First Insights into the Syntax of Slovene Student Writing: A Statistical Analysis of Šolar 3.0 vs. Učbeniki 1.0

Tina Munda
Centre for Language Resources and Technologies,
University of Ljubljana
Večna pot 113, Ljubljana
tina.munda@cjvt.si

Špela Arhar Holdt
Faculty of Arts,
University of Ljubljana
Aškerčeva cesta 2, Ljubljana
spela.arharholdt@ff.uni-lj.si

Abstract

This study investigates the syntactic features of Slovene student writing by comparing essays from the Šolar 3.0 corpus (ages 13–19; primary and secondary school levels) with textbook texts from the Učbeniki 1.0 corpus aligned to the same educational stages. We apply quantitative syntactic analysis at two complementary levels: clause-type frequency (coordination, parataxis, and four types of subordination) and tree-based syntactic complexity measures (number of clauses, clauses per T-unit, and maximum parse-tree depth). Results show that students heavily rely on coordination and specific subordinate clauses (especially object and adverbial), producing more clauses per sentence and per T-unit than textbooks. However, their sentences tend to exhibit flatter syntactic structures, with shallower embedding in primary school and only modest increases in tree depth by secondary school. These findings reveal a divergence between surface-level complexity and hierarchical depth, highlighting developmental trends and instructional targets in written syntactic maturity. We conclude by discussing implications for syntactic development and directions for future research.

1 Introduction

In recent years, the availability of large scale corpora annotated with syntactic relations, improved accuracy of automatic parsing tools, and user-friendly software for corpus-data extraction have opened new opportunities for syntactic research based on automatically processed linguistic data.

This paper investigates the use of syntactic structures in the writing of Slovene primary (age 13–15) and secondary (age 15–19) school students, using the developmental Šolar 3.0 corpus (Arhar Holdt and Kosem, 2024), and compares it to the syntactic patterns found in the corpus of Slovene textbooks—Učbeniki 1.0 (Kosem and Pori, forthcoming).

The Šolar 3.0 corpus comprises 5,485 Slovene texts (1,635,407 words) produced by pupils in grades 6–9 of Slovene primary school (ages 13–15) and by students in Slovene secondary school (ages 15–19). Its composition reflects the variety of classroom genres: essays (58.7%), classroom exercises (15.0%), practical texts (12.6%), and tests (13.7%). Just under one fifth of the texts (19.7%) come from primary students, with the remainder (80.3%) authored by secondary-school students. Most texts (85.4%) were written in Slovene language classes. Although teacher corrections are available for 38.18% of the corpus, our analysis uses the original student productions to capture authentic student syntax. To ensure better comparability in genre and communicative purpose, we restrict our analysis to texts labeled as *esej ali spis* (essay), which constitute the majority text type in Šolar.

The Učbeniki 1.0 corpus consists of 127 Slovene-language textbooks (4,302,857 words) covering grades 1–9 of primary school and all years of secondary school, across 16 subjects. Primary school textbooks make up 71.6% of the corpus, with secondary school titles accounting for 28.4%. For direct comparability with Šolar 3.0, we restrict our analysis to grades 6–9 of primary school and all secondary school textbooks, excluding readers and early grade textbooks. This alignment ensures that both corpora reflect the pedagogical materials and student texts relevant to the same educational stages.

Although the two corpora differ in genre, comparing them is sensible: textbooks—written for these age and comprehension levels—represent the intended or desired level of language competence, while students' texts reveal their actual writing skills. We cannot, however, be sure that textbook authors intentionally or successfully tailored their language to the target age group.

Situated within the domain of literacy develop-

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 105–114
August 29, 2025 ©2025 Association for Computational Linguistics

ment, this study aims to provide a clearer picture of how syntactic competence manifests in student writing compared to the standard Slovene patterns promoted through pedagogical materials. The analysis relies on corpus data (Munda et al., 2025a; Munda et al., 2025b) obtained via the Universal Dependencies[1] (UD) framework and represents the first computational syntactic study of student writing in the Slovene context. As a starting point, we ask: To what extent do student texts differ syntactically from textbook texts, and how can these differences be meaningfully interpreted? Can such insights yield pedagogically useful guidance for developing writing competence? To answer these questions, our quantitative analysis proceeds on two complementary levels:

- Clause-level frequency comparison: we compare the raw and normalized frequencies of coordination, parataxis, and four subordination types (subject, object, adverbial, and relative clauses) across Šolar 3.0 and Učbeniki 1.0, separately for primary and secondary school subsets.

- Tree-based syntactic complexity profiling: to capture the deeper structural features of students' sentences, we compute three tree-based syntactic complexity measures—number of clauses, clauses per T-unit, and maximum parse tree depth—derived from UD parses.

