Comparing Human and AI Literary Summaries: Insights from Lapitch

Petra Bago

Faculty of Humanities and Social Sciences, University of Zagreb

https://doi.org/10.17234/9789533792910.20

Abstract: Literary summaries, both in their formal and informal uses within educational contexts, play a notable role in students' learning, yet little is known about how AI-generated summaries compare to the human-authored versions students use daily. From an Information & Communication Sciences (ICS) and Digital Humanities (DH) perspective, this paper addresses that gap through a quantitative analysis. We compare three human-written summaries of the Croatian children's novel *Čudnovate zgode šegrta Hlapića* with three ChatGPT summaries generated via distinct prompting strategies: Single-Pass (SP), Incremental Updating (IU), and Hierarchical Merging (HM). Our analysis of structural, lexical, and content features reveals that AI-generated summaries form a distinct linguistic register. Despite high topical alignment (Cosine similarity), AI summaries exhibit greater lexical density and diversity, and significantly lower vocabulary overlap (Jaccard similarity) with their human counterparts. Prompting strategy proves to be a critical variable: SP produces verbose text, HM creates short, dense sentences, while IU most closely approximates human sentence length but remains stylistically different (e.g. more noun-heavy). We conclude that AI summaries are not direct replacements for human study guides but a parallel genre that may increase cognitive load. Consequently, the

prompting strategy is a key determinant of an AI summary's readability and utility for educational purposes.
Keywords: computational text analysis; computational stylistics; lexical diversity; readability; large language models

1. Introduction

When it comes to school required reading, everyday literary engagement by non-professional audiences (notably pupils and parents), often happens through *summaries*: short, purpose-built texts that scaffold recall, support homework, or partially substitute for the full work when time is scarce. From an Information & Communication Sciences (ICS) and Digital Humanities (DH) perspective, this practice is an ideal lens on how people encounter literature outside professional settings. Here, we use natural language processing (NLP) techniques to examine summary writing as an everyday reading activity and to compare human-authored summaries with those produced by a contemporary large language model (LLM) (Jurafsky and Martin 2023).

Recent work in NLP and DH has pushed long-form literary processing forward, including book-length summarisation and analysis of narrative structure and character information (Chang et al. 2024; Jang and Jung 2024; Kim et al. 2024; Yu, Liu, and Xiong 2024; Yuan et al. 2024; Zhao, Wang, and Wang 2024). These studies primarily benchmark models or propose datasets; fewer ask how AI-generated outputs compare to the human summaries that circulate in everyday educational use. Addressing that gap, we study Ivana Brlić-Mažuranić's *Čudnovate zgode šegrta Hlapića* (The Brave Adventures of

Lapitch), a canonical Croatian children's novel widely encountered via school-oriented websites. We contrast three human-authored summaries (found on websites lektira.hr[1], lektire.hr[2], sjedi5.com[3]) with three ChatGPT summaries generated under distinct prompting strategies. To control for temporal variation in model behaviour, all AI outputs were produced on 31 August 2025.

Our analysis targets measures that capture structure, vocabulary, and content overlap using well-established NLP/IR (information retrieval) tools. We report average sentence length (a proxy for syntactic packaging/readability) and average word length (a coarse indicator of lexical sophistication); lexical density as the proportion of content words, a common readability and information-load signal (McNamara et al. 2014); and lexical diversity via Type–Token Ratio (TTR) and its length-robust counterpart Mean Segmental TTR (MSTTR) and Root TTR (RTTR) (Johnson 1944; Jurafsky and Martin 2023; McNamara et al. 2014). To profile stylistic emphasis, we examine part-of-speech (POS) distributions in line with register-analytic practice (Biber 1995). For pairwise text similarity, we use Jaccard similarity over lemma sets to capture vocabulary overlap between summaries, and Cosine similarity computed on term frequency–inverse document frequency (TF-IDF) vectors to capture their topical alignment, both standard vector-space methods in information retrieval (Jurafsky and Martin 2023; Manning, Raghavan, and Schütze 2008; Salton and Buckley 1988).

[1] https://www.lektira.hr/
[2] https://www.lektire.hr/
[3] http://sjedi5.com (As of 7 September 2025, the site is down. However, some content of the site can be found on the Wayback Machine by Internet Archive, https://web.archive.org/.)

Because prompting conditions content selection in LLMs, we compare three summarisation strategies in ChatGPT: Single-Pass (SP) over the full text; Incremental Updating (IU), which iteratively revises a single evolving summary as new chunks arrive; and Hierarchical Merging (HM), which summarises parts first and then merges those summaries into a final whole. Operational details are specified in Section 3, where we implement both IU and HM with a four-chunk workflow tailored to the Croatian source text.

Our study addresses three research questions (RQs) central to understanding everyday reading in the age of generative AI:
RQ1. How do ChatGPT-generated literary summaries compare to human-authored summaries on structural and lexical measures?
RQ2. Do SP, IU, and HM strategies yield distinct stylistic and content profiles in the resulting summaries?
RQ3. What do these differences imply about how AI “reads” and packages narrative for lay audiences who rely on summaries in informal learning contexts?

The remainder of the paper is organised as follows. Section 2 situates the study within ICS, DH, and NLP, connecting prior work on long-form summarisation. Section 3 details the corpus, human sources, and three ChatGPT strategies (SP, IU, HM), implements a four-chunk workflow, and defines all measures. Section 4 reports results i.e. tables of structural/lexical metrics and similarity. Section 5 provides a discussion of the results, while Section 6 outlines limitations. Section 7 concludes and sketches future work. Appendices provide text of prompts for possible reproducibility of the experiment.

2. Related Work

This study approaches everyday literary reading through the prism of ICS and NLP, treating summaries as information artefacts that mediate access to narrative for non-professional audiences. Our perspective is explicitly quantitative: we operationalise stylistic and structural properties of summaries with reproducible NLP measures and compare three human-authored and tree AI-generated summaries of a single novel as a focused case study (Jurafsky and Martin 2023).

Methodologically, we rely on well-established text analytics. Lexical diversity is estimated with TTR and its length-robust variant MSTTR, and RTTR, while lexical density indexes information load via the proportion of content words (Johnson 1944; Jurafsky and Martin 2023; McNamara et al. 2014). Part-of-speech distributions profile register and stylistic emphasis (e.g. action orientation (verbs) versus description (adjectives)) in line with quantitative register analysis (Biber 1995). For content overlap, we use Jaccard similarity over lemma sets (set-based overlap) and Cosine similarity on TF–IDF vectors (vector-space proximity), following core IR practice (Jurafsky and Martin 2023; Manning, Raghavan, and Schütze 2008; Salton and Buckley 1988).

