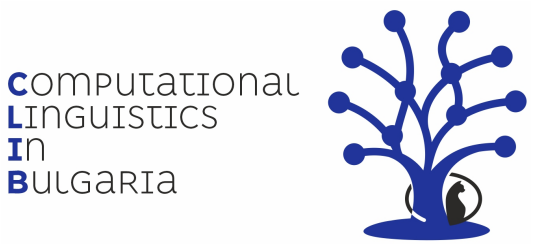
Institute for Bulgarian Language

BULGARIAN ACADEMY OF SCIENCES

Department of Computational Linguistics

Computational Linguistics in Bulgaria



Volume 2, **Issue** 1 • 2026

Editor-in-Chief: Prof. Dr. Svetla Koeva

Sofia, Bulgaria

PUBLICATION AND CATALOGUING INFORMATION

Journal title:	*Computational Linguistics in Bulgaria*
Volume and issue:	Volume 2, Issue 1, 2026
ISSN:	3033-1382 (print) / 3033-2397 (online)
Published and distributed by:	Institute for Bulgarian Language Bulgarian Academy of Sciences Department of Computational Linguistics
Editorial address:	Institute for Bulgarian Language Bulgarian Academy of Sciences Department of Computational Linguistics 52 Shipchenski Prohod Blvd., Bldg. 17 Sofia 1113, Bulgaria Phone/Fax: +359 2/ 872 23 02
Copyright:	The copyright of each article published in the journal belongs to its author(s).
	All content is licensed under a Creative Commons Attribution 4.0 International Licence (CC BY 4.0).
	
	License details: http://creativecommons.org/licenses/by/4.0
	Cover artwork by Bozhidar Chemshirov.
	©2026 Institute for Bulgarian Language, Bulgarian Academy of Sciences, Department of Computational Linguistics

Table of contents

Detecting AI-Generated Bulgarian Text: A Two-Step Multi-Class Classification Approach

Boyan Bogdanov
Sofia University "St. Kliment Ohridski"
Sofia, Bulgaria
boyan.bogdan@gmail.com

Daniel Georgiev
Sofia University "St. Kliment Ohridski"
Sofia, Bulgaria
georgiev.daniel00@gmail.com

Dimitar Dimitrov
Sofia University "St. Kliment Ohridski"
Sofia, Bulgaria
ilijanovd@fmi.uni-sofia.bg

Ivan Koychev
Sofia University "St. Kliment Ohridski"
Sofia, Bulgaria
koychev@fmi.uni-sofia.bg

Preslav Nakov
Mohamed bin Zayed University of
Artificial Intelligence
Abu Dhabi, UAE
preslav.nakov@gmail.com

This paper introduces a novel two-step multi-class classification system to identify varying degrees of machine involvement in Bulgarian text. As Large Language Models (LLMs) proliferate, distinguishing original human writing from machine-assisted or machine-generated content is crucial to prevent misinformation and preserve educational integrity. We developed a comprehensive dataset in Bulgarian encompassing purely human-written texts and four distinct mixed-content categories, such as machine-continued and machine-polished text. Our proposed pipeline utilises a binary ensemble classifier combining stylometric features, XLM-RoBERTa, and an adapted Binoculars model to first filter out human-written text. Subsequently, a specialised multi-class Support Vector Machine categorises the remaining machine-involved texts. Our findings indicate that this hierarchical approach improves the reliable identification of human authorship and enhances the overall accuracy of multi-class text detection compared to single-step methods.

Keywords: AI-generated text detection, large language models, benchmark evaluation, Bulgarian language

1. Introduction

The rapid spread of sophisticated Large Language Models (LLMs) has made it increasingly difficult to distinguish between text written by humans and text generated by machines, posing substantial problems in many areas. When LLM-generated content is passed off as

https://doi.org/10.47810/JCLIB.2.2026.01

original human work, it can fuel the spread of false information, violate intellectual property rights, and lower educational quality. Furthermore, the inability to reliably detect machine-generated text risks contaminating public datasets, which in turn degrades the quality of training data for future language models and hinders downstream linguistic research.

To tackle this, our paper introduces a new system designed to classify Bulgarian text into multiple categories. We built an extensive collection of texts that includes purely human-written content, purely machine-generated content, and four different types of mixed content: text that started with human input and was finished by a machine, text written by humans and refined by a machine, text generated by a machine but then made to sound more human, and text produced from deeply mixed human and machine involvement.

Our study tests both established machine learning techniques and newer LLM-based classifiers using this unique, multi-category dataset. Our aim is to create a reliable standard for identifying different levels of machine involvement in Bulgarian text, moving beyond simple yes/no detection to offer a more precise and context-sensitive analysis.

In summary, our main contributions are:

- The creation of a novel, multi-category Bulgarian dataset featuring purely human, machine-generated, and deeply interwoven text classes.
- An evaluation of the limitations of single-step multi-class models in distinguishing authentic human text from machine-assisted text.
- The proposal of a high-performing, two-step hierarchical classification pipeline that significantly improves the recall of human-written Bulgarian text.

2. Related work

The detection of machine-generated text is deeply rooted in traditional authorship attribution and stylometry, where machine learning techniques established the foundational methodologies. Early research heavily relied on feature-based methods such as n-gram frequencies, TF-IDF representations, and stylometric markers, features demonstrated to effectively distinguish writing styles (Stamatatos 2009), fed into classic classifiers like Support Vector Machines (SVMs), decision trees, and random forests (Joachims 1998; Potthast et al. 2017; Kestemont et al. 2019). These conventional models demonstrated significant success in identifying authorial style and, in some cases, even outperformed early neural networks in identifying the specific Large Language Model (LLM) used for generation. As the field evolved toward binary classification (human vs. machine) of increasingly sophisticated generated text, fine-tuned transformer models such as RoBERTa have consistently shown superior performance (Uchendu et al. 2020).

A significant factor influencing detection difficulty is the scale of the generator model; text produced by larger, more complex LLMs is inherently more challenging to identify (Jawahar, Abdul-Mageed, and Lakshmanan 2020; Solaiman et al. 2019; Crothers, Japkowicz, and Viktor 2023). While detectors trained on output from smaller models show moderate success in identifying text from larger ones, the inverse is more effective. Feature-based methods relying on stylometry and linguistic properties like text perplexity remain relevant,

but typically require longer text samples and are sensitive to the generator's configuration settings (Crothers, Japkowicz, and Viktor 2023).

Modern detection efforts heavily leverage transformer-based architectures. The 2024 PAN Voight-Kampff Generative AI Authorship Verification Task (Bevendorff et al. 2024) saw a majority of successful systems using embeddings from models like BERT, DeBERTa, and RoBERTa, often feeding these features into classifiers like LSTMs or SVMs (Bevendorff et al. 2024). The top-performing solution employed an ensemble of fine-tuned LLMs and the zero-shot detector Binoculars (Hans et al. 2024), which relies on a comparison between two LLMs. This highlights a trend towards combining multiple advanced models for state-of-the-art results.

However, many of these advanced detectors exhibit significant performance degradation when tested on out-of-domain data (Wang et al. 2024; Tufts, Zhao, and Li 2025). Zero-shot models such as Grover (Zellers et al. 2019) are effective primarily within the specific domain for which they were designed, such as news articles. This domain dependency is a recurring challenge, with some research exploring adversarial training (e.g., DANN) to create more domain-agnostic models (Abassy et al. 2024).

The task of multi-class classification, which mirrors the work in this paper, was recently defined by the 2025 PAN Voight-Kampff Generative AI Detection shared task (Bevendorff et al. 2025). The results from this task, as shown in Table 3 of the cited paper, indicate that fine-tuned, large-scale models like DeBERTa-v3-Large and Qwen-4B are favored by the top-performing teams (Macko 2025; Li, Qi, and Yan 2025; Zheng et al. 2025). Many successful approaches involve sophisticated fine-tuning strategies, including data augmentation for underrepresented classes, multi-head attention mechanisms, and Mixture of Experts (MoE) architectures (Li, Qi, and Yan 2025; Teja, Yadagiri, and Pakray 2025; Voznyuk, Gritsai, and Grabovoy 2025).

