

KidLit by Numbers: Linguistic Profiles of Original and Translated Russian 0+ Books

Daria Gavrilik Aleksandra Ostraya

Peter the Great St. Petersburg Polytechnic University (SPbPU)
St. Petersburg, Russia
daria.gavrilic@gmail.com aleksandra.ostraya@mail.ru

Abstract

We present **KidLit v0.1**, the first balanced corpus of contemporary Russian 0+ children’s literature: 100 texts (2015–2026), 50 original works (**RUS-O**) and 50 Russian translations (**RUS-T**). We analyse 35 linguistic metrics (volume, lexical diversity, morphosyntax, readability) using the Mann–Whitney U test with Benjamini–Hochberg correction (FDR, $q = 0.05$). Despite a large volume difference (RUS-O are 3 times longer than RUS-T, $d = 0.91$), the two groups are indistinguishable on 4 readability indices, basic POS distribution, syntactic depth, and active voice. Apart from volume, the only robust effects are the inverse relationship between normalised (TTR, hapax – higher in RUS-T) and absolute (Shannon entropy – higher in RUS-O) lexical density, and a higher proportion of coordinated predicates in RUS-O. We interpret this pattern as evidence that the grammatical and syntactic profile of 0+ children’s literature is set by the genre-age niche rather than by the origin of the text.

1 Introduction

Reading plays a foundational role in the cognitive, linguistic, and social development of children. In line with Vygotsky’s Zone of Proximal Development framework (Vygotsky, 1978), a text addressed to the youngest readers must balance accessibility with novelty without inducing excessive cognitive load. The age label “0+” formally marks books recommended for children from birth to 3+ years old. In practice, such books are designed to be read aloud by an adult: preschool children are not yet able to read independently but they actively listen and engage with the text and illustrations. This age category brings together texts that have been minimally studied quantitatively: existing Russian-language corpora of children’s literature focus on school-age material and do not treat the 0+ segment as an independent object.

At the same time, the 0+ segment has a complex and poorly documented structure. A substantial share of pre-school children’s books on the Russian market consists of translations from foreign languages. These translated texts undergo extensive editorial and translational adaptation (O’Sullivan, 2005; Lathey, 2010), which formally renders them Russian-language works but not original in the authorial sense. A linguistic comparison of original and translated texts within the same age niche allows us to answer a key question: do translations preserve the stylistic and grammatical features of the source genre, or do they adapt to the norms of Russian-language children’s literature, converging on the characteristics of original texts?

This work addresses the following research question:

How do original Russian-language texts (RUS-O) and Russian-language translations of foreign literature (RUS-T) for the 0+ age group differ in narrative volume, lexical diversity, morphological and syntactic profile, and standard readability indices?

Our contribution is fourfold. (i) We introduce **KidLit v0.1**, the first balanced corpus of contemporary (2015–2026) Russian-language 0+ children’s literature, annotated by origin type. (ii) We provide a comparative quantitative profile over 35 linguistic metrics with Benjamini–Hochberg correction for multiple comparisons (Benjamini and Hochberg, 1995). (iii) We uncover a non-trivial pattern: apart from differences in volume and normalised lexical density, RUS-O and RUS-T are linguistically aligned, which we interpret as a trace of the genre-age editorial standard for 0+ literature. (iv) We report the first manual validation of a standard NLP pipeline for Russian on 0+ children’s material, bounding the error of sentence segmentation,

tokenisation, POS tagging, and coordination detection on this domain.

2 Related Work

2.1 Automatic complexity assessment and readability indices

Morozov et al. (2022) and Iomdin and Morozov (2021) carry out large-scale empirical studies of Russian-language children’s school literature, identifying textual features that improve automatic complexity classification. They combine readability indices (Russian adaptations of Flesch, FOG, ARI) with morphological, syntactic, punctuational, and frequency features, complemented by topic modelling. On this basis, a neural network for automatic complexity classification was trained and released as open source (Morozov, 2022). However, this work targets school-age material; for the 0+ segment, the calibration and validation of Russian readability indices remains an open task.

In a study of book age ratings and readability, Glazkova (2021) shows, on a collection of 5,516 sample fragments, that the mean values of readability indices (Flesch–Kincaid, Coleman–Liau, ARI, SMOG, Dale–Chall) increase monotonically from the 0+ category to 16+, yet unexpectedly drop for the 18+ category. Importantly, Glazkova works on sample fragments (10% of the full text) rather than full texts, which limits the precision of the metrics, especially for short 0+ texts. In our study we apply the same indices to the full texts of KidLit v0.1 and find no significant differences between RUS-O and RUS-T within the single 0+ age category (Section 5.5), indicating limited sensitivity of the indices to variation within a narrow age niche.

2.2 Corpora of children’s and young-adult literature

Maslinsky et al. (2021) developed **DetCorpus**, a representative annotated corpus of 20th–21st century Russian children’s literature comprising more than 2,000 works. This resource covers predominantly school-age literature; no dedicated 0+ sub-corpus has been carved out.

The **CYP-LEX** corpus (Korochkina et al., 2024) provides a lexicon of British children’s literature for ages 7+ across 1,200 books and 70 million word occurrences. The **Russian National Corpus** sub-corpus “From 2 to 15” (Savchuk et al., 2024) contains 75 texts selected by child-reader preference. None of the existing Russian-language resources

offers a balance by origin status (original vs. translation), which makes a targeted comparative analysis infeasible on currently available data.

