An Improved Karlin Model Fit Test: Application to English and Uzbek Texts and Challenges

Shahzod Fayzullaev[1,2*]

[1] Novosibirsk State University, Novosibirsk, Russia
[2] Urgench State University, Urgench, Uzbekistan
[*] Corresponding author's email: shahzodjohn1@gmail.com

DOI: https://doi.org/10.53482/2026_60_430

ABSTRACT

Zipf's law and similar frequency laws have been studied in many languages, but their behavior in Uzbek has not been investigated. In this paper, we examine the fit of the generalized Karlin model of Zipf's law to Uzbek texts. For 386 texts consisting of three different genres (prose, poetry, and newspapers), we compute the T_n and H_n statistics and their analogs in the first half of the text and propose a new goodness-of-fit statistic Q_n based on their joint asymptotic behaviour. Our results show that model fit varies systematically with text length and genre: newspapers and poetry, which are typically shorter and more thematically compact, fit the model significantly better than long narrative prose. These results clarify how the Karlin model works in Uzbek texts and provide an empirical baseline for future comparative studies of understudied languages, including agglutinative ones. We also compare the new statistic with two previously proposed tests, based on type and hapax, on ten English reference texts. The results show that Q_n produces virtually identical p-values to the second test and leads to the same accept/reject decisions at standard significance levels, while the first test is systematically more conservative.

Keywords: Karlin model, vocabulary growth, hapax legomena, goodness-of-fit, quantitative linguistics, Uzbek texts

1 Introduction

Hapax legomena are words that appear in a text only once. This term comes from ancient Greek and means "said once". In linguistics, the number of hapax legomena is an important indicator that reflects the stylistic features of the author, the complexity of the text or the structure of the language. This term was first introduced into scientific circulation by Harrison (1921), and it was emphasized that it determines the style of the author. An infinite urn scheme is a probabilistic model in which n balls are thrown into an infinite box with probability p_k, independently and with the same distribution. This scheme, proposed by Karlin (1967), is applied in linguistics as follows: we have a text corpus of size n, consecutive words in the text are like balls, and they correspond to one of the words in an infinite dictionary with probability p_k,

just as a ball falls into one of the infinite boxes. The number of words that occur only once in the text, i.e. hapaxes, is defined as the number of boxes containing only one ball after throwing n balls, and the number of non-empty boxes is defined as the number of different words in the text, which is usually called the number of types or simply types. Petrini et al. (2023) investigated the relationship between word frequency and word length in a wide range of languages. They analyzed the average word length L and showed that more frequent words tend to be shorter, in accordance with Zipf's law of contraction. They also introduced a random baseline L_r and showed that the relation $L < L_r$ holds consistently across languages, which they interpret as direct evidence of compression in human language. Galieva and Vavilova (2021) analyzed the syllable structure in Tatar fiction by comparing the first (main) and last (affixal zone) syllables of polysyllabic words. Using χ^2 tests, they found a highly significant discrepancy: simple CV syllables dominate at the beginning of words, while sonorant syllables (SV, SVS, etc.) predominate at the end of words. Tuzzi et al. (2009) statistically confirmed the validity of Zipf's law on a corpus of New Year's addresses of Italian presidents; the model worked well even when the same texts were sometimes written by different authors. Abebe et al. (2024) proposed a word change point method based on the urn model that counts forward and backward different numbers of words to determine the subject/author change point in a text, demonstrating theoretical consistency and performance with an error rate of $< 3\%$ for English and multilingual corpora. And in a recent study, Abebe (2025) developed a new algorithm to detect two change points in a linked text consisting of three different texts using the same probabilistic models as in the previously mentioned study. Kudryavtseva and Kovalevskii (2025) compared AI-generated texts with human-written texts based on word frequency, the Zipf distribution, and the hapax legomena coefficient. The study found that AI-generated texts had a significantly smaller vocabulary and a significantly lower proportion of rare words. Popescu and Altmann (2008) showed that in a large cross-linguistic sample the proportion of hapax legomena in a text relative to types fluctuates around a nearly fixed value. Another study of the Menzerath-Altmann law by Mačutek et al. (2026) found that while most languages show an inverse relationship between word length and the average syllable length of that word, some languages are exceptions to this rule. We refer you to a study conducted in the field of cognitive linguistics Bao et al. (2025), the results of which show that the level-frequency distribution of DM (discourse marker) satisfies the Zipf-Alexeev law at the corpus and text levels. The study reviewed in Milička (2009) studies the type-token relationship as a property of the text rather than as a property of a particular language, and proposes a model based on a combinatorial description of the distribution of different types in the text. Davis (2018) reexamined the relationship between Zipf's and Heaps' laws and proposed a completely new, logarithmic model for the type-token curve. The study showed that smaller texts do not conform to Zipf's law, and the law becomes more accurate as the text size increases. However, it becomes less accurate with larger text sizes, and it is possible to identify an ideal corpus that conforms to Zipf's law for

larger texts. Uzbek language has an agglutinative structure, and words are formed by adding together successive morphemes. This leads to the formation of numerous variants for a single lemma. A large number of morphological variants increases the number of hapax legomena and the number of different words (types), which significantly affects the robustness of statistical models such as Zipf and Heaps. These statistical models have not yet been tested in Uzbek, and little research has been conducted in other agglutinative languages. In this study, the Karlin model is applied to Uzbek texts for the first time and analyze original, lemmatized, and stemmed versions of the corpus separately in order to examine how morphological normalization affects the model fit.

More specifically, this paper's contribution is threefold. First, we introduce a new fit statistic, Q_n, for the Karlin infinite urn model, derived from the joint asymptotic behavior of the number of types and hapax legomena at positions n and $n/2$.

Second, we apply this test to a large corpus of 386 Uzbek texts (prose, poetry, and newspapers) in three versions (original, lemmatized, and stemmed) and analyze how the fit of the model depends on genre and text length. Third, we compare Q_n with two previously proposed tests, type-based and hapax-based, on ten English reference texts and show that Q_n behaves very similarly to $Q_n^{(2)}$, while $Q_n^{(1)}$ acts as a more conservative criterion.

2 Material

This study analyzes 386 texts in three different genres:

1) prose (158 works and stories);

2) poetry (167 poems and poetry collections); and

3) various newspapers (61 issues).

