens (4.8% counted separately as integration evidence; e.g. *startaplarni* ACC.PL, *onlayn tarzda* ‘in online manner’) demonstrate unimpeded participation of borrowed stems in agglutinative morphology—the strongest structural indicator that these items function as Uzbek words rather than code-switches.

Table 2. Adaptation strategies (share of anglicism tokens, 2010–2024 pooled).

Strategy	Share (%)	Examples
Orthographic assimilation	62.4	onlayn, smartfon, brend, reyting
Unassimilated transfer	24.1	startup, blockchain, stream
Hybrid / loan translation	8.7	media-makon, raqamli kontent
Morphological integration (affixed)	4.8	startaplarni, smartfonlar

4.4. Lexical competition (RQ3)

Table 3 tracks five representative competition pairs. For eight of the fifteen pairs examined, the anglicism’s share of pair occurrences rose above 50% by 2024; the Russian-mediated stratum is the principal loser (e.g. *menejer* displacing *boshqaruvchi* in business contexts while itself continuing an earlier Russian-mediated form). In the remaining pairs the native or earlier item held its ground, typically

where it is short, morphologically productive, and register-neutral—conditions the borrowing literature identifies as protective [4, 10, 12].

Table 3. Selected competition pairs: anglicism share of pair tokens (%).

Anglicism	Established equivalent	2010	2024
onlayn	tarmoqdagi	34	88
kontent	mazmun (in media sense)	12	71
menejer	boshqaruvchi	41	66
reyting	nufuz ko'rsatkichi	55	83
xaridor*	shopper/beyer	91	78

**Pair where the native item retains majority share; the anglicism remains marginal.*

4.5. Qualitative patterns

Three qualitative observations complement the counts. First, semantic clustering: the 2024 intake concentrates in digital economy and media-professional fields (e-commerce, content production, fintech), while everyday domains that anchored earlier Russian-era borrowing (household technology, administration) contribute little new material—the borrowing wave tracks the economy's novel sectors rather than re-lexifying settled ones. Second, discourse functions: unassimilated anglicisms frequently appear with metalinguistic flagging in the earlier subcorpora—glossed in parentheses, introduced with *ya'ni* ('that is'), or quoted—whereas by 2024 the same items occur unflagged in headlines, the strongest register position; flagging rate for the twenty most frequent items fell from 31% of occurrences (2010–2014) to 4% (2022–2024), a measurable conventionalisation signal. Third, derivational productivity: borrowed stems increasingly feed Uzbek word formation beyond inflection, yielding agentives and denominal verbs (*bloggerlik* 'blogger-hood/profession', *reklamalashtirish* 'advertisement-isation'), forms that presuppose full lexemic status of the base.

5. Discussion

The quantitative picture is unambiguous: Uzbek online media are in a phase of rapid, accelerating lexical anglicisation whose 2024 rate (0.57% of running tokens) approaches values reported for well-studied European recipient languages in comparable registers [11, 12]. Three aspects of the pattern carry theoretical and regional significance. First, donor reorganisation: morphological and orthographic evidence shows English items now entering Uzbek directly—adapted to Uzbek Latin conventions without Russian phonological fingerprints that characterised earlier internationalisms—confirming the hypothesised post-Soviet re-routing of contact [5, 6]. The competition data sharpen this: even where the displaced item is itself a Russian-era borrowing, its replacement is an anglicism, indicating prestige realignment rather than purism-driven nativisation.

Second, the integration evidence argues against interpreting the phenomenon as superficial code-mixing. Agglutinative suffixation of borrowed stems, monotonic movement from donor spelling to assimilated spelling, and diffusion from technical into cultural domains together trace the classic conventionalisation pathway by which contact vocabulary becomes recipient-language vocabulary [1, 3, 10]. Uzbek's agglutinative typology appears to facilitate rather than impede intake: since borrowed stems require no inflectional-class assignment, the morphological cost of adoption is low—consistent with cross-linguistic findings on borrowability and morphological type [4, 16].

Third, the results furnish an empirical baseline for Uzbekistan's active language-policy debate. The data neither support alarmist claims of wholesale displacement (anglicisms remain a fraction of a percent of running text, and native items with favourable properties resist competition) nor the assumption that terminology commissions are steering intake: none of the fifteen pairs shows the officially recommended equivalent gaining share. Policy influence on media lexis, on this evidence, is marginal relative to the diffusion dynamics of the digital ecosystem [6, 14].

Interpreted against the necessity–luxury distinction [10, 12], the intake divides instructively. A majority of high-growth items name referents without established Uzbek designations at time of entry (*podkast, kripto, startup*): necessity borrowing accompanying referential novelty, the least ideologically fraught category. The competition pairs, by contrast, document luxury borrowing proper—adoption despite available means—and their trajectory shows prestige and register economy (anglicisms are typically shorter and connotatively “modern”) outcompeting availability. That both categories grow simultaneously suggests the driver is not lexical gap-filling alone but a broader reorientation of the media register toward a globalised stylistic norm, of the kind documented for other peripheral-language mediascapes [8, 9, 11].