Together, these approaches yield a richer account of student syntactic development. The findings offer practical implications for language instruction, highlighting which clause types and complexity dimensions merit explicit teaching, and to what extent students' real world usage aligns with textbook models. This work was conducted within the framework of the *Empirical foundations for digitally-supported development of writing skill*[2] project, which supports applied linguistic research for educational development.

The remainder of this paper is organized as follows. Section 2 provides related work. Section 3 describes data-preparation procedures and the methodology for the quantitative syntactic analyses, including clause-type frequency counts and tree-based syntactic complexity measures. Section 4 reports the results of these analyses for primary and secondary school subsets. Section 5 offers a discussion of our findings. Finally, Section 6 provides our conclusion and outlines avenues for future work.

2 Related Work

Computational syntactic analysis of Slovene corpora has only recently gained significant momentum, primarily due to advancements in syntactic parsing following the Universal Dependencies (UD) framework. This cross-linguistically consistent annotation scheme, having been developing for over the past decade, has enabled more sophisticated analyses of Slovene syntactic structures. The current version of the CLASSLA-Stanza parser (Ljubešić et al., 2024) has achieved remarkably high accuracy rates of 95.54% for UD-relations in Slovene texts, making large-scale automated syntactic analysis increasingly reliable.

Recent studies have begun to leverage UD-annotated corpora for Slovene linguistic research. Dobrovoljc (2024) explored the potential and limitations of UD-relations for analyzing spoken Slovene, highlighting the adaptability of the framework to various contexts. Terčon (2024) demonstrated the value of UD annotations for comparative syntactic analysis by automatically measuring syntactic complexity differences between written and spoken Slovene corpora.

Research on the Šolar developmental corpus has thus far focused primarily on lexical aspects of student writing. Rozman et al. (2018) analyzed collocations occurring within the corpus, while Gantar and Bon (forthcoming) investigated multi-word lexical challenges faced by students in Slovene primary and secondary schools. Arhar Holdt and Rozman (2015) examined the most frequently corrected verbal and pronominal lemmas in the corpus, aiming to develop data-driven instructional materials responsive to actual student needs.

By contrast, syntactic dimensions of student writing remain largely unexplored. This is a notable gap given the importance of syntactic development in educational contexts and the potential for corpus-based findings to inform pedagogical practice. The present study addresses this gap by providing the first quantitative analysis of Slovene student syntax compared to textbook models, leveraging the analytical power of UD annotations to reveal developmental patterns in syntactic complexity and clause usage across educational levels.

Although Slovene-based studies are scarce, re-

[1] https://universaldependencies.org/

[2] https://www.cjvt.si/prop/en/

search on as a first language (L1) writing provides valuable developmental benchmarks and comparative context for assessing syntactic growth across educational levels.

In their studies of adolescent writing in English L1 in the United States, Beers and Nagy (2009); (2011) examined clause density—measured as clauses per T-unit—among other syntactic complexity measures, across various essay types. For Grades 7 and 8 students (ages 12–14), Beers and Nagy (2009) reported a mean of 1.5 clauses per T-unit in narrative essays and 2.0 in persuasive essays. In a broader study of types of school writing, Beers and Nagy (2011) analyzed texts from Grades 3, 5, and 7, finding that Grade 7 (ages 12–13) clause density ranged from 1.21 in descriptive texts to 2.08 in persuasive essays. Since the Šolar corpus essays are not subdivided by types, the average of the four essay types reported by Beers and Nagy (2011)—1.46 clauses per T-unit—offers a useful reference point for interpreting clause density at the primary-school level in the present study. A foundational study by Hunt (1970) further supports these benchmarks. His analysis of 8th-grade (ages 13–14) student writing revealed an average of 1.42 clauses per T-unit, indicating slightly lower syntactic complexity in U.S. school settings compared to the average across essay types reported by Beers and Nagy.

As for the secondary level, Hunt (1970) reported a mean of 1.68 clauses per T-unit for 12th-grade students (ages 17–18), suggesting moderate syntactic growth in late adolescence. This figure provides a relevant benchmark for evaluating clause density in the secondary-school subset of the present study.

These studies also emphasize two important caveats: first, that increased syntactic complexity does not always correlate with higher writing quality; and second, that genre exerts a significant influence on syntactic structures. Different genres give rise to distinct syntactic realizations, depending on their communicative goals. Therefore, syntactic complexity should be interpreted in light of genre conventions rather than as a standalone indicator of writing development.

3 Methodology

3.1 Data Preparation

The syntactic data (Munda et al., 2025a; Munda et al., 2025b) were extracted in advance from both corpora using the STARK tool (Krsnik and Dobrovoljc, 2025), which leverages morphosyntactic and dependency annotations following the UD schema.