Within NLP, long-form literary processing has accelerated, with work on book-length summarisation, evaluation, and literary understanding. Research surveys and benchmarks examine chunking and control strategies, including hierarchical merging and incremental updating, and how these influence coherence and content selection in extended contexts (Chang et al. 2024). Evaluation

frameworks emphasise faithfulness and content selection for long texts (Kim et al. 2024), while new datasets target deep literary engagement via QA (Yu, Liu, and Xiong 2024). Beyond summary quality, studies probe whether models capture character attributes and relations (Jang and Jung 2024; Yuan et al. 2024) and comprehension and analysis of ancient poetry (Zhao, Wang, and Wang 2024).

Against this backdrop, we focus on a concrete, audience-specific setting: Croatian school-oriented summaries of *Čudnovate zgode šegrta Hlapića* (*The Brave Adventures of Lapitch*). While prior work chiefly compares models to models or evaluates against curated references (Chang et al. 2024; Kim et al. 2024; Yu, Liu, and Xiong 2024), far less is known about how an LLM's outputs diverge from or align with the human summaries that circulate in everyday educational use. Our contribution is to supply a controlled comparison between the three human-authored summaries from popular educational websites and the three ChatGPT summaries produced via Single-Pass, Incremental Updating, and Hierarchical Merging prompting (operationalised in Section 3), with all AI experiments dated 31 August 2025 for temporal control.

Prior work primarily benchmarks LLMs against curated references; there is little of it that addresses alignment with human-authored school summaries circulating online. We target that gap by comparing three Croatian school-site summaries (lektira.hr, lektire.hr, sjedi5.com) with three ChatGPT summaries (SP, IU, HM) under a matched Croatian NLP pipeline[4] and a shared metric suite.

[4] CLASSLA-Stanza pipeline (https://github.com/clarinsi/classla)

In sum, we integrate ICS concerns with information access in informal learning and NLP's quantitative toolset to study how summaries package narrative for lay readers. We intentionally foreground quantitative metrics and leave qualitative discourse/stylistic analysis (e.g. rhetorical moves, error typologies, reader reception) to future work.

3. Methodology

Our analysis focuses on three human-authored summaries (from lektira.hr, lektire.hr, and sjedi5.com) and three AI-generated summaries. The AI summaries were created in Croatian on 31 August 2025 using ChatGPT 5 Thinking model, with each employing a distinct strategy: Single-Pass (SP), Hierarchical Merging (HM), and Incremental Updating (IU).

3.1. Corpus and Materials

Source text. Our case study is Ivana Brlić-Mažuranić's *Čudnovate zgode šegrta Hlapića* (*The Brave Adventures of Lapitch*), a canonical text in the Croatian primary school curriculum. Using a consistent Croatian NLP pipeline for tokenisation, lemmatisation, POS-tagging, and sentence splitting, we calculated the baseline profile of the novel, which can be found in Table 1.

Human-authored summaries. The human-authored benchmark consists of three Croatian-language summaries collected from widely used educational websites (lektira.hr, lektire.hr, and sjedi5.com)

whose primary audience is primary school pupils and teachers. Source texts were copied as plain UTF-8 files. No translation was performed; all analysis remained in Croatian. To ensure like-for-like comparability with the AI outputs, these files were processed through the same Croatian NLP pipeline described above.

AI summaries (ChatGPT). We generated three summaries with ChatGPT (ChatGPT 5 Thinking model) using a consistent header that set the persona to an "expert Croatian children's literature teacher" and specified audience, tone, and output length (see Appendices for prompt content). Each strategy below produced one final summary. Access to the model required a paid ChatGPT Plus subscription; at the time of prompting (31 August 2025) the subscription cost was €23 per month.

3.2. Summarisation Strategies

Each prompt started with a standardised header that fixed: (i) language (Croatian), (ii) a persona of a patient, experienced children's-literature teacher for third-grade pupils, (iii) style (simple wording, short sentences, age-appropriate vocabulary), (iv) tone (warm, encouraging, non-technical), and (v) the expected role (produce pupil-oriented summaries and brief analyses). This header was followed by the task for the LLM, which supplied the strategy-specific instructions for Single-Pass, Hierarchical Merging, or Incremental Updating (see Appendix for full prompts; cf. strategy design in Chang et al. 2024; Kim et al. 2024).

3.2.1. Single-Pass Summarisation (SP)

We provided the entire novel to ChatGPT in one turn and requested a classroom-appropriate summary targeted to third-grade readers. This probes unassisted long-context summarisation in a single generative act. Full text prompt for SP is available in Appendix A.

3.2.2. Hierarchical Merging Summarisation

We divided the novel into four chunks by word count to preserve processing load for LLMs and fed the chunks in order of appearance to preserve narrative continuity: chunk A contains 6,207 words, chunk B 6,048 words, chunk C 6,269 words, and chunk D 6,346 words.
The procedure was as follows:
1. Chunk-level summaries: Prompt ChatGPT to summarise each chunk separately for the same audience (S_A, S_B, S_C, S_D).
2. Pair merges: Merge S_A and S_B into summary S_{AB} and S_C and S_D into S_{CD}, with instructions to resolve redundancies and maintain narrative flow (without access to the raw chunks; only S_A–S_D texts available during merging).
3. Final merge: Merge S_{AB} and S_{CD} into the final HM summary with word limit for the final summary.
Full text prompts for HM are available in Appendix B.

3.2.3. Incremental Updating Summarisation

We simulated sequential reading with rolling revision:
1. Initialise with chunk A → produce a first summary.

2. Update the same summary with chunk B (keep key earlier content unless superseded). Each update replaced the previous summary with a full rewritten synthesis.
3. Update with chunk C (keep key earlier content unless superseded). Each update replaced the previous summary with a full rewritten synthesis.
4. Update with chunk D to produce the final IU summary (single output that reflects all four chunks) with an additional instruction for the word limit. Each update replaced the previous summary with a full rewritten synthesis.

3.3. Computational Analysis

All human and ChatGPT summaries were processed with the same Croatian NLP pipeline to ensure comparable units (tokens, lemmas, sentences, POS).