2.1 Sources of texts

Uzbek prose and poetry were downloaded from the publicly available Ziyouz (https://www.ziyouz.com) website. Issues of the *Xalq so'zi* and *Yangi Uzbekistan* newspapers were downloaded from the official websites of https://xs.uz and https://yuz.uz, respectively, and issues of the https://press.natlib.uz/ru newspaper "O'zbekiston ovozi" were downloaded from the electronic archive of national publications of the National Library of Uzbekistan (https://press.natlib.uz/ru).

2.2 Normalization and transliteration

All texts were processed using a special program written in Python before lemmatization and stemming: These steps are performed as follows: All Unicode apostrophes (', ', ‘, ‘) were replaced with the "‘"

symbol. Although UzbekLemma is generally resistant to apostrophe changes, accurate normalization prevents an artificial increase in the number of hapaxes and types in the text. All punctuation marks, with the exception of Uzbek apostrophes, were removed. Page numbers and other layout-related numeric designations were removed, but numeric expressions that were part of the current text (e.g., "2025-yil," "16-bob") were retained as tokens. Proper nouns (personal names, toponyms, and organizational names) and loanwords were not filtered out and were counted as ordinary word types.

2.3 Evaluation of Lemmatization and Stemming Tools

Transliteration of Cyrillic texts into Latin was performed using the *UzTransliterator* Python library. This study additionally tested two existing tools – the *UzbekLemma* lemmatizer and the *uznltk* stemmer – to assess the impact of lemmatization and stemming on the model's results. For this purpose, a small test list of 500 Uzbek words from different categories was compiled, and the results of each tool were manually verified.

According to the evaluation results:

- Overall accuracy of *UzbekLemma*: 73.60 %
- Stemming accuracy of *uznltk*: 87.17 %

Lemmatization errors were observed primarily in the following cases:

- homonyms with similar verb and noun forms,
- verb forms with multiple suffixes (kelayotganimizni, borilmaydi, etc.).

Nevertheless, in many cases, the tools correctly normalize words. To illustrate how lemmatization and stemming behave in practice, Table 1 presents several representative examples from a list of tests.

Table 1: Examples of lemmatization and stemming outputs for Uzbek words. Lemmatization preserves the dictionary form, while stemming applies more aggressive affix removal.

Original word	Lemma (UzbekLemma)	Stem (uznltk)
yil	yil	yil
yashil	yashil	yashil
deputatlar	deputat	deputat
tadbirda	tadbir	tadbir
qiladigan	qilmoq	qil

These results show that the tools perform well with common nouns, adjectives, and regular affixed forms. Despite these errors, they do not materially change the main statistical indicators of the analysis. This is because an incorrectly lemmatized word is often transformed into the same normalized form as its other

variants. Even if this is technically an incorrect labeling, it still combines several surface forms under a single token. In most cases, this reduces excessive dispersion between different word forms and stabilizes the number of types and hapaxes. Thus, although the above percentages indicate imperfections in the tools, the observed inaccuracies do not lead to statistically significant biases in the analysis.

3 Methodology

Let $\{X_n\}_{n=1}^{\infty}$ denote random variables representing consecutive words in a text that satisfy Karlin's elementary probability model, i.e.

$$p_i = \mathbb{P}(X_1 = i) = l(\theta, i)\, i^{-1/\theta}, \quad i \geq 1, \tag{1}$$

where $l(\alpha, i)$ is a slowly varying function, θ is an unknown parameter, and $0 < \theta < 1$.

We introduce the following definitions:

- n – the total number of words in the corpus under consideration, or, in other words, tokens.
- T_n, $T_{[n/2]}$ – types in the text and in the first half of the text, respectively.
- H_n, $H_{[n/2]}$ – hapaxes in the text and in the first half of the text, respectively.
- $\widehat{\theta} = H_n/T_n$ – an estimate of the parameter θ proposed by Chebunin and Kovalevskii (2019) and proven to be strongly consistent.
- τ_n and η_n – expected value of T_n and H_n, respectively

According to Karlin (1967), we have the law of large numbers for T_n and H_n, as well as the asymptotics for τ_n and η_n as follows:

$$\tau_n \sim \Gamma(1-\theta) l(\theta, n) n^{\theta}, \quad \eta_n \sim \theta\Gamma(1-\theta) l(\theta, n) n^{\theta}, \quad \text{as} \quad n \to \infty. \tag{2}$$

The functional central limit theorem, proven by Chebunin and Kovalevskii (2016), forms the basis of our next theorem. We will construct two processes as follows:

$$Y_n(t) = \left(T_{[nt]} - \tau_{[nt]}\right)/\sqrt{\tau_n}, \qquad Z_n(t) = \left(H_{[nt]} - \eta_{[nt]}\right)/\sqrt{\tau_n}.$$

where $t \in [0, 1]$.

In Karlin's model, both the number of types T_n and the number of hapax legomena H_n grow proportionally to n^θ, and the ratio H_n/T_n approaches θ Chebunin and Kovalevskii (2019) and Karlin (1967). Therefore, differences in their growth rates serve as a sensitive indicator of deviations from the model. Comparing the first half of the text $(T_{[n/2]}, H_{[n/2]})$ with the full text (T_n, H_n) provides a practical way to identify local asymmetries in vocabulary development.

Theorem. Suppose that the discrete probabilistic model given by (1) satisfies the condition

$$p_i = ci^{-1/\theta}(1 + o(i^{-1/2})) \quad \text{as } i \to \infty, \tag{3}$$

where $c > 0$ and $0 < \theta < 1$. Then, for the statistic

$$V_n = \frac{\begin{vmatrix} T_n & H_n \\ T_{[n/2]} & H_{[n/2]} \end{vmatrix}}{T_n^{3/2}},$$

there is weak convergence to a centered normal random variable:

$$V_n \xrightarrow{d} V = Y_1(1/2) - 2^{-\theta} Y_1(1) - \theta\, Y(1/2) + \theta\, 2^{-\theta} Y(1),$$

with zero expectation and variance is determined by the expression

$$\mathrm{Var}(V) = \Sigma^2(\theta) = v(\theta)\, G(\theta)\, v^\top(\theta),$$

where $v(\theta) = (\theta 2^{-\theta}, -\theta, -2^{-\theta}, 1)$ and $G(\theta)$ is a (4×4) matrix from Lemma 1 in Fayzullaev and Kovalevskii (2024).