To enable direct comparison by school level, educational level metadata were added to each textbook parse file, allowing the pedagogical corpus to be split into primary- and secondary-school subsets. The Šolar corpus already contained this information, so the same split was applied there. In addition, only the syntactic data from texts tagged as *esej ali spis* (essay) were retained from Šolar 3.0, with other text types excluded to reduce structural variability.

After this step, every syntactic structure appears in four subsets:

- Šolar: primary-school student texts (šolar_PS)
- Šolar: secondary-school student texts (šolar_SS)
- Učbeniki: primary-school textbooks (učb_PS)
- Učbeniki: secondary-school textbooks (učb_SS)

These form the basis of all subsequent quantitative analyses; see Table 1 for their size.

Subcorpus	Size (words)
šolar_PS	195,233
šolar_SS	1,075,409
učb_PS	2,039,313
učb_SS	1,252,755

Table 1: Size in words of the four subcorpora of Šolar 3.0 (šolar) and Učbeniki 1.0 (učb) for primary (PS) and secondary (SS) school levels.

3.2 Quantitative Syntactic Analysis

3.2.1 Clause-Type Frequency Analysis

We compare the frequencies of major relations at the clause level—coordination, parataxis, and four subordination types (subject, object, adverbial, and relative clauses)—across the four data subsets. Observed counts of each relation were tabulated, then normalized per 1,000 tokens to adjust for corpus-size differences.

To test whether students and textbooks differ in their use of each structure, we applied the Chi-square test (χ^2) on 2 × 2 contingency tables (developmental vs. pedagogical), conducted separately for primary and secondary levels. We report $p < 0.0001$ for all comparisons and calculate the Phi coefficient (ϕ) as an effect-size measure to quantify the magnitude of the association between corpus type and each syntactic structure.

3.2.2 Tree-Based Syntactic Complexity Measures

To capture sentence-level structural complexity, we compute three tree-based syntactic complexity metrics on every sentence in each subset. These measures are modeled after Terčon (2024), who applied them to compare syntactic complexity across written and spoken registers of Slovene. These include:

- NR_OF_CLAUSES: total number of finite and non-finite clauses per sentence.
- CLAUSES_PER_T-UNIT: clause density, calculated as the number of clauses divided by the number of T-units in a sentence, indicating how many subordinate or coordinate clauses are packed into each minimal 'idea' unit.
- MAX_TREE_DEPTH: height of the dependency tree, measuring embedding depth; put simply: the largest number of levels from any word up to the main verb in the sentence's dependency tree, showing how many nested syntactic layers the sentence has.

We excluded the three token-based measures—Mean Dependency Distance, Normalized Dependency Distance and Words per Sentence—that were included in Terčon (2024), because the textbook corpus contains noisy segmentation that distorts token-based calculations. The three selected measures, by contrast, rely on parse-tree structure and are more robust to segmentation errors.

For each metric, we first compute descriptive statistics (mean, standard deviation, median) for Šolar vs. Učbeniki at each school level. To assess group differences, we run Mann–Whitney U tests (non-parametric) and report rank-biserial correlations as effect sizes. Finally, to control the overall error rate across the three comparisons at each level, we apply the Holm–Bonferroni correction to the raw p-values.

4 Results

In this section, we present the results of the quantitative analyses, organized by educational level. We first report findings for the primary-school texts, followed by those from the secondary-school level. At each level, we examine both clause-type frequencies and tree-based syntactic complexity measures.

4.1 Primary-School Level

4.1.1 Clause-Type Frequency Analysis

The quantitative analysis of syntactic structures at the primary-school level (see Table 2) shows clear differences between the developmental corpus (Šolar 3.0) and the textbook corpus (Učbeniki 1.0). Coordination structures (*conj*) occur at a notably higher normalized frequency per 1,000 tokens in the Šolar corpus (31.64) compared to textbooks (11.72). Subordination structures also exhibit higher normalized frequency in student texts (38.41) relative to textbook texts (22.54). Among subordinate clause types, adverbial clause *(advcl)* and object clause (*ccomp*) stand out with notably higher normalized frequencies in student texts (16.59 and 11.63, respectively) compared to textbooks (7.54 and 3.21, respectively). Relative clauses (*acl*), however, show relatively similar frequencies across both corpora (9.29 vs. 11.21). Parataxis (*parataxis*) structures appear slightly less frequently in student texts (13.65) compared to textbooks (14.74).

Chi-square tests confirm these differences are statistically significant (see Table 2). The strongest associations, as indicated by Phi (ϕ), appear for coordination (ϕ=0.0488), object (ϕ=0.0379) and adverbial clauses (ϕ=0.0281). Relative and subject clauses, and parataxis show smaller effect sizes, indicating minimal differences between students and textbooks, despite statistical significance.