2.3 Comparing original and translated texts

Linguistic studies of translation draw on the paradigm of translation universals (Baker, 1993), which includes hypotheses of explicitation, simplification, and normalisation in translated texts relative to originals. Empirical support for these hypotheses on adult fiction has been reported by Laviosa (2002) and Olohan (2004). In Russian translation studies, Kunilovskaya et al. (2017) identify four principal universals—simplification, explicitation, normalisation, and interference—and note that the most frequent indicators are measures of lexical diversity (TTR), sentence length, and the frequency of function words. However, whether these universals persist in the 0+ age niche, where editorial constraints are already strong, has not previously been addressed.

3 The KidLit v0.1 Corpus

3.1 Inclusion criteria

Texts were included if they satisfied four conditions: (a) they bear the publisher’s “0+” age marker or its equivalent (“for pre-school age”, “for adults to read to children”); (b) the year of first publication (or, for RUS-T, of actual creation of the text) is no earlier than 2015; (c) the work is a piece of fiction featuring characters, plot, and dialogue (excluding purely educational and instructional material); (d) full text is accessible with extractable metadata.

3.2 Composition and balance

The present version (**KidLit v0.1**) contains 100 texts balanced by origin type: 50 in subgroup **RUS-O** (original works by Russian-language authors) and 50 in subgroup **RUS-T** (Russian translations of foreign literature produced by professional translators for the Russian book market). Translations come from 7 source languages, with English dominating (48%). Texts were sourced from electronic libraries and digitised printed editions of leading Russian children’s publishers. Basic characteristics are summarised in Table 1.

3.3 Preprocessing

The majority of texts were obtained by optical character recognition (OCR) of scanned pages using

Characteristic	RUS-O	RUS-T
Number of texts	50	50
Time span	2015–2026	
Mean words / text	2,067	698
Median words / text	1,346	547
Min / max (words)	188 / 11,461	81 / 3,819
Mean sentences	278	91
Mean pages	41.7	34.6
Source languages	—	7
Publishers in sample	26	21

Table 1: Basic characteristics of the KidLit v0.1 corpus.

the Mistral AI API (Mistral AI, 2024) with subsequent manual verification. Digitally available texts underwent graphematic cleaning: paratextual elements (prefaces, page numbers, colophons) were removed; quotation marks, dashes, and ellipses were normalised. Metadata (title, author, year, publisher, translator, age marker, source language, author gender) are stored in tabular form; each text was assigned a unique identifier.

3.4 A living corpus

Children’s literature is a dynamic object of study because new works are published regularly, which requires a corpus resource capable of continuous updating. KidLit is designed as a living corpus supporting expansion and versioning. This paper presents the first official version, KidLit v0.1 (100 texts, 50 RUS-O + 50 RUS-T). The corpus will grow along 2 lines: (1) the inclusion of new works appearing on the Russian book market, and (2) the retrospective digitisation and annotation of existing texts that meet the inclusion criteria. Subsequent versions (KidLit vN.0, 120+ texts) will be presented in separate publications, ensuring reproducibility of results and continuity of the research programme.

4 Methodology

4.1 NLP pipeline

All texts were processed through a single Python 3.12 pipeline. Tokenisation and sentence segmentation rely on **razdel** (Natasha Project, 2023); morphological analysis was performed with **pymorphy3** (Korobov, 2015); the lexical and syntactic profiles were produced by the spaCy model for Russian **ru_core_news_sm** (Honnibal et al., 2020).

4.2 Metric groups

For each text we computed 35 quantitative metrics in the *main block*, organised into eight groups,

together with 17 metrics in an *extended morphological block*.

- (1) **Volume** (3): word tokens, sentences, pages.
- (2) **Lexical diversity** (6): TTR, STTR (100-token window), hapax ratio, Yule K, Shannon entropy, top-1000 word ratio (Lyashevskaya and Sharov, 2009).
- (3) **POS profile – main** (8): proportions of nouns, verbs, adjectives, adverbs, proper nouns, pronouns, personal pronouns, animate nouns.
- (4) **Syntax** (2): mean dependency tree depth, predicate-coordination ratio (proportion of sentences containing a conj relation in which both dependent and head are verbal).
- (5) **Readability** (4): Flesch (RU), FOG, ARI, Coleman–Liau.
- (6) **Word & syllable** (4): mean word length in characters, mean syllables per word, mean sentence length in words, mean sentence length in characters.
- (7) **Grammar** (6): past, present, future tense and imperative mood proportions; interrogative and exclamatory sentence proportions.
- (8) **Additional** (2): active-voice ratio, punctuation diversity.
- (9) **Extended morphological block** (17, *t*-test): indices of analyticity, verbality, substantivity, pronominality, adjectivity, nominal vocabulary, and the noun-to-verb ratio; proportions of word forms in the genitive and instrumental cases, short adjectives, full and short participles, gerunds, infinitives, numerals, predicatives, and particles.

4.3 Statistical analysis

For each metric, between-group differences were assessed with the **Mann–Whitney U test**. This non-parametric method was chosen because most metrics exhibit right-skewed distributions, which violate the assumptions of parametric tests. Effect size is reported as **Cohen’s *d***: $|d| < 0.2$ negligible, 0.2–0.5 small, 0.5–0.8 medium, > 0.8 large.

To control Type I error under simultaneous testing we applied the **Benjamini–Hochberg** correction (Benjamini and Hochberg, 1995) (FDR $q = 0.05$) *independently* to the main block (35

metrics) and the extended morphological block (17 metrics). As a stricter check, **Holm’s correction** (Holm, 1979) (FWER) was also applied; both procedures yield a consistent set of robust effects. All adjusted p -values reported below are BH-corrected (p_{adj}); raw p -values (p_{raw}) appear in the appendix tables.