The proof of the theorem is in the appendix.

Based on the above theorem, we construct statistics to test the elementary probability model in the form of a p-value $= 2\Phi^{-1}(-|Q_n|)$, where $Q_n = V_n/\Sigma(\widehat{\theta})$ and Φ^{-1} is the quantile function of the standard normal distribution. Since $\widehat{\theta}$ actually depends on n, we can redefine $\Sigma(\widehat{\theta})$ as $\Sigma(\widehat{\theta}) := \Sigma_n(\widehat{\theta})$, and in this case $Q_n = V_n/\Sigma_n(\widehat{\theta})$. If p-value $\geq \varepsilon$, we fail to reject the hypothesis that the elementary probability model corresponds to the text at the significance level ε; otherwise, we reject it.

4 Results

The corpus consists of 386 texts across three genres: 158 prose works (novels and short stories), 167 poetry collections and individual poems, and 61 newspaper articles. For each text, we calculated the values

of n, T_n, $T_{[n/2]}$, H_n, $H_{[n/2]}$, the estimate $\widehat{\theta} = H_n/T_n$, the standardized statistic Q_n, and the corresponding p-value. A link to the full table is provided in the *Data Availability Statement* section, and the original texts are available via the links in the *Materials* section. In addition to analyzing the original texts, we repeated all calculations on two preprocessed versions of the corpus: lemmatized (UzbekLemma) and stemmed (uznltk). This allows us to test the robustness of the inferences to morphological normalization. Throughout Section 4, we use the 10% significance level ($\varepsilon = 0.1$) as a convenient reference threshold for descriptive summaries. For completeness, Figure 4 also indicates the more common levels 0.05 and 0.01; using these stricter thresholds does not change the qualitative conclusions.

4.1 Genre Patterns

At the level of the original texts, the proportions with p-value < 0.1 are as follows:

- newspapers: 8 out of 61 issues;
- poetry: 29 out of 167 texts;
- prose: 108 out of 158 texts.

Thus, most newspaper issues and poems fit Karlin's model well, while approximately 70% of prose works have p-value < 0.1 and therefore demonstrate a weaker fit.

A similar analysis of the lemmatized corpus yields only minor changes in these proportions: 8 newspaper issues, 41 poetry texts, and 99 prose texts have p-value < 0.1. For the stemmed text corpus, we again obtain 8 newspaper issues with p-values < 0.1, while for poetry and prose these figures become 27 and 64, respectively. In all three cases, newspapers and poetry remain the genres with the best overall fit, while prose systematically shows the worst fit.

For newspapers, the classification into "good fit" and "poor fit" is very stable across all three preprocessing options. Of the 61 issues, 49 have p-values ≥ 0.1 in all three versions (the original, lemmatized, and stemmed versions), and 5 issues have p-values < 0.1 in all three versions. Only a small number of borderline cases (with p-value close to 0.1) change their classification when lemmatization or stemming is applied. This shows that the observed differences between texts and genres are due to genuine structural properties of the texts and not to pre-processing stages.

4.2 Examples of poor and good fit in prose

Within the prose subcorpus, the set of texts that fit Karlin's model well, and those that clearly violate it, is remarkably stable across all three preprocessing options (original, lemmatized, and stemmed texts). In particular, several long works yield very low p-values across all three options, while a small group

of shorter prose texts consistently exhibit relatively high p-values. Only a few borderline cases change their classification when lemmatization or stemming is applied. A typical example of a persistent poor match is Oybek's novel "Qutlug' qon". In this novel, p-value in the original version is practically zero (p-value $< 10^{-4}$), and remains very small after lemmatization (p-value ≈ 0.0005) and stemming (p-value ≈ 0.004). As shown in Figure 1, the curves T_k and H_k generally retain a power-law shape, but the number of hapax legomena increases systematically faster than the theoretical prediction $\widehat{\theta} T_k$, especially in the first half of the text. A similar pattern is observed in Risolat Khaidarova's novel "Javzo", where the p-value remains below 0.01 in all three text variants (approximately 0.0002 for the original, 0.0006 for the lemmatized version, and 0.0025 for the stemmed version). These works belong to a small group of seven long prose texts for which the test statistics systematically reject Karlin's model, regardless of whether we analyze the original tokens, lemmas, or stems. This indicates that the rejection is due to the texts' genuine structural properties (length, theme shifts, richness of descriptive fragments), rather than to the peculiarities of morphological preprocessing.

On the other hand, there are also prose works that conform reasonably well to the model in all three variants. For example, in the collections "Sobiq o'g'ri" by Said Ahmad and "Mezon" by Shomirza Turdimov, the p-value exceeds 0.5 for both the original texts and their lemmatized and stemmed versions (for Sobiq o'gri, the p-value is about 0.67 for the original text, 0.58 for the lemmatized version, and 0.75 for the stemmed version; for Mezon, the corresponding p-values are about 0.68, 0.67, and 0.52). In these texts, after a short initial phase, the growth in the number of hapax legomena becomes roughly proportional to the growth in the number of types. For the work "Sobiq o'g'ri", Figure 2 shows that H_k closely follows the theoretical curve $\widehat{\theta} T_k$, while Figure 3 shows that the process $D_k = H_k - \widehat{\theta} T_k$ oscillates around zero without a clear trend. This behavior is typical for works with relatively homogeneous vocabulary and a high degree of repetition, for example, due to frequent dialogue or repetitive narrative patterns.

However, there are several borderline cases where morphological normalization changes the test result. For example, in Hamid G'ulom's novel "Qoradaryo", the p-value for the original text is slightly below 0.1 (around 0.067), while for the lemmatized and stemmed versions it increases to p-values of around 0.77 and 0.84, respectively. Here, lemmatization and stemming remove some of the superficial morphological variability and make the trajectories of T_k and H_k more regular, so the text is classified as compatible with Karlin's model after normalization. However, such cases are rare compared to the highly consistent and highly inconsistent texts discussed above. Overall, these examples demonstrate that long prose works tend to present the greatest challenge to Karlin's elementary probabilistic model. If the text contains strong internal variation, shifting themes, abrupt topic shifts, and bursts of rare words, the test rejects the model in all three preprocessing modes. Conversely, if the prose text has a more homogeneous vocabulary and a

high degree of repetition, the model can provide a good approximation, and this result remains stable under lemmatization and stemming.