4.1.2 Tree-Based Syntactic Complexity Measures

Tree-based syntactic complexity measures reveal additional structural insights into primary-school-level sentence constructions (see Table 3, Figure 1).

The number of clauses per sentence is notably higher in the Šolar corpus (mean=2.22) than in textbooks (mean=1.73). Clause density (number of clauses per T-unit) is slightly higher in student texts (mean=1.45) than in textbooks (mean=1.29). These differences are visually evident in the corresponding boxplots (Figure 1), where student data show a broader range and higher median values for both measures, indicating more frequent use of multi-clause structures and denser clause packaging in learner writing.

Interestingly, the mean maximum tree depth—a measure of syntactic embedding—does not differ

Structure	šolar_PS		učb_PS		Φ
	Ofq	Nfq	Ofq	Nfq	
coordination	6,177	31.64	23,905	11.72	0.0488
subordination	7,499	38.41	45,973	22.54	0.0293
subject c.	177	0.91	1,196	0.59	0.0036
object c.	2,270	11.63	6,544	3.21	0.0379
adverbial c.	3,239	16.59	15,369	7.54	0.0281
relative c.	1,813	9.29	22,864	11.21	0.0052
parataxis	2,665	13.65	30,050	14.74	0.0025
Total	16,341	83.7	99,928	49.00	

Table 2: Observed (Ofq) and normalized (Nfq; per 1,000 tokens) frequencies of syntactic structures in the primary school (PS) subset of Šolar 3.0 and Učbeniki 1.0, along with Phi (Φ) effect sizes for differences in syntactic structure use across both corpora in the primary school (PS) subset. All differences are statistically significant (χ^2, p < 0.0001).

substantially between corpora (4.10 for Šolar vs. 4.28 for textbooks), with nearly identical medians and overlapping interquartile ranges. This suggests that while students produce more clauses, they do not build significantly deeper syntactic structures.

Mann–Whitney U tests confirm statistically significant differences for clause-related metrics (U=1.84×10^9, r_{rb}=–0.363* for number of clauses; U=1.68×10^9, r_{rb}=–0.260* for clauses per T-unit), but not for maximum tree depth (U=1.45×10^9, r_{rb}=0.005). These findings indicate that student essays are structurally more clause-heavy, but not more syntactically embedded than textbooks.

4.2 Secondary-School Level

4.2.1 Clause-Type Frequency Analysis

The syntactic-structure frequencies at the secondary-school level (see Table 4) also exhibit notable differences between the Šolar and textbook corpora. Coordination structures again occur more frequently in the Šolar corpus (30.85 per 1,000 tokens) compared to textbooks (11.16). Similarly, overall subordination structures are used more frequently by students (44.67) than in textbooks (23.00). Among subordinate clause types, adverbial clause (16.23) and object clause (11.99) remain prominently higher in students' texts compared to textbooks (7.17 and 2.94, respectively). Relative clauses have slightly higher frequencies in student texts (14.53) compared to textbooks (12.29), while parataxis structures appear slightly less often in the Šolar corpus (12.37 vs. 14.28 in textbooks).

Chi-square tests confirm the statistical significance of these differences (see Table 4). Coordination (Φ=0.0697), subordination (Φ=0.0605), object clause (Φ=0.0537), and adverbial clauses (Φ=0.0426) show moderate effect sizes, reflecting stronger corpus differences. Relative clause, subject clause, and parataxis have minimal effect sizes despite statistical significance.

Compared to the primary school subset, coordination and subordination frequencies remain similarly elevated in student texts, but effect sizes are notably higher in the secondary-school data—suggesting a stronger divergence in how advanced students structure their writing. Additionally, whereas relative clause frequencies were nearly identical across corpora in the primary school subset, secondary-school students use relative clauses more frequently, though the effect size remains small.

4.2.2 Tree-Based Syntactic Complexity Measures

At the secondary-school level, tree-based syntactic complexity measures show nuanced differences (see Table 5, Figure 2).

Student writing features a markedly higher number of clauses per sentence (mean=2.66 vs. 1.78), indicating a greater tendency to produce multi-clause constructions. In addition, students demonstrate higher clause density, measured as clauses per T-unit (mean=1.62 vs. 1.31), suggesting that more subordinate or coordinate clauses are packed into individual minimal idea units. These findings are supported by the boxplots (Figure 2), which reveal a broader distribution and higher medians for both the number of clauses and clause density in student texts.