Structural measures.

- Average sentence length (tokens/sentence) approximates syntactic packaging/readability.
- Average word length (characters/word) is a coarse proxy for lexical sophistication.

Lexical diversity and repetition.

- TTR = *types / tokens*; sensitive to length (Jurafsky and Martin 2023).
- MSTTR averages TTR over fixed-length segments (100 tokens) to stabilise length effects (Johnson 1944).

- Root TTR (RTTR) = *types / √(tokens)* (Jurafsky and Martin 2023) that roughly cancels length effect.
- Hapax and dis legomena (types occurring once/twice) index vocabulary dispersion.

Lexical density and content focus.

- Lexical density = *content words* (NOUN, PROPN, VERB, ADJ, ADV) / *all tokens excluding punctuation*; higher values indicate greater information load (McNamara et al. 2014).
- Verb-to-word ratio profiles action orientation; for this measure we count main verbs without auxiliaries (Biber 1995).

Part-of-speech profile.

- POS distributions (e.g. NOUN/VERB/ADJ/ADV/PROPN) serve as quantitative register/style indicators (Biber 1995), highlighting whether summaries are more action-driven (verbs) or description-heavy (adjectives/nouns).

Text similarity.

- Jaccard over lemma sets captures unique-vocabulary overlap.
- Cosine similarity over TF–IDF vectors captures topical proximity in a vector space (Salton and Buckley 1988; Manning, Raghavan, and Schütze 2008; Jurafsky and Martin 2023).
    - TF–IDF weights emphasise terms distinctive to a document within the comparison set (Salton and Buckley 1988).

Baseline reference. We use the novel's profile (values listed in Table 1) as a contextual anchor when interpreting summary metrics (e.g. expected compression of average sentence length, shifts in POS ratios). Because all measures are pipeline-dependent, comparisons are within-pipeline and like-for-like across human and AI summaries.

4. Results

This section reports quantitative findings for six summaries: three human-authored (lektira.hr, lektire.hr, sjedi5.com) and three ChatGPT summaries generated via Single-Pass (SP), Incremental Updating (IU), and Hierarchical Merging (HM), as well as metrics for the book the summaries were based on. All AI outputs were produced on 31 August 2025. IU and HM followed the four-chunk workflow (6,207 / 6,048 / 6,269 / 6,346 words). We interpret results from an Information & Communication Sciences and Digital Humanities perspective, focusing on measures explained earlier: structural packaging (average sentence/word length), lexical diversity (TTR, MSTTR, RTTR, hapax/dis), information load (lexical density), action orientation (verb-to-word ratio), and content overlap (Jaccard over lemmas; Cosine over TF-IDF).

Table 1. Structural and Lexical Metrics for the Source Novel (*Lapitch*), Three Human-Authored Summaries, and Three ChatGPT-Generated Summaries.

Metric	Sjedi5.com	Lektira.hr	Lektire.hr	ChatGPT SP	ChatGPT IU	ChatGPT HM	Lapitch
Tokens	2126	2080	1741	2972	2557	2289	29886
Types	619	609	533	817	851	695	2221
Average Word Length (chars)	4.19	4.15	4.28	3.95	**4.34**	4.10	*3.67*
Average Sentence Length (tokens)	16.76	15.18	**17.95**	12.08	17.51	9.30	*14.35*
Hapax Legomena	372	372	326	464	**528**	399	*900*
Dis Legomena	104	97	87	138	**146**	116	*347*
Verb-to-Word Ratio	0.140	0.147	**0.153**	0.133	0.102	0.142	*0.13*
Type-Token Ratio (TTR)	0.286	0.293	0.306	0.275	**0.333**	0.304	*0.074*
Mean Segmental TTR (MSTTR)	0.649	0.640	0.657	0.679	**0.745**	0.699	*0.59*
Root TTR (RTTR)	13.425	13.353	12.774	14.99	**16.829**	14.527	*12.847*
Lexical Density	0.522	0.503	0.515	0.519	**0.576**	0.568	*0.461*

Key patterns. Human summaries are shorter overall; ChatGPT-SP is the most verbose. Sentence packaging diverges by strategy: IU produces long, human-like sentences (17.51 tokens; comparable to human 15–18 and close to the novel's 14.35 baseline), whereas HM

yields very short sentences (9.30), consistent with information condensed through multi-stage merging. In vocabulary, IU shows the highest diversity (TTR 0.333; MSTTR 0.745) and the highest counts of hapax/dis legomena, consistent with more varied phrasing; we cannot attribute cause without further analysis. Lexical density is highest for IU (0.576) and HM (0.568), indicating a greater proportion of content words and thus a denser information load than the human summaries. The verb-to-word ratio is lowest for IU (0.102), implying a comparatively noun/adjective-heavy style, while HM (0.142) sits closer to human ratios (~0.14–0.15), despite its shorter sentences.

Table 2. Pairwise Text Similarity Scores (Jaccard and Cosine) Across Human and AI Summaries.

Summary Pair	**Jaccard Similarity**	**Cosine Similarity**
Sjedi5 + Lektire.hr	0.408	0.925
Sjedi5 + Lektira.hr	0.410	0.923
Lektire.hr + Lektira.hr	**0.606**	**0.982**
Sjedi5 + ChatGPT SP	0.280	0.845
Sjedi5 + ChatGPT IU	0.256	0.746
Sjedi5 + ChatGPT HM	0.327	0.796
Lektire.hr + ChatGPT SP	0.244	0.796
Lektire.hr + ChatGPT IU	0.214	0.666
Lektire.hr + ChatGPT HM	0.293	0.693
Lektira.hr + ChatGPT SP	0.273	0.801
Lektira.hr + ChatGPT IU	0.224	0.657
Lektira.hr + ChatGPT HM	0.309	0.684
ChatGPT SP + ChatGPT IU	0.257	0.743
ChatGPT SP + ChatGPT HM	0.327	0.818
ChatGPT IU + ChatGPT HM	0.351	0.834

Interpretation. Human-to-human pairs show high overlap: Cosine values ≥0.92 (shared topical space under TF-IDF weighting) and Jaccard ≥0.41 (substantial unique-lemma overlap). Human-to-ChatGPT pairs are consistently lower on both metrics (all Jaccard <0.33; Cosine mostly 0.66–0.85), indicating that while themes align (Cosine), the lexicons diverge (Jaccard). Among ChatGPT strategies, HM and IU are most similar to each other (Jaccard 0.351; Cosine 0.834), suggesting that multi-pass workflows converge towards a common content selection even as their surface styles differ (short, dense HM vs. longer, diverse IU). From an ICS/DH perspective, these contrasts imply that AI summaries constitute a distinct register relative to the human study-guide style used by pupils and teachers, with strategy choice exerting measurable influence on readability and information packaging.