5 Results

Of 35 main-block metrics, nine are significant before correction ($p_{\text{raw}} < 0.05$); after BH correction, **seven** remain robustly significant (Table 2). Structurally, these seven effects reduce to four substantive contrasts discussed below. Full results are presented in Appendix A.

Metric	RUS-O	RUS-T	d	p_{adj}
Word count	2,067	698	+0.91	< .001
Sentence count	278	91	+0.77	< .001
Shannon entropy	8.20	7.20	+1.10	< .001
TTR	0.57	0.66	-0.85	.001
Hapax ratio	0.42	0.53	-0.82	.001
Predicate coord.	0.25	0.18	+0.79	.002
Pages	41.7	34.6	+0.42	.005

Table 2: Seven robust effects after BH FDR correction ($q = 0.05$). Sign of d : “+” higher in RUS-O, “-” higher in RUS-T.

5.1 Volume of the narrative

Original Russian texts are significantly longer than translations across all three volume metrics: 2,067 words vs. 698 words ($d = 0.91$, $p_{\text{adj}} < .001$); 278 sentences vs. 91 ($d = 0.77$, $p_{\text{adj}} < .001$); 41.7 pages vs. 34.6 ($d = 0.42$, $p_{\text{adj}} = .005$). The ratio of mean volumes is approximately 2.96:1 in words and 3.05:1 in sentences, but only 1.21:1 in pages, indicating fundamentally different text-per-page densities. Distributions are shown in Figure 1.

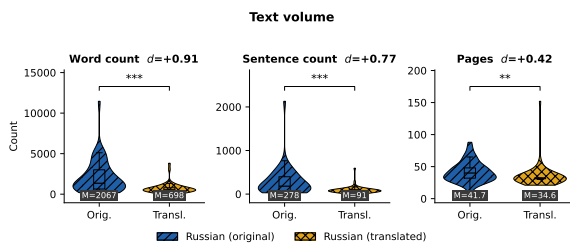


Figure 1: Volume comparison between RUS-O and RUS-T. Asterisks show BH-corrected significance: *** $p_{\text{adj}} < .001$, ** $p_{\text{adj}} < .01$.

5.2 Lexical diversity: the normalisation paradox

Lexical diversity metrics exhibit a classic normalisation paradox (Figure 2). On a per-unit basis, translations are lexically denser: TTR = 0.66 vs. 0.57 ($d = -0.85$, $p_{\text{adj}} = .001$) and hapax ratio = 0.53 vs. 0.42 ($d = -0.82$, $p_{\text{adj}} = .001$).

However, the absolute measure of lexical richness points in the opposite direction: Shannon entropy is 8.20 bits in RUS-O vs. 7.20 bits in RUS-T ($d = 1.10$, $p_{\text{adj}} < .001$). RUS-O contain more unique vocabulary in absolute terms, but this vocabulary is “diluted” across a narrative three times larger, which lowers normalized density measures.

STTR (standardised TTR, 100-token window) shows the same trend ($d = -0.53$, $p_{\text{raw}} = .016$) but does not survive correction ($p_{\text{adj}} = .071$). This is consistent with the design of STTR: the index compensates for length differences, and under large volume disparities it gives the most conservative estimate of lexical density. We treat the STTR effect as a trend requiring verification on an expanded sample.

Yule K index and top-1000 word ratio (Lyshevskaya and Sharov, 2009) are identical between groups ($p_{\text{adj}} > 0.81$): the concentration of frequent words and the shape of the frequency-distribution tail coincide. This means that both groups rely on a structurally comparable frequency profile, consistent with the hypothesis of alignment by a genre-age standard.

5.3 Coordinative syntactic organisation

Of the two syntactic metrics, the only robust difference concerns coordination. The proportion of sentences containing coordinated predicates is 0.251 in RUS-O vs. 0.177 in RUS-T ($d = 0.79$, $p_{\text{adj}} = .002$). Original texts string homogeneous predicates together appreciably more often than translations do.

We restrict the claim to predicates deliberately. Manual annotation (Section 5.6) shows that coordination of full clauses, each with its own subject, occurs in only about 3% of sentences and at the same rate in both subcorpora. The contrast is therefore not one of clause-level complexity: texts in both groups are built around a single predication, and they differ in how many predicates that predication carries. Read together with the volume result of Section 5.1, this describes two ways of advancing a narrative: RUS-O extend the sentence by attach-