Interestingly, mean maximum tree depth is also slightly higher in student texts (4.84 vs. 4.56),

Measure	šolar_PS			učb_PS			U ($\times 10^9$)	r_{rb}
	mean	sd	median	mean	sd	median		
NR_OF_CLAUSES	2.22	1.36	2	1.73	1.36	1	1.84	-0.363*
CLAUSES_PER_T-UNIT	1.45	0.66	1	1.29	0.55	1	1.68	-0.260*
MAX_TREE_DEPTH	4.10	1.43	4	4.28	2.21	4	1.45	0.005

Table 3: Tree-based syntactic complexity measures for the primary school (PS) subset of the Šolar 3.0 (šolar) and Učbeniki 1.0 (učb) corpora. Values are reported as mean, standard deviation (sd), and median per sentence. Mann–Whitney U test statistics (scaled $\times 10^9$) and rank biserial correlations (r_{rb}) are included. All comparisons are statistically significant at $p < 0.0001$ after Holm–Bonferroni correction. Asterisks indicate comparisons that remain significant at $\alpha = 0.05$.

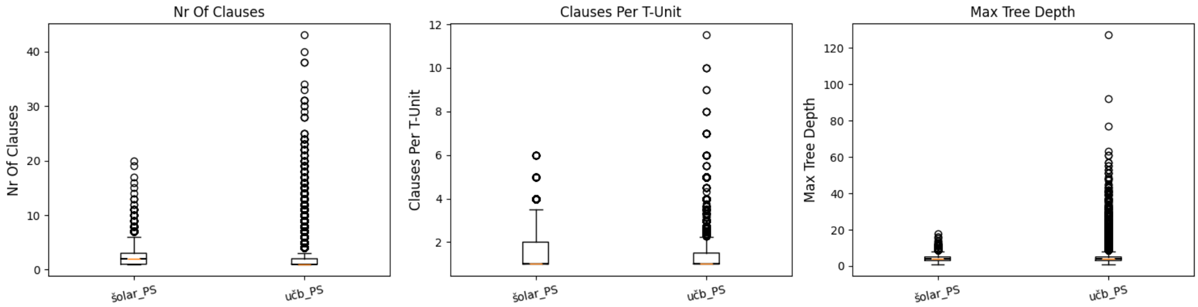

Figure 1: Boxplots of three tree-based syntactic complexity measures—number of clauses, clauses per T-unit, and maximum tree depth—for primary school texts (Šolar_PS vs. Učbeniki_PS). The figure illustrates higher clause density in student writing and a wider range of tree depths and clause counts in textbook data, likely due to segmentation noise.

although the difference is less pronounced and medians remain identical (both 5). However, the broader range and right-skewed distribution in the textbook corpus—reflected in more extreme outliers—suggest occasional structurally deeper sentences, probably due to segmentational noise.

Mann–Whitney U tests confirm statistically significant differences for all three measures (p < 0.0001, adjusted using Holm–Bonferroni correction). The effect sizes (rank-biserial correlation, r_{rb}) are strongest for the number of clauses (–0.363) and clauses per T-unit (–0.260), while the difference in maximum tree depth shows a smaller effect (–0.111), consistent with the visual overlap in distributions.

5 Discussion

This study provides novel insights into syntactic patterns characterizing Slovene students' written language in comparison to pedagogical patterns represented in textbooks. By employing a dual-methodological framework—quantitative analyses of clause-type frequencies and tree-based syntactic complexity measures—we uncover distinct patterns in how students utilize syntactic structures at both primary and secondary school levels.

Before interpreting these findings, it is important to consider the genre makeup of each corpus. The Šolar 3.0 corpus used in this study includes only student essays, while the textbook corpus encompasses a broader mix of descriptive, expository, and narrative texts. This genre imbalance partially explains clause chaining in student writing, which may amplify the frequency of coordination and subordination in comparison to textbooks.

5.1 Primary School: Student Texts vs. Textbooks

At the primary school level, students show a clear preference for coordination and subordination strategies. Compared to textbooks, student texts contain more than double the normalized frequencies of coordination (31.64 vs. 11.72) and adverbial clauses (16.59 vs. 7.54), with relative and object clauses also more frequent. These differences are statistically significant across all clause types, with Phi effect sizes ranging from small to moderate (Φ=0.0025–0.0488). This suggests that primary students are already making active use of a range of clause-linking devices in their writing.