5. Discussion

From an ICS/DH vantage, our results indicate a consistent separation between human-authored and ChatGPT summaries across multiple quantitative lenses. Human–human pairs exhibit high topical proximity (Cosine on TF–IDF) and substantial lexical overlap (Jaccard over lemmas). In contrast, human–ChatGPT pairs are consistently lower on both measures. This suggests the model's outputs constitute a distinct register rather than a mimicry of school-oriented study guides (Manning, Raghavan, and Schütze 2008; Salton and Buckley 1988).

Human vs. ChatGPT: where the differences lie. Two families of measures help explain the divergence. First, overlap metrics reveal key differences. A lower Jaccard similarity in human-ChatGPT pairs indicates the model selects a different vocabulary, even when Cosine similarity suggests broadly similar topics. This reflects alternative content packaging and weighting under TF-IDF (Manning, Raghavan, and Schütze 2008). Second, the texts differ in form and density. ChatGPT outputs, especially IU and HM, show higher lexical density and diversity (TTR/MSTTR). In contrast, the human texts tend to use more formulaic, pedagogy-friendly phrasing. This aligns with evidence that instruction-tuned LLMs use more nominalisations and adopt a more informationally dense, noun-heavy style than human writing (Reinhart et al. 2025).

Anchoring these contrasts to the source novel reveals additional patterning. All summaries compress *Lapitch* to 6–10 % of its length yet increase lexical density from 0.461 to 0.503–0.576 and average word length from 3.67 to 3.95–4.34. Human sentences are longer than the novel (≈15–18 vs. 14.35 tokens), SP shorter (12.08), and HM shortest (9.3). POS profiles also show a shift: the summaries feature noun inflation and a contraction of verbs, pronouns, and adverbs. For instance, the IU summary has the lowest verb-to-word ratio (0.102 vs. the novel's 0.13), indicating more nominalised packaging. ChatGPT's summaries also show a sharp drop in auxiliaries and subordinators. This, combined with higher punctuation rates, yields a flatter clause structure. See Appendix D for full POS frequency distributions.

What prompting strategy changes. Strategy choice leaves a measurable stylistic fingerprint:

- SP is most verbose and least human-like in aggregate similarity, consistent with unconstrained one-shot compression of a long narrative.
- HM yields short, choppy sentences and high density, a hallmark of multi-stage condensation where local compressions accumulate during merges. The result is compact but less narratively fluid.
- IU produces human-like sentence lengths and the highest diversity (TTR/MSTTR). However, it also has a lower verb-to-word ratio, indicating a tilt towards nominal description over event progression. This stylistic difference has potential consequences for readability and plot salience.

Implications for everyday reading. For everyday school use, human summaries appear optimised for accessibility. They feature moderate sentence lengths, stable lexical choices, and a higher action salience (verb ratio) that foregrounds who-does-what. While information-rich, ChatGPT summaries may impose a higher cognitive load due to their greater lexical density and diversity. Depending on the strategy, they can also either over-condense the narrative (HM) or over-elaborate on it (SP). The IU strategy comes closest to human packaging but still diverges lexically and in rhetorical balance. From an ICS standpoint, these findings position AI summaries as a parallel informational genre that coexists with, rather than replaces, human study guides.

Methodological takeaways. The metric bundle used here (sentence/word length, lexical density, TTR/MSTTR/RTTR, verb-to-word ratio, and Jaccard/Cosine) offers a replicable scaffold for comparing human and model outputs at scale (Jurafsky and Martin 2023; McNamara et al. 2014).

Bottom line: ChatGPT can summarise effectively, but not in the same register as human classroom guides. The choice of strategy (SP vs. IU vs. HM) systematically shifts readability and information packaging. This underlines that how we prompt an AI is just as consequential as what we ask it to summarise.

6. Limitations

Our findings are bounded by several constraints relevant to Information & Communication Sciences and Digital Humanities:

- **Corpus Scope**: The analysis is based on a single Croatian novel and its school-oriented summaries, which limits the generalisability of our findings across different genres, age groups, and languages.
- **Model Scope**: The paper compares only three ChatGPT summaries to three human summaries. The results may not transfer to other models.
- **Temporal Specificity**: All AI outputs were generated on a single day (31 August 2025). While this mitigates some variability, it does not eliminate the potential effects of model-update drift over time.
- **Chunking Design**: The specific four-chunk design used for the Incremental Updating (IU) and Hierarchical Merging (HM) strategies may have shaped the outcomes. Different chunking methods could yield different stylistic results.
- **Methodological Dependence**: The results depend on a single Croatian NLP pipeline and a specific set of quantitative metrics (e.g. TTR/MSTTR, lexical density, Jaccard/Cosine).

- **Content Analysis**: No analysis of faithfulness or plot accuracy was performed on the AI-generated summaries.
- **Non-Deterministic Outputs**: We did not control for stochasticity in decoding; repeating prompts may yield variance in wording and metrics.

7. Conclusion and Future Work

From an Information & Communication Sciences and Digital Humanities perspective, we compared three human summaries with three ChatGPT summaries generated using Single-Pass, Incremental Updating, and Hierarchical Merging strategies. The results show that the AI outputs, generated on 31 August 2025, form a distinct register rather than simply reproducing the style of school-oriented study guides. Quantitatively, ChatGPT differs in lexical diversity (TTR/MSTTR), lexical density, sentence packaging, and vocabulary overlap (Jaccard/Cosine). IU comes closest to human stylistics but still diverges; SP tends towards verbosity; HM compresses aggressively into short, dense sentences. For IU/HM, the four-chunk workflow likely shaped stylistic outcomes.