The orthographic findings intersect with Uzbekistan’s unfinished script reform in ways that merit comment. Because the Latin standard has itself been revised repeatedly, incoming anglicisms face a moving assimilation target, and the corpus contains competing adapted spellings (*blogger/bloger/blogyer*) whose resolution tracks, with a lag, each official orthographic adjustment. Anglicisms thus function as a natural probe of orthographic norm diffusion: their spelling variance in a given year indexes how completely the current standard has propagated through editorial practice. For lexicography, the practical implication is that dictionaries of contemporary Uzbek cannot wait for full spelling stabilisation; frequency-based inclusion criteria of the kind used in this study offer a defensible interim policy, and the corpus’s top-frequency stratum—items like *onlayn, kontent*, and *smartfon* that now outnumber their alternatives severalfold—constitutes an evidence-based candidate list for codification.

Limitations include the register restriction to edited news (social media would likely show higher rates and more code-switching), the portal sample’s urban and national orientation, reliance on section metadata for domain assignment, and the exclusion of semantic borrowing (calqued meanings on native forms), which the loan-translation category captures only partially. The competition analysis, further, samples high-frequency pairs and cannot support inferences about the long tail.

6. Conclusions

Drawing on a 2.4-million-token diachronic corpus, this study provided the first systematic quantification of anglicisms in Uzbek online media: a trebling of normalised frequency between 2010 and 2024 (18.4 to 56.7 per 10,000 tokens), a domain hierarchy led by technology but with fastest proportional growth in cultural discourse, dominant orthographic assimilation (62.4%) with productive agglutinative integration, and majority displacement of established equivalents in eight of fifteen competition pairs. Uzbek has entered a direct-contact relationship with English in which the digital media register functions as the primary intake channel, and the borrowing wave shows the structural signatures of durable lexical change rather than transient fashion. Future work should extend the design to social-media Uzbek, incorporate phonological adaptation in spoken broadcasting, model item-level survival with mixed-effects diachronic methods, and compare parallel trajectories in Kazakh and Kyrgyz media to establish whether the dynamics documented here are language-specific or areal.

References

1. Haugen E. The analysis of linguistic borrowing. *Language*. 1950;26(2):210–231.
2. Weinreich U. *Languages in Contact: Findings and Problems*. The Hague: Mouton; 1953.
3. Winford D. *An Introduction to Contact Linguistics*. Oxford: Blackwell; 2003.
4. Haspelmath M, Tadmor U, editors. *Loanwords in the World’s Languages: A Comparative Handbook*. Berlin: De Gruyter Mouton; 2009.
5. Fierman W. *Language Planning and National Development: The Uzbek Experience*. Berlin: Mouton de Gruyter; 1991.

6. Landau JM, Kellner-Heinkele B. Politics of Language in the ex-Soviet Muslim States. London: Hurst; 2001.
7. Schlyter BN. Language policies in present-day Central Asia. *International Journal on Multicultural Societies*. 2001;3(2):127–136.
8. Androutsopoulos J. Mediatization and sociolinguistic change: Key concepts, research traditions, open issues. In: Androutsopoulos J, editor. *Mediatization and Sociolinguistic Change*. Berlin: De Gruyter; 2014. p. 3–48.
9. Zenner E, Backus A, Winter-Froemel E, editors. *Cognitive Contact Linguistics*. Berlin: De Gruyter Mouton; 2019.
10. Matras Y. *Language Contact*. Cambridge: Cambridge University Press; 2009.
11. Onysko A. Anglicisms in German: Borrowing, Lexical Productivity, and Written Codeswitching. Berlin: De Gruyter; 2007.
12. Görlach M, editor. *English in Europe*. Oxford: Oxford University Press; 2002.
13. Sjöberg AF. *Uzbek Structural Grammar*. Bloomington: Indiana University Press; 1963.
14. Dietrich A. Language policy and the status of Russian in the Soviet Union and the successor states outside the Russian Federation. *Australian Slavonic and East European Studies*. 2005;19(1–2):1–27.
15. Dunning T. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*. 1993;19(1):61–74.
16. Wohlgemuth J. *A Typology of Verbal Borrowings*. Berlin: Mouton de Gruyter; 2009.
17. Thomason SG. *Language Contact: An Introduction*. Edinburgh: Edinburgh University Press; 2001.
18. Poplack S, Sankoff D, Miller C. The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics*. 1988;26(1):47–104.
19. Muysken P. *Bilingual Speech: A Typology of Code-Mixing*. Cambridge: Cambridge University Press; 2000.
20. McEnery T, Hardie A. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press; 2012.
21. Baker P. *Using Corpora in Discourse Analysis*. London: Continuum; 2006.
22. Crystal D. *English as a Global Language*. 2nd ed. Cambridge: Cambridge University Press; 2003.
23. Boeschoten H. Uzbek. In: Johanson L, Csátó ÉÁ, editors. *The Turkic Languages*. London: Routledge; 1998. p. 357–378.