While many recent top-performing systems employ single-step architectures, the concept of a two-step or hierarchical approach is also an established approach in machine generated text detection. Previous studies have utilized multi-stage pipelines to first separate human-written text from generated content before applying more granular categorization, such as identifying the specific generator model or pinpointing the exact boundary where human authorship transitions into machine continuation (Kushnareva et al. 2024; Zeng et al. 2024). We attempt to build upon this hierarchical framework, adapting it to the multi-class setup containing deeply interwoven, machine-polished, and machine-humanised texts, which remains an underexplored task for low-resourced languages such as Bulgarian.

Research on Bulgarian text detection reveals unique challenges and is an actively developing field. Early efforts in identifying machine-generated Bulgarian content on social media demonstrated the difficulty of distinguishing textual deepfakes from human discourse, highlighting the need for robust, language-specific tools (Temnikova et al. 2023). More recently, models tailored specifically for Bulgarian, such as the Siamese Transformer model BuST, have been proposed to address these nuances (Maslo and Gargova 2025). Furthermore, a study using the M4 dataset found that detectors for Bulgarian perform best when trained exclusively on Bulgarian data, even in cross-generator scenarios (Wang et al. 2024).

The multi-class detection task is complicated further by the specific characteristics of the Bulgarian language and the varying quality of generated texts. Research shows

that modern multilingual LLMs often rely on English-centric vector spaces, leading to outputs that resemble literal translations and exhibit syntactic calquing (Li et al. 2025). For Bulgarian specifically, studies have observed that models sometimes hallucinate by introducing vocabulary and grammatical structures from more widely resourced Slavic languages, such as Russian (Berbatova and Salambashev 2023). However, the contemporary models utilized in our dataset (e.g., Gemini 2.5 Flash, o4-mini) often generate fluent and grammatically correct text. Consequently, the detection task has become significantly more difficult as classifiers can no longer rely solely on superficial grammatical mistakes or signs of literal translation. Instead, they must capture subtle stylistic rigidities, repetitive phrasing, and the unnatural blending of styles that occurs when human and machine inputs are interwoven, making multi-class detection a non-trivial problem.

Moreover, the problem is uniquely challenging for Bulgarian due to significant resource constraints. Unlike English and other high-resource languages, which benefit from numerous large-scale datasets specifically for machine-generated text detection, Bulgarian lacks comparable annotated corpora for multi-class classification. The majority of existing machine-generated text detection datasets are English-centric, and adapting these resources to Bulgarian is non-trivial due to linguistic differences.

3. Dataset

In this section, we describe in details how we construct our multi-class Bulgarian dataset. We first describe the sources and extraction processes for compiling the original human baseline. Subsequently, we outline the specific prompting strategies and large language models employed to generate the four distinct classes of machine-involved content. Finally, we discuss the text preprocessing steps required to prepare this diverse corpus for the classification pipeline.

3.1 Human-written texts

To ensure a diverse collection of human-written texts, we selected several sources to guarantee sufficient variation in content properties such as domain, author, topic, and length. We then used a portion of this collected data to generate the four other classes of text, employing multiple Large Language Models for each text.

We chose the online library of New Bulgarian University (NBU),[1] as a primary source for human-written texts due to its large volume of published content and relative accessibility enabling filtering texts by various fields.

As a first step, we filtered documents by their publication year, with a cutoff date of 2020. We did this to avoid contaminating the human-written dataset with texts that could have potentially been created using LLMs. To simplify the text extraction process, we further narrowed down the selection to only PDF files, excluding other materials like audio-visual content, entire books, and presentations. The library's search engine paginates the results

1 https://eprints.nbu.bg/

but also provides a JSON file containing hyperlinks to the articles, along with their titles, unique identifiers, and other metadata.

Class	Illustrative Snippet
Human-written	Като общо правило, материята на публичното право постоянно повдига множество практически и теоретични въпроси. Причините за това могат да се търсят в няколко насоки... *As a general rule, the subject matter of public law constantly raises numerous practical and theoretical questions. The reasons for this can be sought in several directions...*
Machine-continued	Като общо правило, материята на публичното право постоянно повдига множество практически и теоретични въпроси. Причините за това могат да се търсят в няколко насоки. Първо, сложността на правната рамка често води до различни интерпретации... *As a general rule, the subject matter of public law constantly raises numerous practical and theoretical questions. The reasons for this can be sought in several directions. First, the complexity of the legal framework often leads to different interpretations...*
Machine-polished	Материята на публичното право постоянно повдига множество практически и теоретични въпроси поради няколко причини. Първо, динамиката на политическите... *The subject matter of public law constantly raises numerous practical and theoretical questions for several reasons. First, the dynamics of political...*
Machine-humanised	Правната доктрина и практика постоянно се стремят да дефинират и обвържат основните категории, които формират публичното право. В центъра на това... *Legal doctrine and practice constantly strive to define and connect the main categories that form public law. At the center of this...*
Deeply-mixed	Като общо правило, материята на публичното право постоянно повдига множество практически и теоретични въпроси. Това се дължи на многообразието на публичноправните субекти... На първо място, динамиката на... *As a general rule, the subject matter of public law constantly raises numerous practical and theoretical questions. This is due to the diversity of public law subjects... First, the dynamics of...*

Table 1
A condensed structural illustration of the five text classes based on a single source paragraph. Human text is highlighted in grey, while machine-generated text is in green.

The filtered articles, typically several pages long, were unsuitable for the detection task, as real-world scenarios rarely involve generating such long texts at once. Therefore, we segmented each article into paragraphs.

Next, we extracted the text from each PDF document. We experimented with several specialized Python libraries, including spacy-layout, pdfminer, PyMuPDF, and PyPDF. However, due to the non-standard format of the articles, these open-source libraries failed to produce satisfactory results. Common issues included the inability to correctly identify paragraph breaks, improper spacing (either missing or excessive), and scrambled lines that rendered the text meaningless. The variety and inconsistency of these errors made it difficult to develop a common data cleanup solution.

We found the most effective system for text extraction to be Azure AI Document Intelligence, a paid cloud service accessible via a Python library.[2] It uses AI to perform tasks such as document classification, table extraction, and identifying document structure, including paragraphs and titles.

In addition to the academic articles from the New Bulgarian University library, we used several other corpora of human-written texts to study the task with a more extensive dataset. Relying on texts from limited domains can lead to classifiers overfitting to these domains and performing poorly on unseen data. For this reason, we chose to include two additional domains in our dataset: news articles and Wikipedia pages, as these domains are also vulnerable to abuse of LLM involvement and offer more variety in text structure and traits. Part of the news articles are reused from the SemEval 2025 Task 10: Multilingual Characterization and Extraction of Narratives from Online News (Rosenthal et al. 2025), while the rest of the news dataset was gathered following the data collection practices described in the SemEval shared task. The Wikipedia pages subset was built using their public API which allows exporting pages into a plain text format.

We followed the same procedure for generating the other text classes from this dataset as we did for the academic articles. Due to the different properties of these new texts, being shorter and more self-contained on average, we adjusted the instructions for the LLMs to improve the text generation quality, e.g. by better reflecting the expected traits of the output.

3.2 Machine-involved texts

For the creation of the text corpus for classes requiring a Large Language Model, we used several contemporary multilingual models: Gemini 2.0 Flash, Gemini 2.5 Flash, OpenAI o4-mini, and BgGPT 2.

The **machine-generated** class is not part of the classification task's definition, but it's required as the basis for the **machine-generated, then machine-humanised** class. We generated machine-generated texts by passing only the title of human-written articles, without using any of their content. To approximate the style of human articles while maintaining diversity, we specified a target number of paragraphs or words in the prompts for some examples. We then split the resulting text into paragraphs, which served as the

2 https://azure.microsoft.com/en-us/products/ai-foundry/tools/document-intelligence

basis for some of the other classes. Examples of titles and prompts used for the generation of the texts of this class can be found in the Appendices in Table 9.