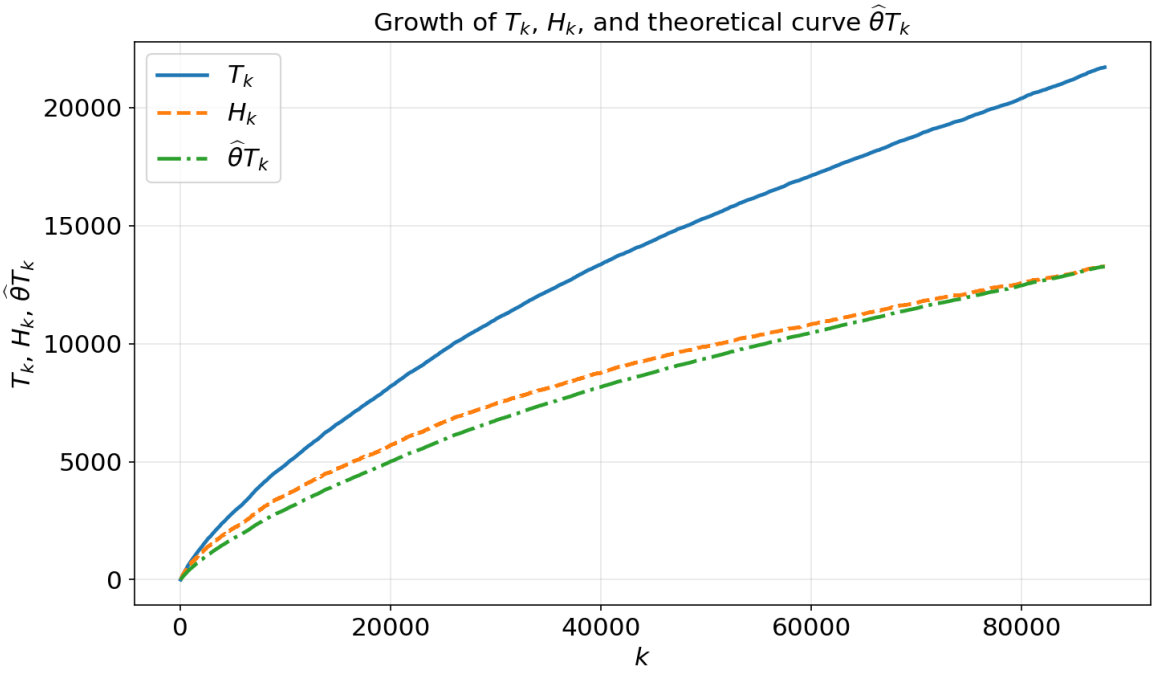

Figure 1: Growth of T_k, H_k, and the theoretical curve $\widehat{\theta T}_k$ in Oybek's novel *Qutlug' qon* (original text). The curve for H_k grows systematically faster than $\widehat{\theta T}_k$, which leads to a very small *p-value*.

4.3 Newspapers and Poetry

For newspapers and poetry, Karlin's model performs significantly better than for prose. In the original newspaper texts, 53 of 61 issues (about 87%) have a p-value ≥ 0.1, and 32 issues (more than half the sample) have a p-value ≥ 0.5. The average p-value in this subcorpus is about 0.52. For poetry, 138 of 167 texts (about 83%) have a p-value ≥ 0.1, and 39 poems achieve a p-value ≥ 0.5, with an average of about 0.34. Thus, shorter and more thematically compact texts tend to demonstrate significantly better matches than longer prose works.

Lemmatization and stemming do not change these findings. In all three versions of the newspaper corpus (original, lemmatized, and stemmed), exactly 8 issues have p-value < 0.1, while the remaining 53 issues remain above 0.1. Five issues are classified as "poorly conforming" in all three versions, and 49 are classified as "well conforming" in all three, so only a small group of seven edge cases change status depending on pre-processing. A similar pattern is observed for poetry: in the original texts, 29 of 167 poems have a p-value < 0.1, while in the lemmatized corpus, 41 poems fall below this threshold, and in the stemmed corpus, 27 do so. However, 15 poems consistently demonstrate a p-value < 0.1 in all three

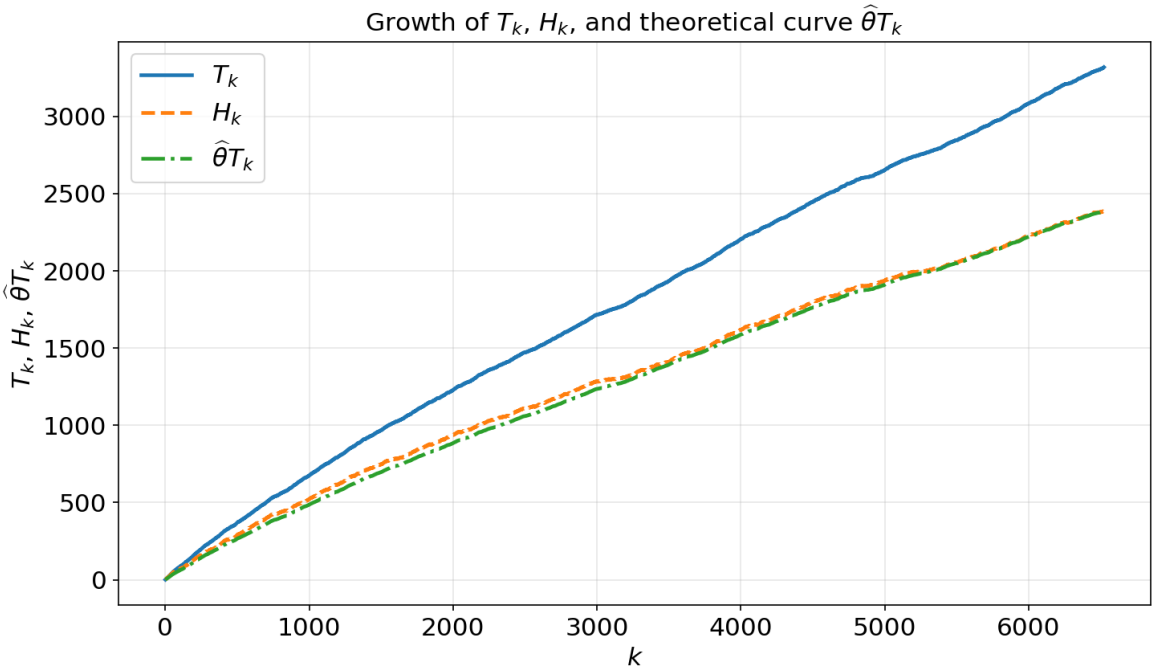

Figure 2: Growth of T_k, H_k, and $\widehat{\theta}T_k$ in Said Ahmad's *Sobiq o'g'ri* (original text). Here H_k closely follows the theoretical curve, indicating a good fit of Karlin's model.

variants, while 116 poems have a p-value ≥ 0.1 in all variants. This stability suggests that the observed differences in model fit are primarily due to the internal structure of the texts, rather than the details of morphological normalization.