Tree-based syntactic complexity measures reveal more nuanced structural patterns. While student sentences contain more clauses per sentence (2.22 vs. 1.73) and show greater clause

Structure	šolar_SS		učb_SS		Φ
	Ofq	Nfq	Ofq	Nfq	
coordination	33,176	30.85	13,983	11.16	0.0697
subordination	48,035	44.67	28,809	23	0.0605
subject c.	2,068	1.92	742	0.59	0.0191
object c.	12,897	11.99	3,684	2.94	0.0537
adverbial c.	17,449	16.23	8,988	7.17	0.0426
relative c.	15,621	14.53	15,395	12.29	0.0097
parataxis	13,307	12.37	17,891	14.28	0.0083
Total	94,518	87.89	60,683	48.44	

Table 4: Observed (Ofq) and normalized (Nfq; per 1,000 tokens) frequencies of syntactic structures in the secondary school (SS) subset of Šolar 3.0 and Učbeniki 1.0, along with Phi (Φ) effect sizes for differences in syntactic structure use across both corpora in the secondary school (SS) subset. All differences are statistically significant (χ^2, p < 0.0001).

density as measured by clauses per T-unit (1.45 vs. 1.29), their maximum tree depth is slightly lower (4.10 vs. 4.28). Both clause metrics show significant differences (r_{rb} = –0.363 and –0.260), while tree depth does not. These results indicate that primary-school students tend to produce more linear clause sequences and pack more information per idea unit, but do so using relatively shallow structures—favoring additive linking and minimally embedded subordination over hierarchical nesting.

In terms of cross-linguistic reference points, the mean clause density of 1.45 clauses per T-unit observed in Slovene primary-school essays closely aligns with values reported for English L1 adolescent writing. Beers and Nagy (2011) found that Grade 7 students (ages 12–13) produced between 1.21 and 2.08 clauses per T-unit depending on text type, with an average of 1.46 across four text types. Hunt (1970) similarly reported a mean of 1.42 for 8th-grade students (ages 13–14). These parallels suggest that Slovene students in this age range exhibit comparable syntactic development in terms of clause density, despite language-specific and curricular differences.

5.2 Secondary School: Student Texts vs. Textbooks

In secondary school, the same overall trends persist but become more pronounced. Students continue to use coordination (30.85 vs. 11.16) and subordination (44.67 vs. 23.00) far more frequently than textbooks. This includes object (11.99 vs. 2.94) and adverbial clauses (16.23 vs. 7.17) both of which exhibit moderate effect sizes (Φ=0.0537 and 0.0426). These elevated frequencies suggest that students are increasingly using complex sentence structures to develop arguments and articulate relationships between ideas.

Tree-based syntactic complexity measures reinforce this pattern, showing significantly more clauses per sentence (2.66 vs. 1.78) and higher clause density (1.62 vs. 1.31) in student writing, with large effect sizes (r_{rb}=–0.363 and –0.260). In contrast to the primary level, however, maximum tree depth is also greater in student texts than in textbooks (4.84 vs. 4.56), and while the effect size is smaller (r_{rb}=–0.111), the difference remains statistically significant. These findings point to a developmental progression: secondary-school students not only use more clauses but also begin to imbed them more hierarchically—an indication of growing syntactic proficiency and control.

These clause density findings correspond well with those reported for English L1 writing at this educational level. Hunt (1970) found that 12th-grade students (ages 17–18) produced a mean of 1.68 clauses per T-unit, which aligns closely with the 1.62 observed in Slovene secondary-school essays. This supports the interpretation that clause-per-T-unit density may reflect a broader developmental milestone in adolescent writing, observable across typologically different languages.

5.3 Developmental Trends across School Levels

A comparison of student writing across school levels reveals a developmental trajectory in syntactic maturity. From primary to secondary school, the mean number of clauses per sentence increases (2.22 to 2.66), as does clause density per T-unit (1.45 to 1.62), reflecting a growing tendency to

Measure	šolar_SS			učb_SS			U ($\times 10^9$)	r_{rb}
	mean	sd	median	mean	sd	median		
NR_OF_CLAUSES	2.66	1.67	2	1.78	1.46	1	4.82	-0.363*
CLAUSES_PER_T-UNIT	1.62	0.77	1.5	1.31	0.57	1	4.45	-0.260*
MAX_TREE_DEPTH	4.84	1.72	5	4.56	2.11	4	3.93	-0.111*

Table 5: Tree-based syntactic complexity measures for the secondary school (SS) subset of the Šolar 3.0 (šolar) and Učbeniki 1.0 (učb) corpora. Values are reported as mean, standard deviation (sd), and median per sentence. Mann–Whitney U test statistics (scaled $\times 10^9$) and rank biserial correlations (r_{rb}) are included. All comparisons are statistically significant at $p < 0.0001$ after Holm–Bonferroni correction. Asterisks indicate comparisons that remain significant at $\alpha = 0.05$.