For the **human-initiated, machine-continued** class, we employed two main approaches with several prompt variations. The first involved splitting a human-written paragraph at the end of a sentence to create two parts. We included the first part in the prompt to the LLM, tasking it with completing the paragraph to a length approximately equal to the original human-written one. However, a single paragraph's beginning does not always provide sufficient context for the LLM to generate a meaningful continuation that aligns with the original. To address this, we randomly prepended between 0 and 5 preceding paragraphs to provide additional context. An visualisation of how a paragraph is split into two can be found in Table 8 in the Appendices.

The prompts we used for generating the **human-written, then machine-polished** class were the most straightforward ones, representing instructions commonly used in real-world scenarios. To generate this part of the corpus, we provided a human-written paragraph to the LLM along with a simple instruction such as "improve the readability of this paragraph", "paraphrase this paragraph", or "make this paragraph clearer".

In the **machine-generated, then machine-humanised** class, the setup is similar to the human-written, then machine-polished class, with a few key differences. Most notably, an LLM originally wrote the source text instead of a human. For this purpose, we used text from the machine-generated dataset. We then instructed the generator to revise the text by adopting a persona (e.g., an assistant, a graduate, a student), attempting to simplify it or to make the text seem less like it was written by an LLM. To simulate a real-world situation where a text might be the result of several sequential instructions, we fed the output from the previous step back into the same LLM with a modified instruction in the prompt as shown in Table 10. Each paragraph randomly underwent between 0 and 2 additional LLM humanisation iterations after the initial generation.

The **deeply-mixed** texts, which contain interwoven parts of human-written and machine-generated paragraphs, resemble the human-initiated, machine-continued class. In both cases, we deleted a portion of the human-written text for the LLM to fill in. For the mixed texts, this masked portion can appear one or more times within the paragraph, removing entire sentences. To differentiate the mixed texts from the "human-initiated, machine-continued" class, we avoided masking the very beginning or end of texts. Similarly to the "human-initiated, machine-continued" class, we sometimes appended several preceding paragraphs at the start of the current one to serve as context, enabling the LLM to fill in the missing parts with more relevant and accurate content. Again, we used several different prompts to maintain diversity in the dataset. Table 11 showcases the steps in preprocessing the text in order to generate e deeply-mixed version of it.

Table 2 specifies the number of texts from the different sources in all classes. Table 3 shows the distribution per LLM used.

	Source Domain			
Text Class	**NBU**	**News**	**Wikipedia**	**Total**
Human-written	13,919	23,779	7,667	45,365
Machine-generated	23,970	36,889	9,220	70,079
Machine-continued	28,242	601	0	28,843
Machine-polished	28,319	2,850	0	31,169
Machine-humanised	17,885	1,445	0	19,330
Deeply-mixed	15,835	0	0	15,835
Total	**128,170**	**65,564**	**16,887**	**210,621**

Table 2
Distribution of texts by class and source domain.

	Language Model				
Text Class	**BgGPT-27B**	**Gemini 2.5 Flash**	**Gemini 2.0 Flash**	**OpenAI o4-mini**	**Total**
Human-written	–	–	–	–	45,365
Machine-generated	24,504	6,480	7,642	31,453	70,079
Machine-continued	9,148	9,526	601	9,568	28,843
Machine-polished	9,260	9,498	501	11,910	31,169
Machine-humanised	5,009	6,166	274	7,881	19,330
Deeply-mixed	0	7,863	0	7,972	15,835
Total	**47,921**	**39,533**	**9,018**	**68,784**	**210,621**

Table 3
Distribution of texts by class and LLM used.

3.3 Text Preprocessing and Tokenization Analysis

The raw extracted texts required specialized natural language processing to be viable for the multi-class Support Vector Machine (SVM). We utilized the `spaCy` library, configured specifically for the Bulgarian language, to perform tokenization, lemmatization, and stop-word removal.

A critical finding during our preliminary experiments was the inverse relationship between aggressive text normalization and classification accuracy for machine-generated text. Specifically, applying lemmatization slightly degraded the model's performance. The generalization of tokens achieved through lemmatization inadvertently stripped away morphological variations and specific word forms that serve as crucial stylistic markers (artifacts of generation) left by the LLMs.

Similarly, the removal of stop words presented a trade-off. While removing the approximately 1,000 stop words from the 280,000-word corpus shifted the vocabulary focus toward content-heavy words, it led to heavy overfitting during the multi-class SVM training. The

model achieved near-perfect scores on the training set but failed to generalize, indicating that LLMs often expose their nature through the structural use of highly frequent, seemingly "meaningless" stop words. Consequently, our final Bag-of-Words representation relies on simple `spaCy` tokenization without lemmatization or stop-word removal. The natural language processing is performed by using parts of the bespoke pipeline for Bulgarian developed by (Berbatova and Ivanov 2023).

4. Two-step classification pipeline

Our preliminary studies demonstrated the limitations of a single-step approach. When we trained a multi-class SVM to directly classify texts into all five categories, it struggled to reliably distinguish purely human-written content from the various forms of machine-involved text. This approach yielded a recall of 0.729 for the human-written class, as shown in the Single-Step Recall column of Table 6, indicating that a significant portion of human-authored texts were being misclassified as machine-generated or machine-assisted.

The primary motivation for developing the two-step pipeline is to address this specific shortcoming. By introducing a specialised, high-performance binary classifier in the first stage, we aim to more accurately filter out human-written content, thereby increasing the recall for this critical class and reducing the number of false positives that are unnecessarily passed to the second-stage classifier.

To address the nuanced task of identifying different forms of machine involvement in text generation, we propose a two-step classification pipeline. This hierarchical approach is designed to first distinguish purely human-written text from any text involving machine generation and then to further categorise the specific nature of the machine's contribution. The pipeline processes a given text and assigns it to one of five classes.

4.1 Data Splitting and Leakage Prevention

A fundamental principle in training and evaluating machine learning models is preventing data leakage between the training and testing sets. Because the lengthy articles from the New Bulgarian University (NBU) library were segmented into multiple individual paragraphs to match the generation constraints of LLMs, a naive randomized split introduces a severe vulnerability.

If a simple split is employed, different paragraphs from the same source article can simultaneously appear in both the training and testing subsets. This creates an artificial overlap in highly specific vocabulary, formatting, and authorial style. In a real-world environment, the system would not have prior access to adjacent paragraphs of an unseen text. To simulate a realistic operational environment and ensure objective evaluation, we implemented a "parent-child" grouping strategy during the dataset split. We enforce a strict constraint ensuring that all paragraphs ("children") originating from the same article ("parent") are grouped exclusively in a single subset, thereby preventing any stylistic or thematic data leakage.

Applying this parent-child grouping strategy for the first stage of the classification pipeline, the dataset was partitioned into training and testing subsets. The training set is made up of 112,451 samples while the test set contains 15,586 samples, maintaining an approximate 90:10 ratio.

Following this initial parent-child grouped split, the training set underwent a secondary partitioning to facilitate ensemble training. The 112,451 training samples were divided into two subsets: one for training the base models from the stacking ensemble model (approximately 96,865 samples) and another for training the meta-model (approximately 15,586 samples). By holding out a separate evaluation set, we ensure that the ensemble performance metrics reflect genuine generalization on completely unseen data, maintaining the integrity of our results and preventing the ensemble from having an unfair advantage through exposure to its training distribution. The final ratio of the three subsets is 80:10:10.

4.2 Binary ensemble

The first stage of our pipeline employs a binary classifier to perform the initial, high-level distinction between human-written and machine-involved text. This classifier is an ensemble model that leverages three distinct detection signals: an XGBoost classifier trained on stylometric text features, a fine-tuned XLM-RoBERTa instance and a version of Binoculars adapted to Bulgarian by replacing the original Falcon observer and performer LLMs with SambaLingo-Bulgarian Base and Instruct LLMs (Csaki et al. 2024), which in turn are pretrained Llama models. We evaluated these three solutions and found they complement each other, motivating us to implement a logistic regression ensemble that classifies by scaling the output of each model by a weight learned during training on a held-out subset.