Answers to RQs. RQ1: Compared with the three human summaries, ChatGPT outputs show higher lexical density (0.568–0.576 vs. 0.503–0.522), greater lexical diversity (MSTTR 0.699–0.745 vs. 0.640–0.657), strategy-dependent sentence packaging, and markedly lower lemma overlap with humans (all human–ChatGPT Jaccard < 0.33; cosine mostly 0.66–0.85). RQ2: Strategy matters: SP is most verbose and least aligned with human summaries, HM yields the shortest, densest sentences, and IU is closest on sentence length and diversity yet

remains noun-heavy (lowest verb-to-word ratio = 0.102). RQ3: For lay readers who rely on summaries, these differences indicate that AI outputs constitute a parallel register that may raise cognitive load; IU appears the most promising starting point but still benefits from editorial adaptation for classroom use.

Practical implication: For teacher-facing materials, IU-based summaries are the closest stylistic match to human guides but remain lexically denser; add light scaffolding (shorter sentences, more verb-forward phrasing, brief vocabulary glosses) before distributing to pupils.

Looking ahead, we will: (1) broaden the corpus (genres, grade levels, languages) and compare additional models; (2) vary chunking granularity and prompting (few-shot, Chain-of-Thought) to test robustness; (3) extend metrics to discourse coherence, faithfulness, and coverage; (4) incorporate human evaluation with teachers and pupils (readability/comprehension); and (5) conduct a planned qualitative analysis (rhetorical moves, error typologies, reception). Together, these steps will clarify when and how AI summaries can complement everyday reading.

8. Acknowledgments

This work was supported in part by the Croatian Science Foundation under the project number HRZZ-IP-2022-10-7697 (MWE-Cro: Multiword Expressions in Croatian – Lexicological, Computational Linguistic and Glottodidactic Approach), Faculty of Humanities and Social Sciences, and Petra Bago.

9. References

Biber, Douglas. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Chang, Yapei, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. “BooookScore: A Systematic Exploration of Book-Length Summarization in the Era of LLMs.” arXiv preprint. doi:10.48550/arXiv.2310.00785.

Jang, Woori, and Seohyon Jung. 2024. “Evaluating LLM Performance in Character Analysis: A Study of Artificial Beings in Recent Korean Science Fiction.” In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, edited by Mika Hämäläinen, Emily Öhman, So Miyagawa, Khalid Alnajjar, and Yuri Bizzoni. 339–51. Miami, USA: Association for Computational Linguistics. doi:10.18653/v1/2024.nlp4dh-1.34.

Johnson, Wendell. 1944. “Studies in language behavior: A program of research.” *Psychological Monographs* 56 (2): 1-15.

Jurafsky, Daniel, and James H. Martin. 2023. *Speech and Language Processing (3rd ed., draft)*. Accessed September 7, 2025. https://web.stanford.edu/~jurafsky/slp3/.

Kim, Yekyung, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. “FABLES: Evaluating Faithfulness and Content Selection in Book-Length Summarization.” arXiv preprint. doi:10.48550/arXiv.2404.01261.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

McNamara, Danielle S., Arthur C. Graesser, Phillip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge: Cambridge University Press.

Reinhart, Alex, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. "Do LLMs Write like Humans? Variation in Grammatical and Rhetorical Styles." *Proceedings of the National Academy of Sciences* 122 (8): e2422455122. https://doi.org/10.1073/pnas.2422455122.

Salton, Gerard, and Christopher Buckley. 1988. "Term-weighting approaches in automatic text retrieval." *Information Processing & Management* 24 (5), 513–23.

Yu, Linhao, Qun Liu, and Deyi Xiong. 2024. "LFED: A Literary Fiction Evaluation Dataset for Large Language Models." In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, edited by Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. ELRA and ICCL. https://aclanthology.org/2024.lrec-main.915/.

Yuan, Xinfeng, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. "Evaluating Character Understanding of Large Language Models via Character Profiling from Fictional Works." 2024. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, edited by Yaser Al-Onaizan, Mohit Bansal, and Yun-

Nung Chen. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.456.
Zhao, Cheng, Bin Wang, and Zhen Wang. 2024. “Understanding Literary Texts by LLMs: A Case Study of Ancient Chinese Poetry.” arXiv preprint. doi:10.48550/arXiv.2409.00060.

10. Appendices

10.1. Appendix A: ChatGPT Single-Pass (SP) Summarisation Prompt

Croatian	English (translation)
Jezik: hrvatski	Language: Croatian
Osobnost: Ti si strpljiv i iskusan učitelj književnosti, specijaliziran za dječju lektiru za učenike trećeg razreda osnovne škole. Tvoja je glavna zadaća pomoći djeci u razumijevanju i analizi obvezne lektire.	Personality: You are a patient and experienced literature teacher, specialized in children’s assigned reading for third-grade elementary school students. Your main task is to help children understand and analyze the compulsory reading.
Stil: Koristi jednostavan i jasan jezik, s kratkim rečenicama i rječnikom primjerenim za djecu od devet godina.	Style: Use simple and clear language, with short sentences and a vocabulary suitable for nine-year-old children.
Ton: Tvoj ton mora biti topao, poticajan i ohrabrujuć, kao da razgovaraš s djetetom koje uči. Nikad ne koristi složene ili stručne	Tone: Your tone must be warm, encouraging, and supportive, as if you are talking to a child who is learning. Never use complex or

izraze. Uvijek budi podrška.	technical terms. Always be supportive.
Očekivana uloga: Od tebe se očekuje da pišeš sažetke i analize lektira, prilagođene djeci trećih razreda osnovne škole.	Expected role: You are expected to write summaries and analyses of reading assignments, tailored to third-grade elementary school children.
Zadatak: Na temelju priloženog teksta, generiraj opsežan sažetak i analizu djela koji će pomoći učeniku da ispuni svoj dnevnik čitanja.	Task: Based on the attached text, generate a comprehensive summary and analysis of the work that will help the student complete their reading journal.
Upute za sadržaj: Tekst treba biti dugačak između 1500 i 1900 riječi. Tekst treba sadržavati sljedeće dijelove, jasno označene naslovima: 1. Kratak sadržaj: Napiši jasan i detaljan sažetak radnje. 2. Analiza likova: Odaberi i analiziraj glavne i sporedne likove koji su ključni za radnju, objašnjavajući njihove uloge i osobine. 3. Podaci o djelu: Na kraju teksta, u posebnom odjeljku, navedi sljedeće informacije o djelu: * Književni rod	Content guidelines: The text should be between 1,500 and 1,900 words. The text should contain the following parts, clearly labeled with headings: 1. Short summary: Write a clear and detailed summary of the plot. 2. Character analysis: Choose and analyze the main and supporting characters who are key to the plot, explaining their roles and traits. 3. Information about the work: At the end of the text, in a separate section, list the following information about the work: * Literary category

