Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas

Eureka Journal of Language, Culture & Social Change (EJLCSC)

ISSN 2760-4926 (Online) Volume 2, Issue 2, February 2026

https://eurekaoa.com/index.php/3

LINGUISTIC MODELLING OF LOANWORDS IN UZBEK TEXT-TO-SPEECH SYSTEMS

Abduraxmanova Nazokat Azamjonovna
Doctoral Researcher (PhD Candidate) at Fergana State University
E-mail: nazokat.abduraxmanova21@gmail.com
F-ORCID ID 0009-0008-2456-1423

Abstract

This paper discusses how loanwords in Uzbek – especially Russian/international and Arabic – Persian strata – should be handled in the linguistic front-end of text-to-speech (TTS) systems, using evidence from the Uzbek electronic corpus. The study aims to identify and systematize orthography – pronunciation mismatches that cause typical synthesis errors: bi-phonemic graphemes (the Uzbek letter <j> encoding both affricate and fricative realizations), homography driven by the lack of distinct letters for foreign phonemes (e.g., the letter <o> in certain borrowed forms), vowel sequences in loanwords that are realized with inserted glides (y/v), and stress- and syllable-structure-dependent variants. The methodology is based on descriptive analysis, phonetic – morphological interpretation, identification and analysis of typical word combinations in the electronic corpus, as well as a comparative approach to practical experiences in Turkish, Kazakh, and Tatar languages. The results propose a set of practical front-end strategies: enriched pronunciation lexicon for loanwords, context-based homograph disambiguation, explicit glide-insertion rules for vowel sequences, and curated exception lists for fricative-j items. Overall, corpus-driven linguistic normalization is argued to be the key factor for improving naturalness and reducing errors in Uzbek TTS.

Keywords: Uzbek electronic corpus, loanwords, homograph, bi-phonemic grapheme, pronunciation variant, vowel sequence, morphophonology, linguistic front-end, text-to-speech (TTS).

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Language, Culture & Social Change (EJLCSC)
ISSN 2760-4926 (Online) Volume 2, Issue 2, February 2026

https://eurekaoa.com/index.php/3

Introduction

In Uzbek text-to-speech (TTS) systems, achieving a natural-sounding output often becomes problematic not at the level of the "voice model" itself, but rather at an earlier stage – the preparation of text for pronunciation. This stage is commonly referred to as the linguistic front-end, where textual units (words, abbreviations, numerals, terms, borrowings) are converted into their phonetic representations. Loanwords emerge as a particular challenge at this stage. Although their orthography may be adapted to the Uzbek writing system, they frequently reflect heterogeneous phonetic layers. As a result, discrepancies arise between spelling and actual pronunciation, complicating the grapheme-to-phoneme conversion process and, consequently, affecting the naturalness of synthesized speech.

The article focuses on three major issues related to loanwords as central objects of analysis: (1) grapheme – phoneme mismatch (for example, the representation of two distinct phonemes by the single letter *j*), (2) homography (units that share identical orthographic forms but differ in pronunciation and meaning), and (3) vowel sequences and stress-related variation in borrowed terminology. Following the discussion of each issue, its implications for TTS systems are examined. In particular, possible solutions are evaluated – whether rule-based approaches, lexical lists, or context-sensitive analysis – with reference to evidence derived from electronic corpus data.

Literature Review

In Turkic languages, the linguistic foundation of TTS systems often rests on the integration of morphological analysis and a pronunciation lexicon. In Turkish, the pronunciation lexicon has been developed in conjunction with a two-level morphological analyzer, demonstrating that the output of morphological analysis can directly inform pronunciation modeling (Oflazer, 1992). In rule-based

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Language, Culture & Social Change (EJLCSC)
ISSN 2760-4926 (Online) Volume 2, Issue 2, February 2026

https://eurekaoa.com/index.php/3

Turkish grapheme-to-phoneme (G2P) approaches, exception lists and specialized rules for loanwords play a crucial role (Altınok, 2016).