Class	Precision	Recall	F1
0 (Human)	0.954	0.838	0.892
1 (Machine)	0.943	0.985	0.963
Macro Avg	**0.948**	**0.911**	**0.928**

Table 4
Performance of the first-stage binary ensemble classifier.

The performance of this first stage is shown in Table 4. If the ensemble classifies a text as human-written, it is assigned its final label, and no further processing occurs. However, if the text is flagged as machine-involved, the system passes it to the second stage of the pipeline for a more granular analysis.

4.3 Multi-class SVM

The system forwards texts identified as machine-involved by the first step to a second-stage classifier for more detailed analysis. It is important to note that this second classifier is a separate multi-class Support Vector Machine (SVM), distinct from the one used in our initial

single-step experiments. We trained this new SVM specifically on a dataset containing only the four machine-involved classes, rather than all five. The purpose of this specialised SVM is to differentiate between the four distinct categories of machine-assisted text generation.

For feature representation, the SVM utilizes a Bag-of-Words model. Each text is converted into a numerical vector that represents the frequency of words from a vocabulary built upon the training corpus. This traditional, yet effective approach allows the model to capture lexical patterns and stylistic markers that are characteristic of each generation method.

Results can be seen in Tables 5 and 6.

Model Architecture	Acc	Rec	F1
Single-Step SVM	0.633	0.659	0.657
Two-Step Pipeline	0.712	0.692	0.695

Table 5
Overall multi-class performance comparison between the single-step SVM baseline and the proposed two-step pipeline (macro averages).

Type of Text	Single-Step Recall	Two-Step Recall
human-written	0.723	0.838
machine-continued	0.637	0.685
machine-polished	0.723	0.758
machine-humanised	0.734	0.707
deeply-mixed	0.470	0.474
Macro recall	**0.659**	**0.692**

Table 6
Per-class recall comparison of the Single-Step SVM and the Two-Step Pipeline.

By combining the binary and multiclass classification stages, our pipeline provides comprehensive classification, first isolating purely human work and then providing a detailed breakdown of how a machine may have been involved in the creation of a given text. The two-step approach yields a significant improvement across all primary metrics over the single-step baseline for all classes except machine-humanised texts.

4.4 Class-Specific Difficulty and Error Analysis

While the macro-averaged metrics demonstrate the superiority of the two-step pipeline, examining the recall on a per-class basis reveals the details around the challenges of AI text detection. As demonstrated in Table 6, the **deeply-mixed** class proved to be the most resilient to detection, yielding a recall of under 0.50 even in the optimized pipeline. This indicates that a mixed text is statistically more likely to be misclassified as one of the other four categories than to be correctly identified. The seamless weaving of human and machine sentences

disrupts the continuous stylistic patterns that both stylometric features and n-gram models rely upon.

Conversely, the **machine-humanized** class-where an LLM text is recursively fed back into an LLM to sound more human-was consistently the easiest to identify across all classical classifiers. This suggests that repeated interactions and iterative prompting do not successfully mask the machine's origin; rather, they compound the artificial stylistic markers, making the resulting text highly distinguishable from authentic human writing.

Limitations

While our proposed dataset and two-step pipeline demonstrate strong improvements in detecting machine-involved text, we acknowledge several limitations in our study. First, the dataset generation process heavily relies on specific proprietary and open-weights models (such as Gemini and OpenAI models). As these LLMs are continuously updated or deprecated, exact reproduction of the generated text distributions may become challenging. Second, our study is strictly limited to the Bulgarian language. While the methodology is intended to be language-agnostic, the specific performance of models like the adapted Binoculars zero-shot detector and the underlying text extraction tools may vary significantly if applied to other languages with different properties. Finally, the stylometric features utilized in the first-stage binary ensemble are sensitive to text characteristics, such as length, part of speech and punctuation frequency, which leaves part of the ensemble more vulnerable to style change attacks.

5. Conclusion

As Large Language Models become increasingly adept at generating and refining Bulgarian text, standard binary detection techniques are no longer sufficient to capture various cases of human-machine collaboration. In this paper, we introduced a novel, multi-category dataset designed to represent varying degrees of machine involvement, ranging from purely human-authored texts to deeply interwoven human-machine content.

Our preliminary evaluations demonstrated the limitations of utilizing a single-step classifier to navigate this complex landscape, primarily due to its tendency to misclassify authentic human writing as machine-assisted. To resolve this, we implemented a hierarchical two-step classification pipeline. By leveraging a binary ensemble combining stylometric features, an adapted Binoculars model, and XLM-RoBERTa, we were able to accurately filtered out purely human-written text in the first stage. This allowed our specialised second-stage multi-class SVM to focus exclusively on categorizing the specific nature of machine generation.

The results confirm that the two-step methodology effectively ensures high recall of human-written content while providing a measurable improvement in overall multi-class accuracy. Future work will focus on expanding the dataset across more diverse text domains and exploring the direct integration of open-source, Bulgarian-specific LLMs into the second stage of the pipeline to further elevate multi-class detection capabilities.

Acknowledgments

This work is partially supported by the project UNITe BG16RFPR002-1.014-0004 funded by PRIDST and EU NextGenerationEU project, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project SUMMIT, No BG-RRP-2.004-0008.

References

Abassy, Mervat, Kareem Elozeiri, Alexander Aziz, Minh Ngoc Ta, Raj Vardhan Tomar, Bimarsha Adhikari, Saad El Dine Ahmed, Yuxia Wang, Osama Mohammed Afzal, Zhuohan Xie, Jonibek Mansurov, Ekaterina Artemova, Vladislav Mikhailov, Rui Xing, Jiahui Geng, Hasan Iqbal, Zain Muhammad Mujahid, Tarek Mahmoud, Akim Tsvigun, Alham Fikri Aji, Artem Shelmanov, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. LLM-detectAIve: a tool for fine-grained machine-generated text detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 231–243, Association for Computational Linguistics.

Berbatova, Melania and Filip Ivanov. 2023. An improved Bulgarian natural language processing pipeline. In *Annual of Sofia University "St. Kliment Ohridski", Faculty of Mathematics and Informatics*, volume 10, pages 37–50.

Berbatova, Melania and Yoan Salambashev. 2023. Evaluating hallucinations in large language models for Bulgarian language. In *Proceedings of the 8th Student Research Workshop associated with the International Conference Recent Advances in Natural Language Processing*, pages 55–63.

Bevendorff, Janek, Daryna Dementieva, Maik Froebe, Bela Gipp, Andre Greiner-Petter, Jussi Karlgren, Maximilian Mayerl, Preslav Nakov, Alexander Panchenko, Martin Potthast, Artem Shelmanov, Efstathios Stamatatos, Benno Stein, Yuxia Wang, Matti Wiegmann, and Eva Zangerle. 2025. Overview of PAN 2025: Generative AI detection, multilingual text detoxification, multi-author writing style analysis, and generative plagiarism detection. In *Proceedings of the 47th European Conference on Information Retrieval (ECIR)*.

Bevendorff, Janek, Matti Wiegmann, Jussi Karlgren, Luise Dürlich, Evangelia Gogoulou, Aarne Talman, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2024. Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2024. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, volume 3740 of *CEUR Workshop Proceedings*, pages 2486–2506, CEUR-WS.org.

Crothers, Evan N., Nathalie Japkowicz, and Herna L. Viktor. 2023. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002.

Csaki, Zoltan, Bo Li, Jonathan Lingjie Li, Qiantong Xu, Pian Pawakapan, Leon Zhang, Yun Du, Hengyu Zhao, Changran Hu, and Urmish Thakker. 2024. SambaLingo: Teaching large language models new languages. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 1–21, Association for Computational Linguistics.

Hans, Abhimanyu, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs with binoculars: zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, JMLR.org.