Lexical diversity

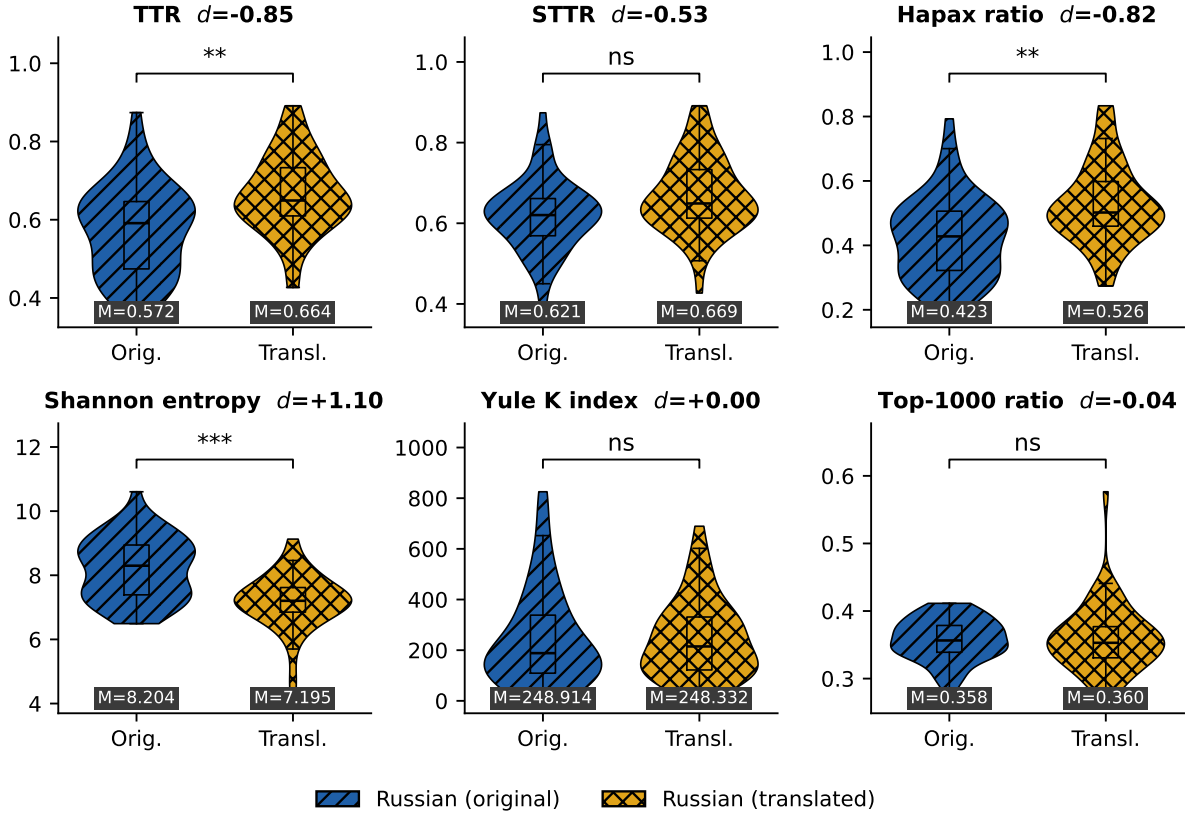


Figure 2: Lexical diversity metrics for RUS-O and RUS-T. Asterisks show BH-corrected significance: *** $p_{adj} < .001$, ** $p_{adj} < .01$; ns = not significant after BH FDR correction.

ing further actions to the same actor, while RUS-T close the sentence and open a new one. Subordination was not measured directly, and mean dependency tree depth shows no significant difference after correction ($p_{adj} = .347$): syntactic “depth” in the traditional sense is not a marker of origin in our corpus.

5.4 Morphological alignment

An extended analysis of the morphological profile across 17 metrics shows that no morphological metric survives the multiple-comparisons correction (all $p_{adj} > 0.07$), as shown in Figure 3. The closest to significance is the analyticity index (function-word ratio: 0.176 in RUS-O vs. 0.163 in RUS-T, $d = 0.58$, $p_{raw} = .005$, $p_{adj} = .080$).

The basic POS distribution is virtually identical: nouns (0.235 vs. 0.236), verbs (0.192 vs. 0.189), adjectives (0.076 vs. 0.067; all $p_{adj} \geq .17$). Fine-grained indicators (proportions of infinitives, short adjectives, participles, gerunds, predatives,

oblique case forms) also show no significant differences ($p_{adj} > 0.35$). Noun-to-verb ratio: 1.58 vs. 1.63 ($d = -0.13$, $p_{adj} = .80$). The distribution of verb tenses shows an observational trend: RUS-O lean toward past tense (0.58 vs. 0.48), RUS-T toward present (0.21 vs. 0.29); neither survives correction ($p_{adj} = .184$ and $.218$, respectively). Active and passive voice proportions are identical (0.76 vs. 0.77, $p_{adj} = 1.00$). Full morphological results are in Appendix B.

5.5 Readability: a robust null result

Four readability indices (Flesch RU, FOG, ARI, Coleman–Liau) show no significant differences between groups (all $p_{adj} > 0.39$; $|d| \leq 0.37$; see Figure 4 in Appendix C). This aligns with the known critique of English-language readability indices and their Russian adaptations (Morozov et al., 2022; Iomdin and Morozov, 2021; Glazkova, 2021): these metrics are calibrated on school-age and adult texts and may be insensitive to the