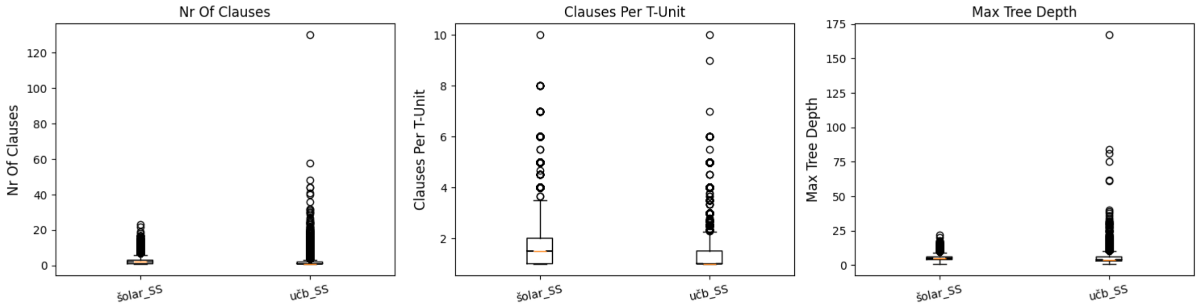

Figure 2: Boxplots of three tree-based syntactic complexity measures—number of clauses, clauses per T-unit, and maximum tree depth—for secondary school texts (šolar_SS vs. učb_SS). The figure illustrates higher clause density in student writing and a wider range of tree depths and clause counts in textbook data, likely due to segmentation noise.

elaborate ideas through clause integration.

At the same time, maximum tree depth rises only modestly (4.10 to 4.84), suggesting that while older students use more clauses, their embedding depth increases more gradually. This indicates that syntactic development may initially proceed through horizontal expansion—adding and linking clauses—before progressing to deeper hierarchical structuring.

Taken together, these developmental trends support the interpretation that Slovene students, as they mature, expand their syntactic repertoire primarily by increasing clause quantity and density, rather than embedding complexity. Instructional efforts might focus on helping students transition from additive, chain-like structures toward more varied and hierarchically integrated syntax.

It is important to interpret these results in light of the reference standard. While in the study, textbooks serve as a syntactic benchmark, they do not represent a neutral writing model. Sentence structures may be intentionally simplified for pedagogical clarity, which can inadvertently limit syntactic exposure and under-challenge learners. The observed disparities thus reflect not only student development but also the editorial and instructional conventions that shape textbook style.

6 Conclusion and Future Work

6.1 Conclusion

This study provides the first large-scale syntactic comparison between Slovene student writing and textbook models, using dependency-parsed data from the Šolar 3.0 and Učbeniki 1.0 corpora. Focusing on essays produced by students in Slovene primary and secondary school, we examined clause-type usage and three tree-based syntactic complexity measures: number of clauses, clauses per T-unit, and maximum tree depth.

The results reveal consistent differences between student writing and textbook texts across educational levels. Slovene students in both primary and secondary school employed significantly more coordination and subordination—particularly object and adverbial clauses—than textbooks, reflecting the clause-chaining tendencies of essayistic expression. At the same time, students produced significantly more clauses per sentence and per T-unit, while maximum tree depth remained comparable or shallower, suggesting a preference for linear rather than deeply embedded syntactic structures.

These findings highlight the developmental nature of student writing: increased clause density signals growing syntactic fluency, yet deeper hierarchical structuring lags behind. The comparison with textbook models also underscores potential mismatches between pedagogical input and student output, calling for closer alignment between instructional materials and authentic student language.

Ultimately, this study demonstrates the value of corpus-based syntactic analysis for identifying developmental patterns and informing pedagogical practice.

6.2 Future Work

This analysis has already provided valuable insights into student writing, but there remain several directions to deepen and extend our understanding:

Corpus cleaning and improved annotation: The Učbeniki 1.0 corpus currently contains missegmented sentences and non-alphanumeric artifacts from PDF conversion. A thorough cleaning—removing stray characters and correcting sentence boundaries—would enable more accurate automatic UD annotation and allow us to reintroduce token-based syntactic complexity measures (MDD, NDD). These measures could be applied to enrich the overall picture of syntactic development.

Fine-grained conjunction patterns: A detailed examination of conjunction usage—specifically the distribution and frequency of individual conjunctions within each subtype of coordination and subordination—could uncover broader usage patterns and register effects in student writing.

Exploratory and machine-learning analyses: Beyond hypothesis-driven tests, an exploratory, data-driven approach—using clustering or classification techniques—could uncover hidden patterns of clause use, parse-tree configurations, or connective preferences.

Longitudinal developmental studies: Tracking students over time, from primary through secondary and into higher education, would illuminate how syntactic and lexical competencies evolve. Such longitudinal data could reveal critical periods for particular structures or register shifts as students encounter more advanced writing tasks.

Pursuing these avenues will not only validate and refine our current findings but also chart a richer map of syntactic development in educational contexts.

Limitations

Our findings should be interpreted with certain methodological limitations in mind. First, the Učbeniki 1.0 corpus suffers from noise introduced during PDF conversion: many sentences are missegmented or contain stray characters. This inadvertently lead to segmentation errors and parsing mistakes.