Jawahar, Ganesh, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2020. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain*, pages 2296–2309, International Committee on Computational Linguistics.

Joachims, Thorsten. 1998. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML'98, page 137–142, Springer-Verlag, Berlin, Heidelberg.

Kestemont, Mike, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast, and Benno Stein. 2019. Overview of the cross-domain authorship attribution task at PAN 2019. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2019), Lugano, Switzerland*, volume 2380, pages 1–15.

Kushnareva, Laida, Tatiana Gaintseva, German Magai, Serguei Barannikov, Dmitry Abulkhanov, Kristian Kuznetsov, Eduard Tulchinskii, Irina Piontkovskaya, and Sergey Nikolenko. 2024. AI-generated text boundary detection with RoFT. ArXiv: 2311.08349.

Li, Bohan, Haoliang Qi, and Kai Yan. 2025. Team Bohan Li at PAN: DeBERTa-v3 with R-drop regularization for human-AI collaborative text classification. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain*, volume 4038 of *CEUR Workshop Proceedings*, pages 3763–3769, CEUR-WS.org.

Li, Yafu, Ronghao Zhang, Zhilin Wang, Huajian Zhang, Leyang Cui, Yongjing Yin, Tong Xiao, and Yue Zhang. 2025. Lost in literalism: How supervised training shapes translationese in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12875–12894.

Macko, Dominik. 2025. mdok of KInIT: Robustly fine-tuned LLM for binary and multiclass AI-generated text detection. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain*, volume 4038 of *CEUR Workshop Proceedings*, pages 3819–3826, CEUR-WS.org.

Maslo, Andrii and Silvia Gargova. 2025. BuST: A Siamese transformer model for AI text detection in Bulgarian. In *Proceedings of Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models*, pages 45–52.

Potthast, Martin, Francisco Manuel Rangel-Pardo, Michael Tschuggnall, Efstathios Stamatatos, Paolo Rosso, and Benno Stein. 2017. Overview of PAN'17: Author identification, author profiling, and author obfuscation. In *Lecture Notes in Computer Science*, volume 10456, pages 275–290, Springer.

Rosenthal, Sara, Aiala Rosá, Debanjan Ghosh, and Marcos Zampieri, editors. 2025. *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.

Solaiman, Irene, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *CoRR*, abs/1908.09203.

Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Teja, Lekkala Sai, Annepaka Yadagiri, and Partha Pakray. 2025. Team CNLP-NITS-PP at PAN: advancing generative AI detection: Mixture of experts with transformer models. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain*, volume 4038 of *CEUR Workshop Proceedings*, pages 3904–3915, CEUR-WS.org.

Temnikova, Irina, Iva Marinova, Silvia Gargova, Ruslana Margova, and Ivan Koychev. 2023. Looking for traces of textual deepfakes in Bulgarian on social media. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, Varna, Bulgaria*, pages 1151–1161.

Tufts, Brian, Xuandong Zhao, and Lei Li. 2025. A practical examination of AI-generated text detectors for large language models. In *Findings of the Association for Computational Linguistics (NAACL 2025), Albuquerque, New Mexico*, pages 4839–4856, Association for Computational Linguistics.

Uchendu, Adaku, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Association for Computational Linguistics.

Voznyuk, Anastasia, German Gritsai, and Andrey Grabovoy. 2025. Team advacheck at PAN: multitasking does all the magic. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025)*, volume 4038 of *CEUR Workshop Proceedings*, pages 4007–4014, CEUR-WS.org.

Wang, Yuxia, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, Association for Computational Linguistics.

Zellers, Rowan, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32, pages 9054–9065, Curran Associates, Inc.

Zeng, Zijie, Lele Sha, Yuheng Li, Kaixun Yang, Dragan Gašević, and Guangliang Chen. 2024. Towards automatic boundary detection for human-AI collaborative hybrid essay in education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22502–22510.

Zheng, Miaoji, Yong Zhong, Fen Liu, Tufeng Xian, Meifang Xie, Weidong Wu, Zhiliang Zhang, and Qiyuan Sun. 2025. StarBERT: a hybrid neural network model for human-AI collaborative text classification. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain*, volume 4038 of *CEUR Workshop Proceedings*, pages 4080–4085, CEUR-WS.org.

Appendices

A. LLM parameters used during generation

Generator Model	Configuration / Hyperparameters
BgGPT 2	`model` = insait/insait-gemma-2-27b-bg-mixed-19847, `temperature` = 0.7, `max_tokens` = 2048, `top_k` = 20, `repetition_penalty` = 1.1, `stop` = [”<end_of_turn>”, ”<eos>”]
Gemini (2.0 Flash & 2.5 Flash)	Default API settings (no explicit parameters passed)
OpenAI o4-mini	Default API settings (no explicit parameters passed)

Table 7
Configuration parameters for the Large Language Models used in the generation of the machine-involved text classes.

B. Prompts and steps for generating texts

Human-written paragraph	Split paragraph into two
“Друга специализирана система...”	“Друга специализирана система... е FACS... Тя позволява кодиране на микромимики...”
“Another specialized system...”	*“Another specialized system... is FACS... It enables the coding of micro-expressions...”*

Table 8
An illustration of how a human-written paragraph is split into two parts.

Title of article	Prompt
"Херменевтика и чуждоезиково обучение"	"Напиши научна статия от около 14 параграфа със заглавие "Херменевтика и чуждоезиково обучение". Не включвай заглавието на статията или заглавия на секции."
"Hermeneutics and Foreign Language Teaching"	*"Write a scientific article of approximately 14 paragraphs with the title "Hermeneutics and Foreign Language Teaching". Do not include the article title or section headings."*
"Стилистика на езика, стилистика на речта или семантическа стилистика"	"Напиши научна статия от около 2564 думи със заглавие "Стилистика на езика, стилистика на речта или семантическа стилистика". Не включвай заглавието на статията или заглавия на секции."
"Stylistics of Language, Stylistics of Speech or Semantic Stylistics"	*"Write a scientific article of approximately 2564 words with the title "Stylistics of Language, Stylistics of Speech or Semantic Stylistics". Do not include the article title or section headings."*

Table 9
A selection of NBU human-written article titles alongside the corresponding prompts given to the large language model o4-mini for generating paragraphs in the "machine-generated" class. The original titles and prompts in Bulgarian as well as English translations are provided.

Instruction	Result
Машинен текст, използван за вход *Machine-generated text used as input*	"Препоръчва се комбиниран подход, който съчетава традиционните техники..." *"A combined approach integrating traditional manual pattern-making..."*
"Промени този параграф да звучи сякаш е писан от човек..." *"Rewrite this paragraph to sound as if it were written by a human..."*	"Комбиниран подход, който обединява класическото ръчно моделиране..." *"A combined approach uniting classical manual pattern-making..."*
"Промени го да звучи сякаш е писан от студент." *"Rewrite it to sound as if it were written by a student."*	"Смятам, че най-добре работи комбинацията..." *"I think the best approach is the combination..."*
"Промени го да звучи по-просто." *"Rewrite it to sound simpler."*	"Най-добре е да комбинирам ръчно моделиране..." *"The best approach is to combine manual pattern-making..."*

Table 10
The sequence of steps applied to a machine-generated text for the "machine-generated, then machine-humanized" class. The original instructions and texts are in Bulgarian; English translations are provided for the reader.

Step	Text
Текст, написан от човек *Human-written text*	"Друга специализирана система... е FACS... FACS представлява..." *"Another specialized system... is FACS... FACS is an anatomically based system..."*
Замаскиран текст *Masked text*	"⟨⟨ПРОПУСНАТА ЧАСТ⟩⟩. FACS представлява..." *"⟨⟨OMITTED PART⟩⟩. FACS is an anatomically based system..."*
Резултат от големия езиков модел *Large language model output*	"Табутата при одита на правното звено водят до игнориране... FACS представлява..." *"Taboos in the audit of the legal unit lead to the ignoring... FACS is an anatomically based system..."*

Table 11
An example of a text with an omitted beginning that the large language model must fill in to produce a "deeply mixed" text.