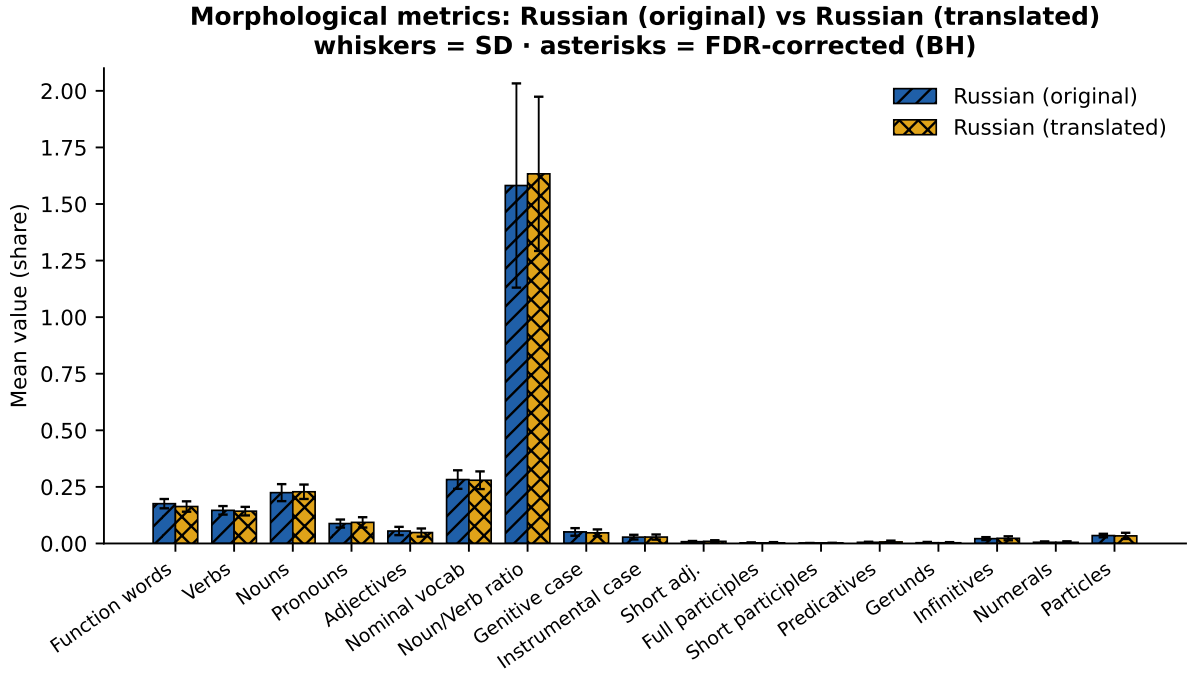


Figure 3: Extended morphological profile: RUS-O vs. RUS-T (17 metrics, t -test). Whiskers show SD. No metric is significant after BH FDR correction (all $p_{\text{adj}} > 0.07$).

specifics of pre-school literature. A possible cause may be the work of translators and editors in aligning texts on “reading difficulty”, which is discussed in Section 6.

5.6 Validation of the NLP pipeline on 0+ material

The tools we rely on are trained on general Russian corpora, so we measured their behaviour on our own material. We drew a stratified sample of 40 passages (200 sentences, 2,042 tokens), one per text from 40 of the 100 books, balanced between subcorpora and oversampling verse, and annotated it manually and blind to the system output.

Sentence segmentation was correct for all 200 sentences, including the 40 verse sentences. Tokenisation was correct for 2,039 of 2,042 tokens (99.85%), with no difference between subcorpora ($p = .11$). The volume, lexical-diversity and readability metrics therefore rest on a segmentation and tokenisation layer whose error rate is below 0.2%.

Part-of-speech tagging reached 86.4% (95% CI [82.9, 89.5]), and 92.2% if pairs that UD conventions for Russian leave unsettled (predicatives, determiner-pronouns) are treated as acceptable variants. The most frequent substantive error is characteristic of the genre: sentence-initial capitalised common nouns naming animal or object protago-

nists are read as proper nouns. Accuracy did not differ between subcorpora ($p = .16$), and recomputing the POS ratios from the gold annotation shifts them by at most .023, leaving the between-group differences within the validation sample essentially unchanged (nouns $-.087$ against $-.086$; adjectives $+.017$ against $+.019$). The null results of Section 5.4 cannot be attributed to tagging noise.

A broader rule counting any conj or cc relation reaches only .433 precision (recall .975)—it also fires on sentence-initial discourse conjunctions—and its accuracy differs between subcorpora (.64 against .84, $p = .002$); the predicate-based definition of Section 4.2 reaches .875 with no such difference ($p = .67$). Coordination of full clauses occurs in about 3% of sentences, equally in both groups.

The validation sample is sized to estimate tool accuracy, not to re-test the between-group contrasts; with 100 sentences per subcorpus it is not powered to reproduce them independently.

6 Discussion

Our results reveal a clear pattern. Despite a substantial difference in volume (RUS-O are nearly three times longer than RUS-T), the two groups are indistinguishable on most grammatical, syntactic, and formal parameters. This contrast between

narrative scale and grammatical alignment is the principal linguistic finding of this work.

The paradox between normalized (TTR, hapax) and absolute (Shannon entropy) lexical density follows directly from the volume difference. RUS-O contain more unique vocabulary in absolute terms, but this vocabulary is diluted across a large narrative mass, so on any given page of a RUS-O text the reader encounters fewer novel words than on a page of a RUS-T text. This substantively corresponds to two different narrative strategies: extended, repetitive narration in RUS-O and compact, dense delivery in RUS-T.

The only stable syntactic contrast is the higher proportion of coordinated predicates in RUS-O relative to RUS-T. Mean dependency tree depth does not differ between groups after correction for multiple comparisons.

The methodologically most significant findings are negative. The absence of differences across the four readability indices admits two readings: either translators and editors do in fact align texts on “reading difficulty” (which would be an important observation about editorial practice), or the indices miss substantive differences. Disentangling these two interpretations requires parallel expert validation of the indices on 0+ material, which we regard as a direction for future work.

A similar result obtains for morphology, where 17 morphological metrics, including the basic POS distribution and fine-grained grammatical indices, do not distinguish RUS-O from RUS-T. We interpret this as a trace of the 0+ genre-age standard, in which the grammatical and syntactic profile is set by the text’s address to a very young child rather than by its origin. The absence of morphological differences suggests that translators and editors may actively adapt translations to the norms of Russian children’s literature, a hypothesis that requires direct verification with parallel corpus data.

In terms of translation universals (Baker, 1993; Kunilovskaya et al., 2017), our data give a mixed picture. The simplification hypothesis is partially supported: RUS-T are significantly shorter than RUS-O in words and sentences, which matches one of the expected markers of simplification. However, the classical marker of simplification such as reduced lexical diversity is contradicted by our material: translations show higher values of TTR and hapax ratio. This points to the specificity of the 0+ age niche, where translation strategy does not fully conform to universals derived from adult literature.