Individual examples illustrate these trends. In "Xalq so'zi" issue 179, the p-value is approximately 0.60 for the original text, 0.95 for the lemmatized version, and 0.62 for the stemmed version. Analysis of the trajectories of T and H shows that despite local irregularities caused by shifts between different topics and article types, their growth remains roughly proportional, so Karlin's model provides a satisfactory approximation for this issue. In the poetry subcorpus, Jamal Sirojiddin's "Tanbur" collection represents a typical case of very good fit: the p-value is 0.75 for the original text and increases to 0.96 and 0.99 for the lemmatized and stemmed versions, respectively. Here, the number of hapax legomena closely follows the theoretical curve $\widehat{\theta}T_k$ throughout the text, and morphological normalization makes this proportionality even more regular.

Overall, newspapers and poetry support the interpretation that Karlin's elementary probabilistic model works best for relatively short and structurally homogeneous texts. In such cases, the results remain stable for both lemmatization and stemming, while long text with uneven internal structure continues to be the main source of systematic deviations from the model.

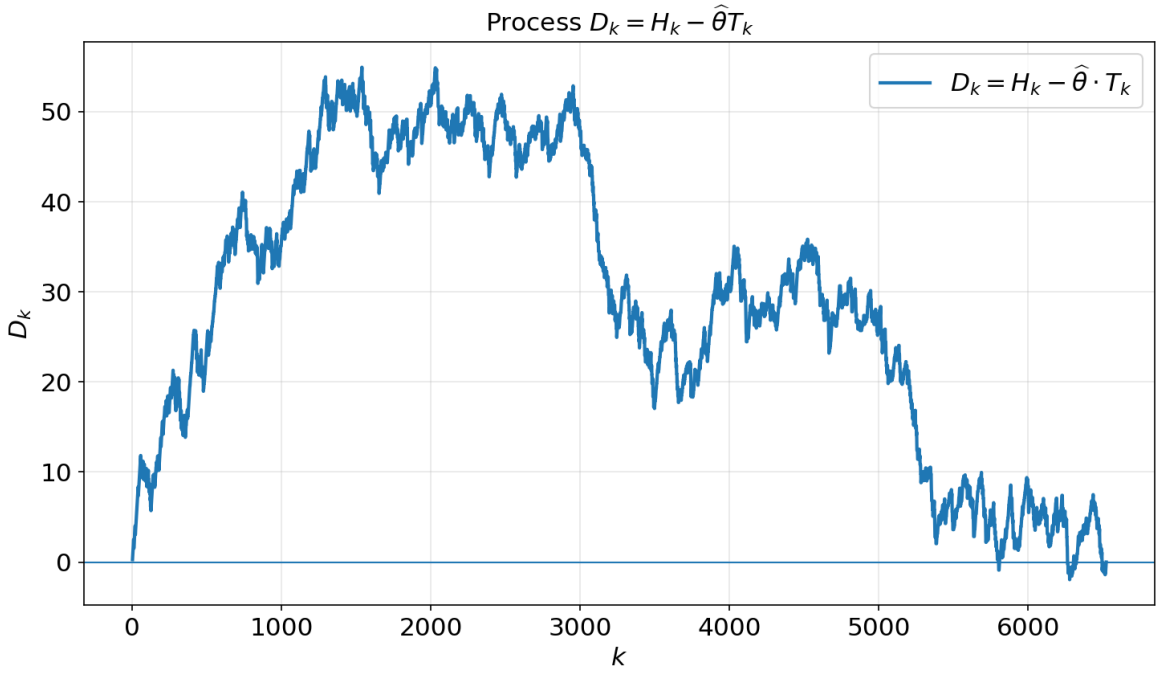

Figure 3: Behaviour of the process $D_k = H_k - \widehat{\theta} T_k$ for Said Ahmad's *Sobiq o'g'ri* (original text). The process fluctuates around zero without a clear trend, which is consistent with the model.

4.4 Dependence on Text Length

The relationship between the p-value and text length n is shown in Figure 4, where each point corresponds to one text from the corpus (based on the original, unnormalized version). The horizontal axis shows the number of tokens n, and the vertical axis the corresponding p-value. The dashed horizontal lines indicate levels p = 0.1, p = 0.05, and p = 0.01.

The figure shows a clear dependence of model fit on text length. For relatively short texts (up to approximately 10^4 words), the points are widely distributed across the interval (0,1), and many texts have p-values significantly higher than 0.1, sometimes close to 1. Poetry and newspapers are well represented in this domain. As n increases, the point cloud gradually shifts downward: for texts of medium length, the proportion of p-values ≥ 0.1 decreases, while for very long texts (with n on the order of 10^4–10^5 words), most p-values lie below 0.1, with many p-values being very small. Almost all of these very low p-values correspond to long prose works.

This pattern is consistent with the general behavior of statistical tests. In particular, for huge samples, classical goodness-of-fit tests may reject the null hypothesis even under very small departures from the model, because test power increases with n. This "large-sample" sensitivity has been discussed in quantitative linguistics; see, for example, Mačutek and Wimmer (2013). Therefore, part of the downward shift of p-values with increasing text length may reflect this general statistical effect. As sample size n increases, a test based on the Q_n statistic becomes more sensitive to systematic differences between

the empirical trajectories T_k and H_k and the theoretical predictions of the Karlin model. For short texts, random fluctuations can mask such differences, and the model is rarely rejected. However, for very long texts, localized spikes in rare words (e.g., due to topic changes or extensive descriptive fragments) accumulate across the document, pushing p-values closer to zero.