Second, all analyses rely on automatic Universal Dependencies annotation. Although UD provides a consistent framework, off-the-shelf parsers can mislabel complex or ungrammatical constructions—especially in student data, where nonstandard usage may further confuse the parser and introduce annotation errors.

Third, there is a text-type mismatch between corpora: Šolar is dominated by essays, whereas the textbook corpus is largely expository and descriptive. We have noted and taken genre effects into account in our interpretation, but residual differences in discourse conventions may still influence our quantitative measures.

Finally, our syntactic analysis does not include the UD *xcomp* (open clausal complement) relation, which in Slovene often corresponds to object clause. Because *xcomp* cannot be reliably distinguished from other non-clausal complements without manual inspection—and it lacks an overt conjunction—these instances are currently unaccounted for in our object-clause counts. Integrating *xcomp* into future analyses would provide a more complete picture of students' use of object clauses.

Acknowledgments

The paper was produced within the project *Empirical foundations for digitally-supported development of writing skills* (No. J7-3159) and the research programme Language Resources and Technologies for Slovene (No. P6-0411) funded by the Slovene Research and Innovation Agency (ARIS).

The authors would also like to thank the anonymous reviewers for their valuable feedback and constructive suggestions.

References

Špela Arhar Holdt and Iztok Kosem. 2024. Šolar, the developmental corpus of slovene. *Language Resources Evaluation*.

Špela Arhar Holdt and Tadeja Rozman. 2015. Možnosti uporabe podatkov iz korpusa Šolar za pripravo slovarskih priročnikov. In Mojca Smolej, editor, *Slovnica in slovar – aktualni jezikovni opis, del 1*, pages 67–74. Znanstvena založba Filozofske fakultete, Ljubljana.

Scott F. Beers and William E. Nagy. 2009. Syntactic complexity as a predictor of adolescent writing quality: Which measures? which genre? *Reading and Writing*, 22(2):185–200.

Scott F. Beers and William E. Nagy. 2011. Writing development in four genres from grades three to seven: syntactic complexity and genre differentiation. *Reading and Writing*, 24(2):183–202.

Kaja Dobrovoljc. 2024. Uporaba drevesnice sst v raziskavah govorjene slovenščine: prednosti in omejitve. *Jezik in slovstvo*, 69(4):187–209.

Polona Gantar and Mija Bon. forthcoming. Večbesedni leksikalni problemi pri samostojnem tvorjenju besedil v osnovnih in srednjih šolah. *Sodobna pedagogika*.

Kellogg W. Hunt. 1970. Syntactic maturity in schoolchildren and adults. *Monographs of the Society for Research in Child Development*, 35(1):iii–67.

Iztok Kosem and Eva Pori. forthcoming. Prvi koraki do seznama temeljnega besedišča z analizo korpusa slovenskih učbenikov. *Sodobna pedagogika*.

Luka Krsnik and Kaja Dobrovoljc. 2025. Stark: A toolkit for dependency (sub)tree extraction and analysis. In *Proceedings of SyntaxFest 2025*.

Nikola Ljubešić, Luka Terčon, and Kaja Dobrovoljc. 2024. Classla-stanza: The next step for linguistic processing of south slavic languages. In *Proceedings of the Conference on Language Technologies and Digital Humanities (JT-DH 2024)*, Ljubljana, Slovenia.

Tina Munda, Špela Arhar Holdt, Kaja Dobrovoljc, Iztok Kosem, Eva Pori, and Simon Krek. 2025a. Frequency lists of syntactic structures from the učbeniki 1.0 corpus. Slovenian language resource repository CLARIN.SI.

Tina Munda, Špela Arhar Holdt, Kaja Dobrovoljc, Tadeja Rozman, Mojca Stritar Kučuk, Simon Krek, Irena Krapš Vodopivec, Marko Stabej, Eva Pori, Teja Goli, Polona Lavrič, Cyprian Laskowski, Polonca Kocjančič, Bojan Klemenc, Luka Krsnik, and Iztok Kosem. 2025b. Frequency lists of syntactic structures from the Šolar 3.0 corpus. Slovenian language resource repository CLARIN.SI.

Tadeja Rozman, Špela Arhar Holdt, Senja Pollak, and Iztok Kosem. 2018. Kolokacije v korpusu Šolar. *Jezik in slovstvo*, 63(2/3):117–128.

Luka Terčon. 2024. Uporaba šestih mer skladenjske kompleksnosti za primerjavo jezika v govornem in pisnem korpusu. In *Proceedings of the Conference on Language Technologies and Digital Humanities (JT-DH 2024)*, Ljubljana, Slovenia.