7 Conclusion

We have introduced **KidLit v0.1**, the first balanced corpus of contemporary Russian-language 0+ children’s literature, and analysed 35 metrics in the main block and 17 in the extended morphological analysis with Benjamini–Hochberg correction (Benjamini and Hochberg, 1995). After correction, seven effects are robustly significant: three concerning volume (RUS-O are three times longer), three concerning lexical diversity (normalised density higher in RUS-T, absolute density higher in RUS-O), and one concerning syntax (predicate-coordination ratio higher in RUS-O).

Four readability indices, 17 morphological metrics, syntactic depth, and the active-voice proportion do not distinguish the two groups. We regard this negative result as an equally significant contribution, demonstrating that the grammatical and syntactic profile of 0+ children’s literature is set by the genre-age niche rather than by the origin of the text.

Future work. (a) Expanding the corpus to 120+ texts and multivariate modelling of the resulting joint distribution. (b) Building gold-standard annotation layers over KidLit with named entities and sentiment. (c) Domain adaptation: fine-tuning Russian language models on 0+ material and re-running the present comparison with domain-adapted components.

Limitations

Corpus size and stratification. The corpus of 100 texts (50 per group) provides statistical power ≈ 0.85 for effects $d \geq 0.5$ at $\alpha = 0.05$, which is adequate for the main between-group comparisons. However, the corpus does not support reliable subgroup analyses by year of publication, publisher, or source language of translations.

Status of translated texts. The RUS-T subcorpus consists of Russian translations from seven source languages, not original foreign texts. In comparing RUS-O and RUS-T we measure not the difference “Russian vs. other languages” but the difference between two classes of Russian-language texts: those originally written in Russian and those produced through translation and editorial work. All findings are interpreted within this framework.

NLP tool limitations. The tools used (spaCy `ru_core_news_sm` (Honnibal et al., 2020), `pymor-`

phy3 (Korobov, 2015), **razdel** (Natasha Project, 2023)) are trained on SynTagRus (Boguslavsky et al., 2024), which contains predominantly adult news and fiction. Section 5.6 reports a manual validation on 0+ material and bounds the resulting error. It does not cover the extended morphological block, computed with a separate analyser, and the onomatopoeia and diminutive counts in the sample (2 and 8 tokens) are too small to estimate accuracy on those categories.

Applicability of readability indices. The four indices used were originally calibrated on school-age and adult texts, and their Russian adaptations have not been specifically validated on 0+ literature. The negative result may reflect either genuine alignment of texts on “reading difficulty” or insufficient sensitivity of the indices. Disentangling these interpretations requires parallel expert validation.

Univariate comparison model. We use the Mann–Whitney U test as the primary tool for between-group comparison and report Cohen’s d as the effect size. For the 35 simultaneously tested metrics we applied the Benjamini–Hochberg correction ($\text{FDR } q = 0.05$) (Benjamini and Hochberg, 1995) and Holm’s correction (FWER) (Holm, 1979). We do not model the relationships between the metrics. Multivariate treatment (clustering, PCA, classification) is the subject of a companion study on an expanded version of the corpus, where the ratio of observations to features is adequate for such modelling. On 100 texts and 52 metrics we regard univariate testing with explicit correction for multiple comparisons as the defensible choice, and we prefer to report it as such rather than to fit models the present sample cannot support.

Uncontrolled variables. The corpus is balanced by origin type and by age marker, and Section 3 reports the number of publishers and source languages represented. It is not balanced on year of first publication: RUS-O texts are on average more recent than RUS-T texts (mean year 2021.2 vs. 2018.4, $p < .001$), so any effect of changing editorial practice over the period is confounded with origin. Nor is translator identity controlled: several translators contribute more than one text, as do several authors, so observations within each group are not fully independent, which the Mann–Whitney U test assumes. Publisher concentration is low: the largest single imprint accounts for 5 of 50 texts in RUS-O and 7 of 50 in RUS-T. Both constraints fol-

low from the size of the 0+ segment on the Russian market over the period covered, and both call for a substantially larger corpus rather than a different analysis of this one.

Genre composition. The corpus is not stratified by genre subcategory, but the coarse distinction between prose and verse is annotated and does not differ between groups (4 verse texts against 7; Fisher’s exact test, $p = .52$). Since verse texts are shorter than prose in both subcorpora, we verified that the volume contrast is not an artefact of this composition: restricted to prose alone (46 against 43 texts), RUS-O average 2,154 words against 758 in RUS-T, and the difference remains highly significant (Mann–Whitney, $p < .001$). Finer subcategories such as fairy tale, story, didactic book are not annotated, so a residual composition effect cannot be excluded.

Data and Code Availability

The KidLit v0.1 corpus consists of 100 in-copyright books issued by 40 commercial publishers. We do not redistribute the texts. The restriction extends to any representation from which the texts could be reconstructed, including tokenised and morphologically annotated forms. Everything needed to reproduce the analysis is released at <https://github.com/dariagavrilic-creator/KidLit-v0.1>: the bibliographic metadata of all 100 texts, the per-text matrices of all metrics reported here, the processing and statistics notebooks, and the records of the manual validation of Section 5.6. Every statistic in this paper can be recomputed from those files without access to the texts. The pipeline depends on nothing specific to our sample: it uses openly available components and can be applied unchanged to any comparable collection of 0+ literature in Russian, or, with the corresponding language models substituted, in another language. We regard the method and the linguistic profile it yields, rather than the text collection itself, as the transferable contribution of this work.

References

- Mona Baker. 1993. *Corpus linguistics and translation studies: Implications and applications*. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology: In Honour of John Sinclair*, pages 233–250. John Benjamins, Amsterdam.