To ensure that the dependence on n is not a preprocessing artifact, we repeated the same analysis for the lemmatized and stemmed versions of the corpus. In both cases, the distribution of points on the (n, p-value) plane exhibits the same qualitative trend: high p-values are more common among shorter texts, while longer prose works tend to yield low p-values. Morphological normalization slightly shifts some individual texts upward (especially in the newspaper and poetry subcorpora), but does not affect the overall decrease in p-values with increasing n. This suggests that the observed dependence on text length reflects genuine properties of the model and the data, rather than being a byproduct of the preprocessing procedure.

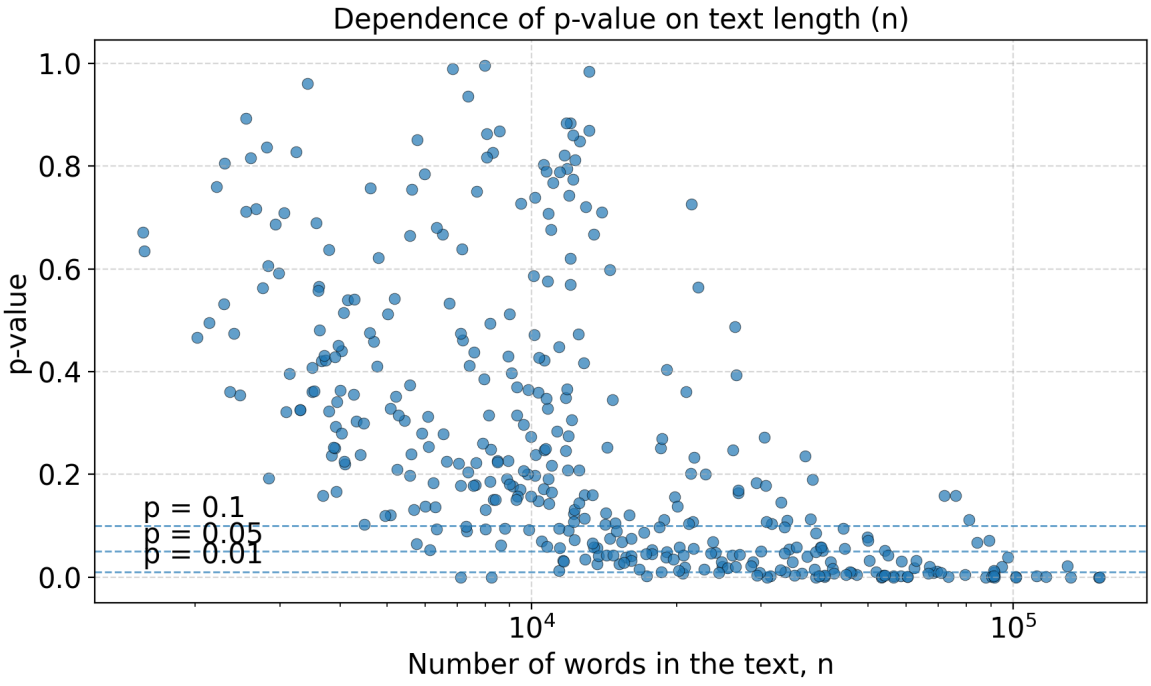

Figure 4: Plot of *p-value* versus text length n (original texts). Each point corresponds to one text from the corpus. Dashed lines indicate levels *p-value* = 0.1, 0.05, and 0.01.

4.5 English texts and comparison with existing tests

To verify that the new Q_n statistic behaves consistently with previously proposed hapax-type tests, we applied all three criteria to ten English texts analyzed in Fayzullaev and Kovalevskii (2024). For six long novels (Alice's Adventures in Wonderland, Dracula, Frankenstein, or the Modern Prometheus, Pride and

Prejudice, The Great Gatsby, and The Scarlet Letter), all three tests yield very low p-values ($p < 0.01$) and therefore confidently reject Karlin's model. In contrast, three shorter, more homogeneous texts (Metamorphosis, Romeo and Juliet, and The Picture of Dorian Gray) have p-values above 0.1 for all three tests, indicating acceptable model fit.

The only borderline case is the popular science text Simple Field Guide to Sabotage. Here, our Q_n and $Q_n^{(2)}$ statistics yield p-values around 0.10, while $Q_n^{(1)}$ yields a smaller p-value of approximately 0.02 and rejects the model at the 10% and 5% levels. Thus, $Q_n^{(1)}$ behaves as a more conservative test, while Q_n is numerically very close to $Q_n^{(2)}$ and leads to the same accept/reject decisions at all standard significance levels.

5 Conclusion

In this paper, we proposed a new fit statistic for the elementary probabilistic Karlin model, based on the joint behavior of numbers of type T_n and hapax legomena H_n at positions n and $[n/2]$, and applied it to a large corpus of Uzbek texts. The corpus includes prose, poetry, and newspapers. For each text, we analyzed three versions: the original token sequence, a lemmatized version, and a stemmed version. This allowed us to study both the theoretical properties of the test and its empirical behavior under different preprocessing conditions.

The main empirical result is that the Karlin model is significantly more suitable for short and structurally homogeneous texts than for long narrative prose. For newspapers and poetry, most texts have p-value ≥ 0.1, with a significant proportion reaching p-values above 0.5. In contrast, only about a third of prose works pass the test at the 0.1 level, and almost all the extremely low p-values in the corpus are generated by long novels and short story collections. The dependence on text length is particularly pronounced: for texts up to 10^4 words long, the points on the $(n, \text{p-value})$ plane are widely scattered, while for very long texts, most p-values cluster near zero. This suggests that Karlin's model captures general patterns of vocabulary growth but becomes sensitive to the long, uneven structure of prose texts.

Analysis of individual texts helps explain this behavior. In several novels with very small p-values, the number of hapax legomena increases much faster than the model predicts, especially in the first half of the text, and the process $D_k = H_k - \widehat{\theta} T_k$ moves away from zero. In other works, such as Said Ahmad's "Sobiq o'g'ri," the growth of H_k quickly becomes proportional to the growth of T_k, and D_k fluctuates around zero without a clear trend, leading to high p-values. These examples demonstrate that deviations from the Karlin model are closely related to localized spikes in rare vocabulary caused by topic shifts, long descriptive passages, or abrupt register changes.