* Književna vrsta * Mjesto radnje * Vrijeme radnje * Glavni lik * Sporedni likovi * Pouka djela	* Literary genre * Setting (place) * Time of the story * Main character * Supporting characters * Moral/lesson of the work
Generiraj kompletan tekst u jednom odgovoru. Ne postavljaj dodatna pitanja.	Generate the complete text in a single response. Do not ask additional questions.
Slijedi tekst književnog djela: [CIJELI IZVORNI TEKST IDE OVDJE]	The following is the text of the literary work: [FULL SOURCE TEXT GOES HERE]

10.2. Appendix B: ChatGPT Hierarchical Merging (HM) Summarisation Prompts

10.2.1. ChatGPT Hierarchical Merging (HM) — Chunk Summarisation Prompts

Croatian	English (translation)
[Identično zaglavlje kao i kod SP: Jezik, Osobnost, Stil, Ton, Očekivana uloga]	[Identical header as for SP: Language, Personality, Style, Tone, Expected role]
Zadatak: Na temelju priloženog teksta koji je isječak iz književnog djela, generiraj opsežan sažetak i analizu isječka iz djela koji će	Task: Based on the attached text, which is an excerpt from a literary work, generate a comprehensive summary and analysis of the

pomoći učeniku da ispuni svoj dnevnik čitanja.	excerpt that will help the student complete their reading journal.
Upute za sadržaj: Tekst treba sadržavati sljedeće dijelove, jasno označene naslovima: 1. Kratak sadržaj: Napiši jasan i detaljan sažetak radnje. 2. Analiza likova: Odaberi i analiziraj glavne i sporedne likove koji su ključni za radnju, objašnjavajući njihove uloge i osobine. 3. Podaci o djelu: Na kraju teksta, u posebnom odjeljku, navedi sljedeće informacije o djelu: * Književni rod * Književna vrsta * Mjesto radnje * Vrijeme radnje * Glavni lik * Sporedni likovi * Pouka djela	Content guidelines: The text should contain the following parts, clearly labeled with headings: 1. Short summary: Write a clear and detailed summary of the plot. 2. Character analysis: Choose and analyze the main and supporting characters who are key to the plot, explaining their roles and traits. 3. Information about the work: At the end of the text, in a separate section, list the following information about the work: * Literary category * Literary genre * Setting (place) * Time of the story * Main character * Supporting characters * Moral/lesson of the work
Generiraj kompletan tekst u jednom odgovoru. Ne postavljaj dodatna pitanja.	Generate the complete text in a single response. Do not ask additional questions.
Slijedi tekst isječka iz književnog djela: [DIO IZVORNOG TEKSTA IDE	The following is the text of the excerpt from the literary work: [CHUNK OF SOURCE TEXT GOES

OVDJE]	HERE]

10.2.2. ChatGPT Hierarchical Merging (HM) — Pairwise Merge Prompts

Croatian	English (translation)
[Identično zaglavlje kao i kod SP: Jezik, Osobnost, Stil, Ton, Očekivana uloga]	[Identical header as for SP: Language, Personality, Style, Tone, Expected role]
Zadatak: Na temelju priloženih sažetaka konzekutivnih isječaka iz književnog djela, generiraj sjedinjeni opsežan sažetak i analizu isječaka iz djela koji će pomoći učeniku da ispuni svoj dnevnik čitanja.	Task: Based on the attached summaries of consecutive excerpts from a literary work, generate a unified comprehensive summary and analysis of the excerpts that will help the student complete their reading journal.
Upute za sadržaj: Tekst treba sadržavati sljedeće dijelove, jasno označene naslovima: 1. Kratak sadržaj: Napiši jasan i detaljan sažetak radnje. 2. Analiza likova: Odaberi i analiziraj glavne i sporedne likove koji su ključni za radnju, objašnjavajući njihove uloge i osobine. 3. Podaci o djelu: Na kraju teksta,	Content guidelines: The text should contain the following parts, clearly labeled with headings: 1. Short summary: Write a clear and detailed summary of the plot. 2. Character analysis: Choose and analyze the main and supporting characters who are key to the plot, explaining their roles and traits. 3. Information about the work: At the end of the text, in a separate

u posebnom odjeljku, navedi sljedeće informacije o djelu: * Književni rod * Književna vrsta * Mjesto radnje * Vrijeme radnje * Glavni lik * Sporedni likovi * Pouka djela	section, list the following information about the work: * Literary category * Literary genre * Setting (place) * Time of the story * Main character * Supporting characters * Moral/lesson of the work
Generiraj kompletan tekst u jednom odgovoru. Ne postavljaj dodatna pitanja.	Generate the complete text in a single response. Do not ask additional questions.
Slijedi tekst jednog sažetka isječka iz književnog djela: [SAŽETAK DIJELA IZVORNOG TEKSTA IDE OVDJE]	The following is the text of one summary of an excerpt from the literary work: [SUMMARY OF CHUNK OF SOURCE TEXT GOES HERE]
Slijedi tekst drugog sažetka isječka iz književnog djela: [SAŽETAK DIJELA IZVORNOG TEKSTA IDE OVDJE]	The following is the text of the second summary of an excerpt from the literary work: [SUMMARY OF CHUNK OF SOURCE TEXT GOES HERE]

10.2.3. ChatGPT Hierarchical Merging (HM) — Final Merge Prompt

Croatian	English (translation)
[Identično zaglavlje kao i kod SP: Jezik, Osobnost, Stil, Ton,	[Identical header as for SP: Language, Personality, Style,

Očekivana uloga]	Tone, Expected role]
Zadatak: Na temelju priloženih sjedinjenih sažetaka konzekutivnih isječaka iz književnog djela, generiraj jedinstven opsežan sažetak i analizu cjelokupnog djela koji će pomoći učeniku da ispuni svoj dnevnik čitanja.	Task: Based on the attached unified summaries of consecutive excerpts from a literary work, generate a single comprehensive summary and analysis of the entire work that will help the student complete their reading journal.
Upute za sadržaj: Tekst treba biti dugačak između 1500 i 1900 riječi. Tekst treba sadržavati sljedeće dijelove, jasno označene naslovima: 1. Kratak sadržaj: Napiši jasan i detaljan sažetak radnje. 2. Analiza likova: Odaberi i analiziraj glavne i sporedne likove koji su ključni za radnju, objašnjavajući njihove uloge i osobine. 3. Podaci o djelu: Na kraju teksta, u posebnom odjeljku, navedi sljedeće informacije o djelu: * Književni rod * Književna vrsta * Mjesto radnje * Vrijeme radnje	Content guidelines: The text should be between 1,500 and 1,900 words. The text should contain the following parts, clearly labeled with headings: 1. Short summary: Write a clear and detailed summary of the plot. 2. Character analysis: Choose and analyze the main and supporting characters who are key to the plot, explaining their roles and traits. 3. Information about the work: At the end of the text, in a separate section, list the following information about the work: * Literary category * Literary genre * Setting (place) * Time of the story