- Yoav Benjamini and Yosef Hochberg. 1995. [Controlling the false discovery rate: A practical and powerful approach to multiple testing](#). *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.
- Igor M. Boguslavsky, Pavel V. Dyachenko, Elena S. Inshakova, Leonid L. Iomdin, Alexey V. Lazursky, Leonid G. Mityushin, Alina A. Movsesyan, Igor P. Rygaev, Vladimir G. Sizov, Sergey P. Timoshenko, Tatiana I. Frolova, and Anastasia V. Chaga. 2024. The current state of the SynTagRus corpus. *Proceedings of the V. V. Vinogradov Russian Language Institute*, (4(42)):141–169. In Russian.
- Anna V. Glazkova. 2021. Book age rating and readability level: On the relationship between two indicators. *Tekst. Kniga. Knigoizdanie*, (26):129–142. In Russian.
- Sture Holm. 1979. [A simple sequentially rejective multiple test procedure](#). *Scandinavian Journal of Statistics*, 6(2):65–70. Accessed 10 August 2026.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength natural language processing in Python](#). <https://spacy.io>. Model ru_core_news_sm 3.7.0. Accessed 10 August 2026.
- Boris L. Iomdin and Dmitry A. Morozov. 2021. [Who can understand “Dunno”? automatic assessment of text complexity in children’s literature](#). *Russkaya Rech’*, (5):55–68. In Russian.
- Mikhail Korobov. 2015. [Morphological analyzer and generator for Russian and Ukrainian languages](#). In *Analysis of Images, Social Networks and Texts (AIST 2015)*, volume 542 of *Lecture Notes in Computer Science*, pages 320–332, Cham. Springer. Library version: pymorphy3 2.0.1. Accessed 10 August 2026.
- Maria Korochkina, Marco Marelli, Marc Brysbaert, and Kathleen Rastle. 2024. [The children and young people’s books lexicon \(CYP-LEX\): A large-scale lexical database of books read by children and young people in the United Kingdom](#). *Quarterly Journal of Experimental Psychology*, 77(12):2418–2438.
- Maria Kunilovskaya, Tatyana Ilyushchenya, and Marina Kovyazina. 2017. Prospects for studying the linguistic features of Russian translated discourse. *PNRPU Linguistics and Pedagogy Bulletin*, (2):17–27. In Russian.
- Gillian Lathey. 2010. *The Role of Translators in Children’s Literature: Invisible Storytellers*. Routledge, New York.
- Sara Laviosa. 2002. *Corpus-based Translation Studies: Theory, Findings, Applications*. Rodopi, Amsterdam.
- Olga N. Lyashevskaya and Sergey A. Sharov. 2009. [Frequency Dictionary of Modern Russian \(Based on the Russian National Corpus\)](#). Azbukovnik, Moscow. Accessed 10 August 2026. In Russian.
- Kirill Maslinsky, Evgenia Lekarevich, Lyubov Aleinik, and Anna Kolotovkina. 2021. Corpus of Russian prose for children and adolescents (DetCorpus). <https://dataverse.pushdom.ru/dataset.xhtml?persistentId=doi:10.31860/openlit-2021.4-C001>. Accessed 10 August 2026.
- Mistral AI. 2024. Mistral OCR API. https://docs.mistral.ai/capabilities/OCR/document_ocr/. Accessed 10 August 2026.
- Dmitry A. Morozov. 2022. books-n-kids: Russian children’s literature complexity classifier. <https://github.com/morozowdmitry/books-n-kids>. Accessed 10 August 2026.
- Dmitry A. Morozov, Anna V. Glazkova, and Boris L. Iomdin. 2022. [Text complexity and linguistic features: Their correlation in English and Russian](#). *Russian Journal of Linguistics*, 26(2):426–448.
- Natasha Project. 2023. razdel: Rule-based token and sentence segmentation for the Russian language. <https://github.com/natasha/razdel>. Accessed 10 August 2026.
- Maeve Olohan. 2004. *Introducing Corpora in Translation Studies*. Routledge, London.
- Emer O’Sullivan. 2005. *Comparative Children’s Literature*. Routledge, London.
- Svetlana O. Savchuk, Timofey A. Arkhangelskiy, Anastasia A. Bonch-Osmolovskaya, Olga V. Donina, Julia N. Kuznetsova, Olga N. Lyashevskaya, Boris V. Orekhov, and Maria V. Podryadchikova. 2024. The Russian National Corpus 2.0: New opportunities and development prospects. *Voprosy Jazykoznanija*, (2):7–34. In Russian.
- Lev S. Vygotsky. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA.