A separate question was whether the observed effects are artifacts of morphological preprocessing. Our additional experiments show that this is not the case. For most texts, the classification into "good" and "bad" matches is robust across the original, lemmatized, and stemmed versions. Lemmatization and stemming slightly improve the match for some borderline cases (e.g., Hamid Gulom's "Qoradaryo"), but do not change the overall picture: newspapers and poems generally conform well to the model, while long prose remains the main source of systematic deviations.

Thus, the proposed statistics provide a convenient and informative tool for testing Karlin's model on real texts and identifying its failures. At the same time, our results show that a single global parameter is often insufficient to describe vocabulary growth in long human-generated texts. Future research will focus on extending the model to locally non-uniform or piecewise regimes (e.g., with parameters changing over the course of the text) and on constructing tests comparing $T_{[tn]}$ and $H_{[tn]}$ at multiple time points $t \in (0, 1)$. Such refinements should allow for a more accurate description of early irregularities while preserving the useful asymptotic structure of Karlin's approach.

In addition to these empirical observations, it is useful to summarize the main methodological properties of the proposed statistics.

In this paper, we introduced a new goodness-of-fit statistic

$$V_n = \frac{T_n H_{\lfloor n/2 \rfloor} - H_n T_{\lfloor n/2 \rfloor}}{T_n^{3/2}}, \qquad Q_n = \frac{V_n}{\Sigma_n(\hat{\theta})},$$

derived from the functional central limit theorem for Karlin's infinite urn scheme. The test exploits the joint behavior of the number of types T_n and the hapax legomena H_n at positions n and $n/2$, not just their finite values. This design makes the statistic particularly sensitive to local variations in vocabulary growth (e.g., topic shifts and spikes in rare words), while remaining robust to small random fluctuations in individual word frequencies.

From a practical standpoint, the test is simple to calculate: it requires only the aggregate values of T_n, H_n, $T_{\lfloor n/2 \rfloor}$, and $H_{\lfloor n/2 \rfloor}$, and it returns a single scalar value Q_n, which can be directly interpreted using the standard normal distribution. This allows for the comparison of a large number of texts and genres on a single scale and the clear identification of systematic deviations from Karlin's model.

For the ten English test texts from Fayzullaev and Kovalevskii (2024), our Q_n statistic produces p-values that are nearly indistinguishable from those of $Q_n^{(2)}$ and gives exactly the same accept/reject decisions at standard significance levels, whereas $Q_n^{(1)}$ systematically produces smaller p-values and hence acts as a more conservative test.

At the same time, we do not claim that the Q_n statistic is universally more powerful than classical goodness-of-fit procedures such as the Kolmogorov-Smirnov test, χ^2, or the likelihood ratio test. Our construction still relies on the simplifying assumptions of Karlin's model, in particular, the approximately independent sample of words. As discussed in the literature on Zipf-type laws (see, e.g., Altmann and Gerlach (2016)), this assumption is known to be violated in real texts due to long-range correlations and topic structure.

Furthermore, Q_n is based only on the number of types and hapax legomena. Consequently, two different texts may have very similar Q_n values even if their full word frequency distributions differ significantly. In this sense, the proposed test is best viewed as a convenient model-based diagnostic measure that complements, rather than replaces, the more general distribution-based tests used in quantitative linguistics.

Acknowledgments

I am grateful to my scientific supervisor, Doctor of Physical and Mathematical Sciences Artem Pavlovich Kovalevskii, for valuable guidance and important recommendations during the preparation of this work. I also acknowledge the support of the "El-Yurt Umidi" Foundation for the Training of Prospective Personnel under the President of the Republic of Uzbekistan, which provided a scholarship during my graduate studies.

I also thank the anonymous referee for careful reading of the manuscript and for constructive comments that helped improve the presentation.

Data availability statement

All datasets used in this study (original, lemmatized, stemmed statistics for 386 texts and the 500-word evaluation list) are publicly available on Zenodo: https://doi.org/10.5281/zenodo.17711166.

References

Abebe, B., Chebunin, M., Kovalevskii, A. (2024). Text segmentation via processes that count the number of different words forward and backward. *Journal of Quantitative Linguistics*, *31*(1), 1–18. https://doi.org/10.1080/09296174.2023.2275342

Abebe, B. (2025). A new method for detecting multiple text change points. *Glottotheory*, *16*(1), 1–15. https://doi.org/10.1515/glot-2025-2003

Altmann, E. G., Gerlach, M. (2016). Statistical laws in linguistics. In M. Degli Esposti, E. G. Altmann, F. Pachet (Eds.), *Creativity and universality in language* (pp. 7–26). Springer International Publishing. https://doi.org/10.1007/978-3-319-24403-7_2

Bao, M., Yan, J., Huang, D. (2025). Zipf's law for discourse markers in spoken Mongolian. *Journal of Quantitative Linguistics*, *32*(2), 166–180. https://doi.org/10.1080/09296174.2025.2463754

Chebunin, M., Kovalevskii, A. (2016). Functional central limit theorems for certain statistics in an infinite urn scheme. *Statistics and Probability Letters*, *119*, 344–348. https://doi.org/10.1016/j.spl.2016.08.019

Chebunin, M., Kovalevskii, A. (2019). Asymptotically normal estimators for Zipf's law. *Sankhya A*, *81*, 482–492. https://doi.org/10.1007/s13171-018-0135-9

Davis, V. (2018). Types, tokens, and hapaxes: A new Heap's law. *Glottotheory*, *9*(2), 113–129. https://doi.org/10.1515/glot-2018-0014

Fayzullaev, S., Kovalevskii, A. (2024). Hapax legomena via stochastic processes. *Glottometrics*, *56*, 22–39. https://doi.org/10.53482/2024_56_415

Galieva, A., Vavilova, Z. (2021). Initial and final syllables in tatar: From phonotactics to morphology. *Glottometrics*, *50*, 57–75.

Harrison, P. N. (1921). *The problem of the pastoral epistles*. Oxford University Press.

Karlin, S. (1967). Central limit theorems for certain infinite urn schemes. *Journal of Mathematics and Mechanics*, *17*(4), 373–401.