In Kazakh, TTS studies based on open-access datasets (e.g., *KazakhTTS*) identify the modeling of the loanword layer-particularly Russian-derived phonemes and stress assignment patterns–as a distinct challenge (Khassanov et al., 2021). Similarly, in Tatar, there have been efforts to develop automatic phonetic transcription based on G2P rules and to integrate these models into speech technologies (*TatarTTS*, n.d.). These experiences demonstrate that loanwords cannot be fully accounted for by "regular" orthographic rules alone; rather, a lexicon- and corpus-based exception layer is essential.

In studies on the Uzbek language, the electronic corpus is interpreted as an important linguistic resource that reflects the real-life usage of language units in authentic speech (Abdurakhmonova et al., 2021). The corpus text base enables the observation of word and word-form frequencies, contextual features, and variation patterns. As such, it serves not only as a foundation for theoretical research but also as a valuable resource for applied speech technologies.

In this process, morphological analysis tools assume particular importance. Specifically, analyzers such as *MorphUz* enable the segmentation of words into stems and affixes and the identification of their grammatical features (tags), thereby revealing the internal structure of linguistic units. This not only facilitates the systematic description of corpus materials but also provides a foundation for making informed linguistic decisions in speech technologies (Abdurakhmonova et al., 2022). In explaining the relationship between orthography and pronunciation, theoretical perspectives from phonetics and graphemics serve as essential sources. Classical studies have extensively examined the phenomena of homography, biphonemic graphemes, and pronunciation variants (Rahmatullayev, 2006; Mirtojiyev, 2013), offering a scientific basis for interpreting discrepancies between written form and spoken realization.

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Language, Culture & Social Change (EJLCSC)
ISSN 2760-4926 (Online) Volume 2, Issue 2, February 2026

https://eurekaoa.com/index.php/3

Methodology

The study draws on three types of data: (a) electronic corpus texts, (b) phonetic and grammatical rules, and (c) prior experience in speech technologies for Turkic languages.

The primary units of analysis include loanwords (Russian-international and Arabic-Persian origin), their homographic forms, grapheme–phoneme inconsistencies related to the letters *j* and *o*, vowel sequences (structures prone to diphthongization), and stress-related variation.

The following methods were employed: descriptive phonetic analysis (articulatory description of sounds and their positional variation); lexical-semantic analysis (distinguishing homonyms and homographs); corpus-based collocational analysis (identifying typical lexical neighbors of a given word); a comparative approach (examining how loanwords are "covered" in Turkish, Kazakh, and Tatar speech technology practices).

Results and Discussion

1. The Grapheme "J" and Biphonemicity: the Fricative *j* in Loanwords

In Uzbek orthography, the letter *j* encodes two distinct consonants: (I) the affricate /dʒ/ (as in *juda*, *jo'ra*), and (II) the fricative /ʒ/ (as in Russian-international loanwords such as *jurnal*, *janr*, *drenaj*). In graphemic theory, this is described as a biphonemic grapheme (Rahmatullayev, 2006:118). As a result, TTS systems relying on letter-by-letter reading tend to process *jurnal*-type words in the same way as *juda*-type words, which leads to artificial pronunciation.

TTS solution (corpus-based): Loanwords containing the fricative *j* are separated into a "special pronunciation list." This list can be automatically extracted from a corpus by identifying: (a) words beginning with *j* and ending in Russian/international morphemes (e.g., *-aj*, *-janr*, *-jurnal*); (b) typical segments such as *-aj* (*drenaj*, *peyzaj*). Corpus frequency helps prioritize and update the list.

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Language, Culture & Social Change (EJLCSC)
ISSN 2760-4926 (Online) Volume 2, Issue 2, February 2026

https://eurekaoa.com/index.php/3

Thus, when the front-end encounters *jurnal*, it selects the fricative pronunciation variant, whereas *juda* remains governed by the general rule.

2. The Grapheme "O" and Homography: the Case of *tom*

Homography in Uzbek often arises from the absence of distinct letters for certain phonetic distinctions. For example, *tom* ("roof of a house") and *tom* ("volume of a book") are orthographically identical but differ in pronunciation and etymology. Rahmatullayev notes that homographs may arise artificially due to the representation of different phonetic values by the same grapheme, particularly *o* (Rahmatullayev, 2006:142).