C. Hyperparameters and training details

Model	Key Hyperparameters & Configuration	Hardware & Runtime
XGBoost	`objective` = binary:logistic, `eval_metric` = logloss, `n_estimators` = 1000, `max_depth` = 6, `learning_rate` = 0.023, `subsample` = 0.722, `colsample_bytree` = 0.843, `gamma` = 0.271, `min_child_weight` = 2, `reg_alpha` = 0.966, `reg_lambda` = 2.617, `early_stopping_rounds` = 50	Intel Core i7 CPU 15 seconds
XLM-RoBERTa	`learning_rate` = 2e-5, `num_train_epochs` = 2, `per_device_train_batch_size` = 16, `per_device_eval_batch_size` = 64, `gradient_accumulation_steps` = 4, `weight_decay` = 0.1, `warmup_steps` = 500, `max_length` = 128, `eval_steps` = 250, `metric_for_best_model` = loss	NVIDIA T4 GPU 2:17 hours
Binoculars	`observer` = SambaLingo-Bulgarian-Base, `performer` = SambaLingo-Bulgarian-Chat	NVIDIA T4 GPU 3:07 hours
Logistic Regression	`max_iter` = 1000	Intel Core i7 CPU 10 seconds
Multi-class SVM	`C` = 0.8, `kernel` = rbf, `gamma` = scale, `class_weight` = balanced, `decision_function_shape` = ovr, `tol` = 0.001, `random_state` = 42	Intel Core i7 CPU 6:45 hours

Table 12
Summary of key hyperparameter configurations and execution environments for the models evaluated in this study.

ELEXIS-Ro – Adding Romanian to the ELEXIS Corpus

Verginica Barbu Mititelu
Romanian Academy, Research Institute for Artificial Intelligence
Bucharest, Romania
vergi@racai.ro

Elena Irimia
Romanian Academy, Research Institute for Artificial Intelligence
Bucharest, Romania
elena@racai.ro

Cătălin Mihăilă
Romanian Academy, Research Institute for Artificial Intelligence
Bucharest, Romania
catalin@racai.ro

Simina Popa
University of Bucharest
Bucharest, Romania
popa.simina2005@gmail.com

Lea Radu
The Saint Sava National College
Bucharest, Romania
leacristiana@icloud.com

Ioana Bucșa
Gheorghe Șincai National College
Bucharest, Romania
ioana.bucsa34@gmail.com

Mihaela Cristescu
University of Bucharest
Bucharest, Romania
mihaela.cristescu@litere.unibuc.ro

The paper presents, against the landscape of corpora dedicated to or also containing the Romanian language, the extension of the ELEXIS corpus with yet another language, Romanian, which both enhances the value of the parallel corpus and offers Romanian new opportunities for comparative and contrastive research within the Romance language family, as well as with the Balkan languages.

The Romanian component has been integrated primarily through automatic methods, followed by systematic manual validation to ensure annotation quality and cross-linguistic consistency. The levels of processing and annotation are: automatic translation of the English corpus into Romanian, automatic tokenization of the sentences, automatic morpho-syntactic annotation and semantic annotation (i.e., annotation with multiword expressions).

The annotation frameworks are widely used, multilingual ones: for morphology and syntax we used Universal Dependencies, while for multiword expressions, we made use of PARSEME guidelines.

Keywords: parallel corpus, Romanian, morpho-syntactic annotation, multiword expressions

https://doi.org/10.47810/JCLIB.2.2026.02

1. Introduction

Parallel corpora, defined as collections of texts aligned across two or more languages, constitute essential linguistic resources for multilingual research and language technology development. By providing direct correspondences between original texts and their translations, parallel corpora enable systematic investigation of how meaning, structure, and discourse features are transferred across languages. These resources facilitate the identification of translation patterns, equivalence relations, and cross-linguistic variation, offering valuable insights into both universal and language-specific phenomena. When translations included in such corpora are carefully produced and validated – either by professional translators or through rigorous quality control procedures – the resulting datasets become reliable repositories of linguistic and conceptual knowledge.

Despite the rapid advancement and widespread adoption of Large Language Models (LLMs) in recent years, parallel corpora remain indispensable. While LLMs demonstrate remarkable capabilities in multilingual understanding and generation, their performance still depends heavily on the availability of high-quality multilingual training data. Parallel corpora provide structured and verifiable bilingual or multilingual alignments that can be used to evaluate, fine-tune, and interpret these models. Furthermore, unlike automatically generated multilingual data, curated parallel corpora offer transparent mappings between languages, allowing researchers to assess the fidelity of meaning transfer and to analyze linguistic phenomena with a higher degree of confidence. In this sense, parallel corpora continue to play a crucial role in grounding data-driven approaches and ensuring that multilingual language technologies remain interpretable and reliable.

By incorporating low-resourced languages into parallel datasets, researchers not only preserve linguistic diversity but also enable cross-lingual transfer learning, domain adaptation, and the development of multilingual systems that perform more equitably across languages. Such corpora can support tasks including machine translation, cross-lingual information retrieval, terminology extraction, and linguistic typology, while also contributing to language preservation and revitalization efforts.

Moreover, the usefulness of a parallel corpus increases substantially with the addition of multiple levels of linguistic annotation. Beyond sentence-level alignment, annotations may include morphological, syntactic, semantic, discourse, and pragmatic information, as well as terminology, named entities, and domain-specific labels. Multi-layer annotation enables researchers to investigate correspondences across linguistic levels, revealing how grammatical structures, semantic roles, discourse relations, and stylistic features are translated or adapted across languages. These enriched corpora facilitate more sophisticated analyses and support downstream applications such as supervised machine translation, cross-lingual parsing, multilingual terminology management, and knowledge extraction.

In this context, richly annotated multilingual parallel corpora represent high-value resources for both theoretical and applied research. They not only support linguistic analysis and comparative studies, but also serve as valuable datasets for training, evaluating, and improving multilingual language technologies. As research increasingly moves toward multilingual and inclusive approaches, the development of high-quality, multi-layered parallel

corpora – especially those including low-resourced languages – remains a priority for advancing both linguistic understanding and language processing.

Against this background, we present the extension of the ELEXIS parallel corpus (Martelli et al. 2021) through the addition of Romanian. This inclusion represents a significant enhancement of the resource, both in terms of linguistic coverage and research potential. With the integration of Romanian, the corpus now includes a fourth Romance language, alongside the existing Romance-language components. This enrichment not only increases the multilingual scope of the corpus, but also opens new opportunities for comparative and contrastive research within the Romance language family. Researchers will be able to investigate similarities and differences in lexical choices, syntactic structures, semantic mappings, and discourse strategies across closely related languages, thereby contributing to a deeper understanding of language variation and evolution within the Romance domain.

Furthermore, the addition of Romanian is particularly valuable from a typological and areal linguistic perspective. Romanian occupies a distinctive position within the Romance language family due to its geographical location in Eastern Europe and its long-standing contact with Slavic and other Balkan languages. As a result, Romanian exhibits features that differentiate it from Western Romance languages, including structural and lexical characteristics influenced by language contact. The availability of Romanian within the ELEXIS parallel corpus therefore enables researchers to explore not only genealogical relationships within the Romance family, but also contact-induced phenomena and convergence patterns.

In this respect, the simultaneous presence of Bulgarian in the corpus further enhances its research potential. Romanian and Bulgarian are both part of the Balkan linguistic area (the Balkan Sprachbund), where languages from different families share common structural features due to prolonged contact and mutual influence. The inclusion of both languages within the same parallel corpus makes it possible to conduct systematic comparisons of Balkan-specific phenomena, such as definiteness marking, analytic constructions, clitic doubling, and other shared morphosyntactic traits. Consequently, the extended corpus supports investigations at multiple levels: genealogical comparisons among Romance languages, typological comparisons across language families, and areal studies within the Balkan linguistic context.