A Full results, main block

Group	Metric	RUS-O	RUS-T	d	p_{raw}	p_{adj}
Volume	Word count	2,067	698	+0.91	< .001	< .001 ***
	Sentence count	278	91	+0.77	< .001	< .001 ***
	Pages	41.7	34.6	+0.42	.001	.005 **
Lexical diversity	Shannon entropy	8.20	7.20	+1.10	< .001	< .001 ***
	TTR	0.573	0.664	−0.85	< .001	.001 **
	Hapax ratio	0.423	0.526	−0.82	< .001	.001 **
	STTR	0.621	0.669	−0.53	.016	.071 ns
	Yule K index	248.9	248.3	+0.00	.622	.814 ns
	Top-1000 ratio	0.358	0.360	−0.04	.652	.814 ns
POS profile	Adverbs	0.086	0.096	−0.49	.028	.110 ns
	Adjectives	0.076	0.067	+0.37	.051	.166 ns
	Proper nouns	0.052	0.061	−0.32	.118	.274 ns
	Animate nouns	0.314	0.343	−0.30	.278	.540 ns
	Pronouns	0.087	0.094	−0.24	.424	.707 ns
	Verbs	0.192	0.189	+0.13	.682	.823 ns
	Nouns	0.235	0.236	−0.04	.904	.989 ns
Syntax	Predicate coord.	0.251	0.177	+0.79	< .001	.002 **
	Tree depth	3.828	3.699	+0.29	.159	.347 ns
Readability	FOG index	14.901	14.163	+0.36	.190	.392 ns
	Flesch (RU)	29.837	32.298	−0.26	.365	.638 ns
	Coleman–Liau	12.847	12.598	+0.11	.839	.979 ns
	ARI	8.757	8.484	+0.16	.904	.989 ns
Word & sentence	Sent. len. (w)	8.358	7.548	+0.39	.052	.166 ns
	Sent. len. (ch)	53.7	48.6	+0.37	.068	.184 ns
	Syllables/word	1.992	1.972	+0.19	.493	.738 ns
	Word length	4.996	4.994	+0.01	.637	.814 ns
Grammar	Past tense	0.577	0.476	+0.55	.067	.184 ns
	Present tense	0.208	0.285	−0.48	.087	.218 ns
	Future tense	0.044	0.047	−0.09	.341	.629 ns
	Interrogative sent.	0.080	0.096	−0.25	.506	.738 ns
	Exclamatory sent.	0.245	0.267	−0.16	.475	.738 ns
	Imperative mood	0.029	0.032	−0.12	.544	.761 ns
Additional	Active voice	0.765	0.769	−0.07	.986	1.000 ns
	Punct. diversity	0.769	0.754	+0.12	.985	1.000 ns

Table 3: Complete results, main block (34 metrics, Mann–Whitney U test, BH FDR $q = 0.05$). Sign of d : “+” higher in RUS-O, “−” higher in RUS-T. $N = 50$ per group. *** $p < .001$, ** $p < .01$, ns = not significant after correction. The personal-pronoun metric returned zero in both subcorpora, which reflects a feature-extraction artefact rather than a property of the texts; it is reported as non-significant and was retained in the correction family, so BH was applied to 35 tests, leaving 34 substantive metrics for discussion.

B Extended morphological analysis

Metric	RUS-O	RUS-T	d	p_{adj}
Analyticity index	.1759	.1632	+.579	.080
Adjectivity index	.0550	.0482	+.376	.358
Short adjectives	.0079	.0094	-.378	.358
Predicatives	.0051	.0066	-.330	.442
Genitive case	.0509	.0470	+.245	.635
Pronominality index	.0880	.0932	-.252	.635
Verbality index	.1464	.1428	+.191	.726
Infinitives	.0213	.0228	-.195	.726
Noun-to-verb ratio	1.582	1.633	-.129	.802
Nominal vocab. index	.2824	.2794	+.075	.802
Substantivity index	.2246	.2284	-.111	.802
Full participles	.0026	.0024	+.086	.802
Gerunds	.0033	.0030	+.079	.802
Short participles	.0011	.0012	-.076	.802
Particles	.0349	.0335	+.130	.802
Numerals	.0053	.0051	+.047	.864
Instrumental case	.0280	.0283	-.027	.893

Table 4: Extended morphological analysis (17 metrics, two-sample t -test, BH FDR $q = 0.05$ applied independently from Table 3). All effects non-significant after correction. Sorted by p_{adj} .

C Readability distributions

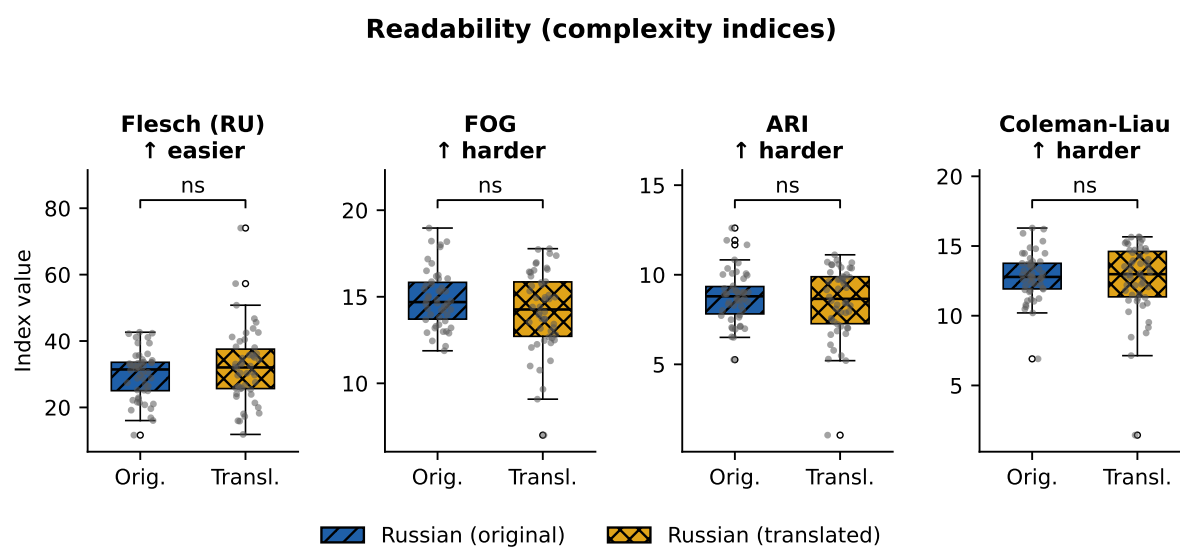


Figure 4: Readability indices: RUS-O vs. RUS-T (Flesch RU, FOG, ARI, Coleman-Liau). No index is significant after BH FDR correction (all $p_{\text{adj}} > 0.39$).