* Glavni lik * Sporedni likovi * Pouka djela	* Main character * Supporting characters * Moral/lesson of the work
Generiraj kompletan tekst u jednom odgovoru. Ne postavljaj dodatna pitanja.	Generate the complete text in a single response. Do not ask additional questions.
Slijedi tekst prvog sjedinjenog sažetka isječaka iz književnog djela: [SAŽETAK SAŽETAKA IDE OVDJE]	The following is the text of the first unified summary of excerpts from the literary work: [SUMMARY OF SUMMARIES GOES HERE]
Slijedi tekst drugog sjedinjenog sažetka isječaka iz književnog djela: [SAŽETAK SAŽETAKA IDE OVDJE]	The following is the text of the second unified summary of excerpts from the literary work: [SUMMARY OF SUMMARIES GOES HERE]

10.3. Appendix C: ChatGPT Incremental Updating (IU) Summarisation Prompts

10.3.1. ChatGPT Incremental Updating (IU) — Initial Prompt

Croatian	English (translation)
[Identično zaglavlje kao i kod SP: Jezik, Osobnost, Stil, Ton,	[Identical header as for SP: Language, Personality, Style,

Očekivana uloga]	Tone, Expected role]
Zadatak: Na temelju priloženog isječka iz književnog djela, generiraj opsežan sažetak i analizu isječka iz djela koji će pomoći učeniku da ispuni svoj dnevnik čitanja.	Task: Based on the attached text, which is an excerpt from a literary work, generate a comprehensive summary and analysis of the excerpt that will help the student complete their reading journal.
Upute za sadržaj: Tekst treba sadržavati sljedeće dijelove, jasno označene naslovima: 1. Kratak sadržaj: Napiši jasan i detaljan sažetak radnje. 2. Analiza likova: Odaberi i analiziraj glavne i sporedne likove koji su ključni za radnju, objašnjavajući njihove uloge i osobine. 3. Podaci o djelu: Na kraju teksta, u posebnom odjeljku, navedi sljedeće informacije o djelu: * Književni rod * Književna vrsta * Mjesto radnje * Vrijeme radnje * Glavni lik * Sporedni likovi * Pouka djela	Content guidelines: The text should contain the following parts, clearly labeled with headings: 1. Short summary: Write a clear and detailed summary of the plot. 2. Character analysis: Choose and analyze the main and supporting characters who are key to the plot, explaining their roles and traits. 3. Information about the work: At the end of the text, in a separate section, list the following information about the work: * Literary category * Literary genre * Setting (place) * Time of the story * Main character * Supporting characters * Moral/lesson of the work
Generiraj kompletan tekst u	Generate the complete text in a

jednom odgovoru. Ne postavljaj dodatna pitanja.	single response. Do not ask additional questions.
Slijedi tekst isječka iz književnog djela: [PRVI DIO IZVORNOG TEKSTA IDE OVDJE]	The following is the text of the excerpt from the literary work: [FIRST CHUNK OF SOURCE TEXT GOES HERE]

10.3.2. ChatGPT Incremental Updating (IU) — Iterative Summary-Update Prompt

Croatian	English (translation)
Zadatak: Ažuriraj postojeći sažetak književnog djela na temelju nastavka dijela priče. Sačuvaj važne ranije događaje, ali prepravi ili proširi sažetak kako bi odražavao cijeli dosadašnji tijek radnje. Generiraj sjedinjeni opsežan sažetak i analizu isječaka iz djela koji će pomoći učeniku da ispuni svoj dnevnik čitanja.	Task: Update the existing summary of the literary work based on the continuation of the story. Preserve the important earlier events, but revise or expand the summary so that it reflects the entire plot so far. Generate a unified comprehensive summary and analysis of the excerpts that will help the student complete their reading journal.
Upute za sadržaj: Tekst treba sadržavati sljedeće dijelove, jasno označene naslovima:	Content guidelines: The text should contain the following parts, clearly labeled with headings:

1. Kratak sadržaj: Napiši jasan i detaljan sažetak radnje. 2. Analiza likova: Odaberi i analiziraj glavne i sporedne likove koji su ključni za radnju, objašnjavajući njihove uloge i osobine. 3. Podaci o djelu: Na kraju teksta, u posebnom odjeljku, navedi sljedeće informacije o djelu: * Književni rod * Književna vrsta * Mjesto radnje * Vrijeme radnje * Glavni lik * Sporedni likovi * Pouka djela	1. Short summary: Write a clear and detailed summary of the plot. 2. Character analysis: Choose and analyze the main and supporting characters who are key to the plot, explaining their roles and traits. 3. Information about the work: At the end of the text, in a separate section, list the following information about the work: * Literary category * Literary genre * Setting (place) * Time of the story * Main character * Supporting characters * Moral/lesson of the work
Generiraj kompletan tekst u jednom odgovoru. Ne postavljaj dodatna pitanja.	Generate the complete text in a single response. Do not ask additional questions.
Slijedi tekst postojećeg sažetka: [SAŽETAK PRETHODNOG DIJELA IZVORNOG TEKSTA IDE OVDJE]	The following is the text of the existing summary: [SUMMARY OF PREVIOUS CHUNK OF SOURCE TEXT GOES HERE]
Slijedi tekst isječka iz književnog djela: [SLJEDEĆI DIO IZVORNOG TEKSTA IDE OVDJE]	The following is the text of the excerpt from the literary work: [NEXT CHUNK OF SOURCE TEXT GOES HERE]