TTS solution (corpus-based): The simplest and most effective disambiguation strategy is contextual analysis. In corpus data, collocations such as "uyning tomi" clearly differ from "I-tom," "II-tom," or "kitobning … tomi." A practical front-end rule may be formulated as follows: if *tom* co-occurs with ordinal markers (I, II, 1-, 2-) or indicator words such as *kitob* ("book") or *jild* ("volume"), the loanword pronunciation variant is selected; otherwise, the native-layer pronunciation is used. The list of contextual indicators can be statistically derived from corpus collocation frequencies.

3. "Hard-Soft" Vowel Variants and Loanwords

Certain Uzbek vowels display positional variation depending on the surrounding consonantal environment. For example, the vowel *o‘* is realized as relatively fronted after *k, g, y, h* but more back after *q, g‘, x* (*ko‘l – qo‘l; go‘r – g‘o‘r; ho‘l – xo‘r*) (Mirtojiyev, 2013:74). Similar positional variants are observed for *a, i,* and *u*.

In loanwords, however, additional shifts occur, such as the backing of *a* after *k/g* (*karta, katalog*), the approximation of *o* toward *o‘* (as in *tom* 'volume'), and softened realizations of *u*.

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Language, Culture & Social Change (EJLCSC)
ISSN 2760-4926 (Online) Volume 2, Issue 2, February 2026

https://eurekaoa.com/index.php/3

TTS solution: This phenomenon cannot be handled solely by simple letter-to-sound rules. Instead, two decision-making sources are required: A loanword pronunciation lexicon (e.g., *karta*, *katalog* stored with specified variants); Corpus-based stylistic tagging. Since loanword terminology is more frequent in scientific and formal registers, stylistic inference (e.g., the presence of affixoids such as *bio-*, *avto-*, *-logiya*) can guide pronunciation selection. Thus, the front-end regulates the "probability of backing" through lexical and stylistic signals.

4. Vowel Sequences and Glide Insertion (y/v)

Loanword terminology frequently contains adjacent vowels (*aero-*, *audio-*, *arxaik*, *teatr*, *biologiya*, *duet*, etc.). Mirtojiyev notes that Uzbek phonotactics generally disfavor consecutive vowels; in speech, such sequences often undergo articulatory compression, becoming prone to diphthongization. In practical pronunciation, glide insertion occurs: *aerodrom* → *ayrodrom, auditoriya* → *avditoriya, teatr* → *teyatr, biologiya* → *biyologiya* (Mirtojiyev, 2013:96).

TTS solution: A rule-based block for vowel sequences is introduced: For V1+V2 sequences (ae, ai, ea, eo, ia, io, iu, oi, ue, ui), test *y*-insertion; For *ao, au*, test *v*-insertion. However, the rule is not universally applicable. The corpus therefore: a) identifies dominant variants via frequency; b) generates exception lists. Thus, if corpus evidence shows that *teyatr* predominates in speech, the front-end approximates that variant; otherwise, it preserves the standard orthographic realization.

5. Loanwords and Morphology: Pronunciation Stability under Affixation

Loanwords frequently accept Uzbek affixes (*tomlar, kataloglar, jurnalda*), yet the stability of their phonological base varies. The challenge lies in whether the front-end recognizes the stem as loanword-derived. If not explicitly marked, general phonetic rules may overapply and produce artificial forms.

TTS solution: The morphological analyzer provides the lemma and affix sequence. If the lemma is tagged as "loanword" versus "native layer," phonetic rules are applied differentially. In Turkish, the integration of pronunciation

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Language, Culture & Social Change (EJLCSC)
ISSN 2760-4926 (Online) Volume 2, Issue 2, February 2026

https://eurekaoa.com/index.php/3

lexicon and morphological analyzer – particularly within a two-level morphological model – has demonstrated that morphological disambiguation can stabilize pronunciation. A similar solution is feasible for Uzbek: automatic proposal of loanword tags from corpus-based terminological signals (*-logiya, -izm, bio-, avto-*), followed by manual validation and lexicon integration.