Kudryavtseva, A., Kovalevskii, A. (2025). Comparative statistical analysis of word frequencies in human-written and ai-generated texts. *Glottometrics*, *58*, 19–34. https://doi.org/10.53482/2025_58_423

Mačutek, J., Nogolová, M., Rovenchak, A., Čech, R. (2026). What does the Menzerath-Altmann law really say? *Journal of Quantitative Linguistics*, *33*(1), 28–43. https://doi.org/10.1080/09296174.2025.2545052

Mačutek, J., Wimmer, G. (2013). Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics*, *20*(3), 227–240. https://doi.org/10.1080/09296174.2013.799912

Milička, J. (2009). Type-token & hapax-token relation: A combinatorial model. *Glottotheory*, *2*(1), 99–110. https://doi.org/10.1515/glot-2009-0009

Petrini, S., Casas-i-Muñoz, A., Cluet-i-Martinell, J., Wang, M., Bentz, C., Ferrer-i-Cancho, R. (2023). Direct and indirect evidence of compression of word lengths. Zipf's law of abbreviation revisited. *Glottometrics*, *54*, 58–87. https://doi.org/10.53482/2023_54_407

Popescu, I.-I., Altmann, G. (2008). Hapax legomena and language typology. *Journal of Quantitative Linguistics*, *15*(4), 370–378. https://doi.org/10.1080/09296170802326699

Tuzzi, A., Popescu, I.-I., Altmann, G. (2009). Zipf's laws in Italian texts. *Journal of Quantitative Linguistics*, *16*(4), 354–367. https://doi.org/10.1080/09296170903211519

Appendix

Proof of the theorem. First, consider the following expression:

$$B_n = \frac{T_n H_{[n/2]} - H_n T_{[n/2]}}{T_n^2} = \frac{H_{[n/2]}}{T_n} - \frac{H_n}{T_n}\frac{T_{[n/2]}}{T_n} = \frac{\frac{H_{[n/2]}-\eta_{[n/2]}}{\tau_n} + \frac{\eta_{[n/2]}}{\tau_n}}{\frac{T_n-\tau_n}{\tau_n}+1} -$$

$$-\frac{\frac{H_n-\eta_n}{\tau_n} + \frac{\eta_n}{\tau_n}}{\frac{T_n-\tau_n}{\tau_n}+1}\frac{\frac{T_{[n/2]}-\tau_{[n/2]}}{\tau_n} + \frac{\tau_{[n/2]}}{\tau_n}}{\frac{T_n-\tau_n}{\tau_n}+1} =$$

$$= \frac{\frac{Z_n(1/2)}{\sqrt{\tau_n}} + \frac{\eta_{[n/2]}}{\tau_n}}{\frac{Y_n(1)}{\sqrt{\tau_n}}+1} - \frac{\frac{Z_n(1)}{\sqrt{\tau_n}} + \frac{\eta_n}{\tau_n}}{\frac{Y_n(1)}{\sqrt{\tau_n}}+1}\frac{\frac{Y_n(1/2)}{\sqrt{\tau_n}} + \frac{\tau_{[n/2]}}{\tau_n}}{\frac{Y_n(1)}{\sqrt{\tau_n}}+1}$$

From (3) and Lemmas 1 and 2 from Chebunin and Kovalevskii (2019), we obtain the following expression for the mathematical expectations T_n and H_n.

$$\tau_n = \Gamma(1-\theta)c^\theta n^\theta + o(n^{\theta/2}),$$

$$\eta_n = \theta\Gamma(1-\theta)c^\theta n^\theta + o(n^{\theta/2}).$$

This gives:

$$\eta_n/\tau_n = \theta + o(n^{-\theta/2}),$$

$$\eta_{[nt]}/\eta_n = t^\theta + o(n^{-\theta/2}),$$

$$\tau_{[nt]}/\tau_n = t^\theta + o(n^{-\theta/2})$$

as $n \to \infty$, $t > 0$, and taking this into account, let us return to the original expression:

$$B_n = \frac{\frac{Z_n(1/2)}{\sqrt{\tau_n}} + \theta 2^{-\theta}}{\frac{Y_n(1)}{\sqrt{\tau_n}}+1} - \frac{\frac{Z_n(1)}{\sqrt{\tau_n}} + \theta}{\frac{Y_n(1)}{\sqrt{\tau_n}}+1}\frac{\frac{Y_n(1/2)}{\sqrt{\tau_n}} + 2^{-\theta}}{\frac{Y_n(1)}{\sqrt{\tau_n}}+1} + o(n^{-\theta/2})$$

If a random variable satisfies the condition $\xi_n \xrightarrow{p} 0$, then $\frac{1}{1+\xi_n} = 1 - \xi_n + o_p(\xi_n)$, and from this:

$$B_n = \left(\frac{Z_n(1/2)}{\sqrt{\tau_n}} + \theta 2^{-\theta}\right)\left(1 - \frac{Y_n(1)}{\sqrt{\tau_n}}\right) -$$

$$-\left(\frac{Z_n(1)}{\sqrt{\tau_n}} + \theta\right)\left(1 - \frac{Y_n(1)}{\sqrt{\tau_n}}\right)\left(\frac{Y_n(1/2)}{\sqrt{\tau_n}} + 2^{-\theta}\right)\left(1 - \frac{Y_n(1)}{\sqrt{\tau_n}}\right) + o_p(n^{-\theta/2}) =$$

$$= \frac{1}{\sqrt{\tau_n}}\left(Z_n(1/2) - 2^{-\theta}Z_n(1) - \theta Y_n(1/2) + \theta 2^{-\theta}Y_n(1)\right) + o_p(n^{-\theta/2})$$

Considering that $V_n = \sqrt{T_n} B_n$, we have

$$V_n = \frac{T_n}{\sqrt{\tau_n}} \left(Z_n(1/2) - 2^{-\theta} Z_n(1) - \theta Y_n(1/2) + \theta 2^{-\theta} Y_n(1) \right) + o_p(n^{-\theta/2})$$

Considering the SLLN for T_n and the FCLT of $Z_n(t)$ and $Y_n(t)$ for the cases $t = 1$ and $t = 1/2$, V_n converges weakly to $V = Y_1(1/2) - 2^{-\theta} Y_1(1) - \theta Y(1/2) + \theta 2^{-\theta} Y(1)$ with a mean value of 0. Its variance is calculated based on the covariance matrix of the vector $(Y(1), Y(1/2), Y_1(1), Y_1(1/2))$. The proof is complete.