10.3.3. ChatGPT Incremental Updating (IU) — Iterative Summary-Update Prompt

Croatian	English (translation)
Zadatak: Ažuriraj postojeći sažetak književnog djela na temelju nastavka dijela priče. Sačuvaj važne ranije događaje, ali prepravi ili proširi sažetak kako bi odražavao cijeli dosadašnji tijek radnje. Generiraj sjedinjeni opsežan sažetak i analizu isječaka iz djela koji će pomoći učeniku da ispuni svoj dnevnik čitanja.	Task: Update the existing summary of the literary work based on the continuation of the story. Preserve the important earlier events, but revise or expand the summary so that it reflects the entire plot so far. Generate a unified comprehensive summary and analysis of the excerpts that will help the student complete their reading journal.
Upute za sadržaj: Tekst treba biti dugačak između 1500 i 1900 riječi. Tekst treba sadržavati sljedeće dijelove, jasno označene naslovima: 1. Kratak sadržaj: Napiši jasan i detaljan sažetak radnje. 2. Analiza likova: Odaberi i analiziraj glavne i sporedne likove koji su ključni za radnju, objašnjavajući njihove uloge i	Content guidelines: The text should be between 1,500 and 1,900 words. The text should contain the following parts, clearly labeled with headings: 1. Short summary: Write a clear and detailed summary of the plot. 2. Character analysis: Choose and analyze the main and supporting characters who are key to the plot,

osobine. 3. Podaci o djelu: Na kraju teksta, u posebnom odjeljku, navedi sljedeće informacije o djelu: * Književni rod * Književna vrsta * Mjesto radnje * Vrijeme radnje * Glavni lik * Sporedni likovi * Pouka djela	explaining their roles and traits. 3. Information about the work: At the end of the text, in a separate section, list the following information about the work: * Literary category * Literary genre * Setting (place) * Time of the story * Main character * Supporting characters * Moral/lesson of the work
Generiraj kompletan tekst u jednom odgovoru. Ne postavljaj dodatna pitanja.	Generate the complete text in a single response. Do not ask additional questions.
Slijedi tekst postojećeg sažetka: [SAŽETAK PRETHODNOG DIJELA IZVORNOG TEKSTA IDE OVDJE]	The following is the text of the existing summary: [SUMMARY OF PREVIOUS CHUNK OF SOURCE TEXT GOES HERE]
Slijedi tekst konačnog isječka iz književnog djela: [SLJEDEĆI DIO IZVORNOG TEKSTA IDE OVDJE]	The following is the text of the final excerpt from the literary work: [NEXT CHUNK OF SOURCE TEXT GOES HERE]

10.4. Appendix D: POS frequency distributions of *The Brave Adventures of Lapitch*, three human-authored summaries, and three ChatGPT-authored summaries

Universal Dependencies POS tags[5]
ADJ: adjective; ADP: adposition; ADV: adverb; AUX: auxiliary; CCONJ: coordinating conjunction; DET: determiner; INTJ: interjection; NOUN: noun; NUM: numeral; PART: particle; PRON: pronoun; PROPN: proper noun; PUNCT: punctuation; SCONJ: subordinating conjunction; SYM: symbol; VERB: verb; X: other.

	Lapitch	Sjedi5.com	Lektira.hr	Lektire.hr	ChatGPT SP	ChatGPT IU	ChatGPT HM
ADJ	1803 (6.03 %)	150 (6.94 %)	162 (7.79 %)	116 (6.66 %)	305 (10.26 %)	281 (10.99 %)	246 (10.75 %)
ADP	1872 (6.26 %)	162 (7.49 %)	165 (7.93 %)	133 (7.64 %)	198 (6.66 %)	182 (7.12 %)	144 (6.29 %)
ADV	2519 (8.43 %)	95 (4.39 %)	108 (5.19 %)	100 (5.74 %)	102 (3.43 %)	75 (2.93 %)	77 (3.36 %)
AUX	2905 (9.72 %)	169 (7.82 %)	267 (12.84 %)	211 (12.12 %)	124 (4.17 %)	55 (2.15 %)	64 (2.80 %)
CCONJ	1714 (5.74 %)	116 (5.37 %)	115 (5.53 %)	99 (5.69 %)	191 (6.43 %)	159 (6.22 %)	151 (6.60 %)
DET	1046 (3.50 %)	82 (3.79 %)	68 (3.27 %)	53 (3.04 %)	61 (2.05 %)	57 (2.23 %)	43 (1.88 %)
INTJ	40 (0.13 %)	0 (0.00 %)	0 (0.00 %)	0 (0.00 %)	0 (0.00 %)	0 (0.00 %)	0 (0.00 %)
NOUN	4201	431	391	340	661	738	546

[5] https://universaldependencies.org/treebanks/hr_set/index.html

	(14.06 %)	(19.94 %)	(18.80 %)	(19.53 %)	(22.24 %)	(28.86 %)	(23.85 %)
NUM	225 (0.75 %)	25 (1.16 %)	13 (0.63 %)	13 (0.75 %)	14 (0.47 %)	9 (0.35 %)	8 (0.35 %)
PART	467 (1.56 %)	20 (0.93 %)	23 (1.11 %)	18 (1.03 %)	39 (1.31 %)	20 (0.78 %)	20 (0.87 %)
PRON	1739 (5.82 %)	127 (5.87 %)	108 (5.19 %)	84 (4.82 %)	154 (5.18 %)	74 (2.89 %)	108 (4.72 %)
PROPN	1383 (4.63 %)	131 (6.06 %)	81 (3.89 %)	74 (4.25 %)	77 (2.59 %)	118 (4.61 %)	107 (4.67 %)
PUNCT	5018 (16.79 %)	282 (13.04 %)	212 (10.19 %)	155 (8.90 %)	590 (19.85 %)	508 (19.87 %)	412 (18.00 %)
SCONJ	1073 (3.59 %)	69 (3.19 %)	62 (2.98 %)	78 (4.48 %)	57 (1.92 %)	18 (0.70 %)	38 (1.66 %)
SYM	0 (0.00 %)	0 (0.00 %)	0 (0.00 %)	0 (0.00 %)	3 (0.10 %)	3 (0.12 %)	0 (0.00 %)
VERB	3875 (12.97 %)	303 (14.01 %)	305 (14.66 %)	267 (15.34 %)	396 (13.32 %)	260 (10.17 %)	325 (14.20 %)
X	6 (0.02 %)	0 (0.00 %)	0 (0.00 %)	0 (0.00 %)	0 (0.00 %)	0 (0.00 %)	0 (0.00 %)
Σ	29886	2162	2080	1741	2972	2557	2289