6. Experience of Related Turkic Languages

In Turkish, despite the extensive loanword layer, pronunciation modeling relies on an "rule + exception list" principle, supported by strong morphological analysis. Oflazer's two-level model demonstrates that automatic morphological parsing contributes to pronunciation stability. Altınok emphasizes the necessity of exception lists and normalization rules for loanwords in rule-based Turkish G2P systems.

In Kazakh, Russian loanwords (especially toponyms and technical terms) exhibit stress and phonotactic deviations. Open resources such as KazakhTTS highlight text–pronunciation alignment as a core issue.

In Tatar, G2P-based automatic phonetic transcription is considered a central component of speech technology pipelines, and corpus-labeled audio data enrich pronunciation lexicons. Across Turkic languages, the general solution converges on a triad: rules + exception lists + corpus-based validation and updating.

Conclusion

Loanwords constitute a critical observation point in the linguistic front-end of Uzbek TTS, where orthography-pronunciation mismatches become particularly evident. The findings indicate that:

1. Biphonemic graphemes (notably *j*) require dedicated pronunciation lists continuously updated via corpus evidence.
2. Homographs (e.g., *tom*) can be effectively disambiguated through statistically derived contextual indicators integrated into simple front-end rules.

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Language, Culture & Social Change (EJLCSC)
ISSN 2760-4926 (Online) Volume 2, Issue 2, February 2026

https://eurekaoa.com/index.php/3

3. Vowel sequences and glide insertion phenomena can be rule-governed, but corpus data are decisive in identifying exceptions and dominant variants.
4. The experience of related Turkic languages confirms the necessity of a "rule + lexicon + corpus" triad for systematic loanword coverage.

Accordingly, corpus-based linguistic normalization and systematic enrichment of the pronunciation lexicon represent the most effective pathway toward improving the naturalness of Uzbek TTS, particularly in stabilizing the synthesis of loanword vocabulary.

References

1. Altınok, D. (2016). Towards Turkish ASR: Anatomy of a rule-based Turkish g2p. arXiv. https://arxiv.org/abs/1601.03783
2. Oflazer, K. (1992). Two-level description of Turkish morphology. In Proceedings of COLING 1992 (pp. 1-6).
3. Yuret, D., & Türe, F. (2006). Learning morphological disambiguation rules for Turkish. In Proceedings of NAACL-HLT 2006 (pp. 328-334). Association for Computational Linguistics. https://doi.org/10.3115/1220835.1220877
4. Khassanov, Y., et al. (2021). KazakhTTS: An open-source Kazakh text-to-speech synthesis dataset. In Proceedings of Interspeech 2021 (pp. 440-444). ISCA. https://doi.org/10.21437/Interspeech.2021-93
5. (Dataset) TatarTTS: Open-source text-to-speech dataset for the Tatar language. (n.d.). Hugging Face Datasets. https://huggingface.co/datasets
6. Abdurakhmonova, N., Tuliyev, U., & Gatiatullin, A. (2021). Linguistic functionality of Uzbek Electron Corpus: uzbekcorpus.uz. In 2021 International Conference on Information Science and Communications Technologies (ICISCT) (pp. 1-4). IEEE.
7. Abdurakhmonova, N., Alisher, I., & Sayfulleyeva, R. (2022). MorphUz: Morphological analyzer for the Uzbek language. In 2022 7th International

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
OPEN ACCESS JOURNALS

Eureka Journal of Language, Culture & Social Change (EJLCSC)

ISSN 2760-4926 (Online) Volume 2, Issue 2, February 2026

https://eurekaoa.com/index.php/3

Conference on Computer Science and Engineering (UBMK) (pp. 61-66). IEEE.

8. Rahmatullayev, S. (2006). Hozirgi adabiy o‘zbek tili. Toshkent: Universitet.
9. Mirtojiyev, M. M. (2013). O‘zbek tili fonetikasi. Toshkent: Universitet.
10. Tog‘ayev I. B. (2025). Computer modeling of phonetic and orthographic transliteration of loanwords in the Uzbek language. Development of Science. https://devos.uz/article.php?id=1844
11. Hamroyeva S. M., & Makhmudjonova G. U. (2025). The significance of G2P models for the low-resource Uzbek language. Қ. Жұбанов атындағы Ақтөбе өңірлік университетінің хабаршысы, 2(80).