Overall, the addition of Romanian strengthens the ELEXIS parallel corpus as a multilingual and multi-purpose resource, broadening its applicability for contrastive linguistics, translation studies, typology, and multilingual Natural Language Processing (NLP).

The ELEXIS corpus is a manually annotated multilingual parallel dataset designed to support both lexicographic research and Word Sense Disambiguation (WSD). The project was first introduced in 2021 and initially developed for ten European languages: Bulgarian, Danish, Dutch, English, Estonian, Hungarian, Italian, Portuguese, Slovene, and Spanish. The dataset consists of parallel sentences extracted from WikiMatrix, an open-access collection of aligned sentences derived from Wikipedia. These sentences were first automatically selected and subsequently manually validated to ensure linguistic quality and semantic clarity. Additional selection criteria were applied, including a minimum length of five words per sentence and the presence of at least two polysemous lexical items, resulting in a carefully curated corpus comprising 2,024 sentences.

The corpus is enriched with five layers of linguistic annotation: tokenisation, subtokenisation, lemmatisation, part-of-speech tagging, and WSD. These annotation layers provide a detailed linguistic representation that supports both fine-grained semantic analysis and cross-linguistic comparison. In particular, the WSD layer plays a central role in the project, as it enables the identification and alignment of word senses across multiple languages. This multilingual semantic annotation supports lexicographic work by refining sense inventories, identifying gaps or inconsistencies in existing lexical resources, and facilitating the development of more precise and comprehensive dictionaries. At the same time, the dataset provides a valuable benchmark for evaluating and improving WSD systems, particularly in multilingual and cross-lingual contexts.

The primary objective of the ELEXIS project is to complete and enhance the semantic annotation across all participating languages while simultaneously improving and expanding existing sense inventories. By offering consistent semantic annotation across multiple languages, the corpus establishes a new reference point for multilingual WSD systems and promotes interoperability among lexical resources. Moreover, the dataset contributes to advancing multilingual NLP by providing high-quality, manually curated data that can be used for training, evaluation, and comparative analysis. This is particularly important given that current NLP resources and models are still heavily biased toward English, resulting in uneven technological development across languages.

In this context, the addition of Romanian represents a valuable and timely extension of the corpus. Romanian remains relatively under-resourced in comparison to major European languages, particularly in the area of semantically annotated multilingual datasets. Including Romanian in the ELEXIS corpus would expand linguistic coverage, enhance the representation of Eastern European languages, and support cross-lingual semantic research involving both Romance and Balkan languages. Furthermore, the availability of Romanian data in a richly annotated multilingual corpus would facilitate the development and evaluation of NLP tools for Romanian, including WSD systems, machine translation, semantic parsing, and lexicographic applications.

The paper is structured as follows: we present the landscape towards which we paint the development of this new corpus, part of a parallel one (Section 2): first, we outline the results of the work in corpora development for Romanian (Subsection 2.1) and then we present other endeavours for extending the ELEXIS corpus (Subsection 2.2). We continue with the steps taken in the development of the ELEXIS-Ro, the Romanian component of the ELEXIS corpus (Section 3): the translation of the English component into Romanian (Subsection 3.1) and then each of the annotation phase in turn: tokenization (Subsection 3.2), morpho-syntactic annotation (Subsection 3.3) and the semantic annotation of the multiword expressions (Subsection 3.4). Concluding the paper, we also announce the future work, i.e., the new annotation levels.

2. Related work

2.1 Romanian corpora

Romanian language corpora have been developed both as components of multilingual resources and as independent monolingual collections, each serving different research

purposes and covering a variety of linguistic domains. These corpora constitute essential resources for NLP, computational linguistics, and corpus-based linguistic studies, providing structured linguistic data for tasks such as machine translation, terminology extraction, and language modeling.

Within the category of multilingual corpora that include Romanian, several significant resources can be identified. One of the most important is the RO-JRC-Acquis (Ceauşu 2008), which contains over 30 million words. This corpus is derived from the European Commission Joint Research Centre's legislative documents and reflects predominantly the legal domain. Due to its parallel structure and terminological consistency, the corpus is particularly valuable for legal language processing and multilingual alignment studies.

Another important multilingual resource is the EUR-Lex Judgements Corpus (Baisa et al. 2016), which contains more than 17 million words. This corpus consists of judicial decisions from European Union institutions and, similarly to RO-JRC-Acquis, reflects the legal domain. Its structured format and domain specificity make it suitable for legal text mining, cross-lingual legal research, and computational legal linguistics.

The EUROPARL Corpus (Koehn 2005) represents another widely used multilingual dataset, containing nearly 10 million Romanian words extracted from the proceedings of the European Parliament. This corpus is particularly valuable for studying formal spoken language, parliamentary discourse, and translation patterns across European languages.

A larger multilingual resource that includes Romanian is OPUS (Tiedemann 2012), which contains approximately 300 million Romanian words. OPUS aggregates multiple parallel corpora collected from various sources, including legislative, administrative, and institutional texts. Due to its size and diversity, OPUS is frequently used for training machine translation systems and for large-scale multilingual language modeling.

In addition to multilingual corpora, several monolingual Romanian corpora have also been developed. Among these, the RoCo-news Corpus (Tufiş and Irimia 2006) contains approximately 7 million tokens and primarily reflects journalistic language. This corpus is particularly useful for studies focusing on contemporary Romanian media discourse and stylistic analysis.

Another significant monolingual resource is the RoWaC (Macoveiciuc and Kilgarriff 2010), which contains nearly 45 million words collected from web sources. RoWaC captures a wide range of informal and semi-formal language usage, making it valuable for studying contemporary linguistic variation, web-based discourse, and emerging vocabulary.

The ROMBAC (Ion et al. 2012) represents a carefully designed balanced corpus containing approximately 36 million words. The corpus is distributed relatively evenly across five domains: journalistic texts, pharmaceutical and medical texts, legal documents, literary history, and fiction. Due to this balanced composition, ROMBAC is particularly suitable for general-purpose linguistic research and domain comparison studies and it was at the core of the design and development of the next major Romanian corpus, CoRoLa.

The culmination of these efforts of corpus creation was the development of the reference corpus designed to reflect contemporary written and spoken Romanian language usage, CoRoLa (Barbu Mititelu, Tufiş, and Irimia 2018), which represented a priority research initiative of the Romanian Academy. This project was coordinated by two of its specialized

institutes, namely the Research Institute for Artificial Intelligence in Bucharest and the Institute for Computer Science in Iași. Despite this institutional coordination, the creation of the corpus evolved into a nationwide collaborative effort. In addition to this, the project also incorporated an important international component. Through the DRuKoLA Project, funded by the Alexander von Humboldt Foundation, the corpus – CoRoLa – benefited from a robust technological infrastructure for indexing and querying its content (Diewald et al. 2016).

At the time of its development, CoRoLa was conceived as a multipurpose resource designed to meet the needs of several user communities. For linguists, the corpus provides empirical evidence for a wide range of linguistic phenomena, allowing researchers to investigate frequency patterns, stylistic variation, domain-specific language use, and other characteristics of contemporary Romanian. The availability of large-scale, naturally occurring textual data enables both qualitative and quantitative linguistic analyses. For language engineers and researchers in NLP, CoRoLa serves as a valuable resource for training computational models, including word embeddings and other statistical representations of language. These representations, extracted from the corpus, support a wide range of applications such as automatic text classification, information retrieval, machine translation, and semantic analysis. At the same time, the corpus is also intended for use by the general public. Users can consult the corpus to identify authentic language usage examples, explore word meanings in context, and examine collocational patterns. This functionality makes CoRoLa a useful tool not only for academic research, but also for language education, lexicography, and professional writing. More recently, international collaborative projects have provided new opportunities to resume and expand the corpus. Particular attention has been paid to ensuring the authenticity and reliability of the data, especially in the context of the growing prevalence of AI-generated text. To minimize the inclusion of artificially generated content, web crawling activities were restricted to materials published prior to 2023. This precautionary measure aims to preserve the corpus as a reliable representation of naturally occurring contemporary Romanian language usage (Irimia et al. 2026).

Together, these corpora provide extensive coverage of Romanian language usage across multiple domains and registers, forming a solid foundation for both linguistic analysis and NLP applications.

2.2 Extension of ELEXIS with other languages

The ELEXIS project[1] has led to the development of a parallel sense-annotated corpus across ten languages (see above, Section 1). During the UniDive COST Action (Savary et al. 2024), the corpus was extended with several new languages, as shown below.

The Danish component, DA-ELEXIS, (Pedersen et al. 2023) is one of the largest sense-annotated corpora available for Danish and constitutes an important component of the broader ELEXIS multilingual infrastructure. The corpus was designed with multiple annotation layers, including tokenisation, sub-tokenisation, lemmatisation, and part-of-speech

1 https://project.elex.is/

(POS) tagging. These initial layers were automatically generated following the Universal Dependencies (de Marneffe et al. 2021) guidelines and subsequently manually verified to ensure annotation accuracy and consistency.

The Danish corpus contains 32,524 tokens, with sense annotations provided for all content words. These annotations include 7,322 nouns, 3,099 verbs, 2,626 adjectives, and 1,677 adverbs[2]. The annotation workflow was supported by a dedicated annotation tool that guided annotators through each token in its sentential context. For each token, the system presented all available senses, allowing annotators to select the most appropriate meaning based on contextual interpretation. This semi-automated annotation process facilitated consistency while maintaining human oversight.

The primary sense inventory used for senses annotation was DanNet, which covers approximately 70,000 Danish lemmas. In cases where compound words were not present in the sense repository, the DA-ELEXIS corpus applied compound splitting into constituent lemmas that could be found therein. This strategy enabled semantic tagging of otherwise unavailable lexical items. However, compounds already present in DanNet were preserved as single units.

Multiword expressions (MWEs) were encoded using a single semantic label. When MWEs appeared discontinuously in the corpus, the entire expression was subsequently identified and linked to lexical resources to ensure consistent annotation.

To evaluate annotation reliability, slightly more than 5% of the corpus (108 sentences) was triple-annotated. Inter-annotator agreement was measured using Cohen's kappa, resulting in an average score of 0.68. After clarifying annotation guidelines and resolving inconsistencies, agreement improved to 0.78 in later stages of the project, indicating satisfactory annotation reliability for semantic tagging tasks.

Krstev et al. (2024) present their work on the Serbian extension of the ELEXIS. The project focuses primarily on semantic annotation and the creation of lexical sense repositories, while also exploring multilingual correspondences across all ten languages included in ELEXIS, particularly for MWEs and named entities (NEs). The sentences from WikiMatrix were translated from English into Serbian using Google Translate, followed by manual verification. Special attention was given to incorrectly translated MWEs, morphological inconsistencies, and phonetic transcriptions of NEs. Tokenisation, lemmatisation, and POS tagging were initially performed automatically and subsequently manually validated to ensure annotation quality.

Because the number of MWEs and NEs varies across languages, expressions from all ten languages were automatically translated into Serbian as phrases rather than word-by-word equivalents. This process revealed structural differences between languages, including instances where MWEs in other languages correspond to single words in Serbian and vice versa. The annotation process produced 529 non-verbal MWE occurrences (i.e. 339 unique ones): 351 nominal MWEs, 133 proper nouns, 44 adverbial expressions, and one adjectival expression. Additionally, 230 occurrences of verbal MWEs were identified, representing 98 distinct expressions. No adjectival or verbal similes were found in this dataset.

2 These numbers can be compared with the corresponding ones in Romanian from Table 2.

Named Entity Recognition (NER) tools identified multiword NEs such as organizations and geopolitical names, which were recognized both by lexical dictionaries and by automatic NER systems. The applied systems are designed to identify new entities when compared with datasets in other languages or following manual validation, allowing the Serbian repository to expand iteratively.

The primary lexical resource used for Serbian word sense annotation was the Serbian WordNet (SrpWN) (Krstev et al. 2004). The English ELEXIS sense inventory is based on the Princeton WordNet (PWN) (Miller 1995; Fellbaum 1998). Because the dataset did not include the WordNet interlingual index, alignment between the two datasets was achieved by comparing definitions. Synonym lists from PWN were automatically translated into Serbian using both Google API and OpenAI technologies. Additional lexical resources from previous research projects were also incorporated to expand the Serbian subset. The results indicated that approximately 39% MWEs from the Serbian ELEXIS dataset were found in the SrpWN. Some entities were present in multiple synsets, while in other cases Serbian used single lexical items where other languages employed MWEs. These findings highlight both cross-linguistic variation and the challenges of multilingual semantic alignment.

3. Methodology

In this section we focus on the work done for adding Romanian to the ELEXIS corpus. At the moment of this writing, the following steps have already been taken: translation of the English corpus, tokenization of the Romanian corpus, morpho-syntactic annotation, semantic annotation with types of MWEs. Each step is described in what follows.

In accordance with the methodology adopted in the ELEXIS parallel corpus, the translation and annotation of the Romanian component are initially performed automatically using a range of specialized tools for machine translation, tokenization, lemmatization, and part-of-speech tagging. These automatically generated annotations are subsequently manually revised, corrected, and validated in order to ensure linguistic accuracy, consistency across languages, and compatibility with the multilingual sense-annotation framework. This combined automatic–manual workflow enables efficient corpus expansion while maintaining high-quality annotation standards required for cross-linguistic semantic analysis and lexicographic applications.

3.1 Translation

Automatic translation from English into Romanian was done using Google Translate. During manual inspection, 1,437 sentences (out of 2024) required no manual interventions, thus were considered correct translations. This means the rest 30% of the sentences were manually corrected.

Machine translation (MT) errors cluster around a limited number of recurring patterns where human intervention becomes systematic and predictable. The first and most frequent category concerns lexical and register-related errors, namely the selection of formally correct equivalents that are nevertheless contextually or stylistically inappropriate:

- MT systems often produce "acceptable" but imprecise or stylistically neutral synonyms, such as *se prezintă* instead of *apare* for the meaning "appears", *au vizitat premiera* instead of *au participat la premiera* "attended (the premiere)", *lucrarea sa ulterioară* instead of *opera sa ulterioară* referring to "his later work", or *rezolvarea completă* instead of *vindecarea completă* for "complete recovery".
- The same phenomenon occurs in terminological contexts, where *cursuri de master* is used instead of *masterclass-uri*, *raze X* instead of *radiografii* for the translation of "X-rays" as a medical investigation procedure, or *grupare funcțională* is used instead of *grupă funcțională*. In such cases, the human translator acts as a filter of contextual appropriateness, selecting forms that correspond to the appropriate register – technical, academic, or colloquial – and to natural Romanian usage, without altering the underlying meaning.
- This category also includes unnatural collocations and literal translations, where MT reproduces source language structures: e.g., *a acordat un rating mare* "gave a high rating", or *împărtășește o graniță* "shares a border", which are subsequently normalized by human intervention into *a lăudat* "praised" and, respectively, *are o graniță comună* "has a common border", as well as non-idiomatic formulations such as *orașul a fost stabilit* instead of *orașul a fost întemeiat* for the English "the city was established", or *se bazează pe* instead of *se inspiră din* for the English "is inspired by".

The second major category involves morphosyntactic and structural errors, where MT transfers patterns from the source language or produces incorrect agreement and syntactic relations. These include:

- agreement errors: *Majoritatea populației este catolic* instead of *catolică* (i.e., MT used a masculine instead of a feminine adjective), *au devenit membră* instead of *membre* (singular instead of pluaral),
- incorrect determiner: *oricui pilot* instead of *oricărui pilot* (*oricui* is not a determiner, unlike *oricărui*),
- incomplete or ambiguous constructions: *Dacă din cauza unui virus…* instead of *Dacă este cauzată de un virus…*,
- issues of word order and fluency also arise when MT preserves source-language structure, requiring reordering for clarity: *vocea lui Nick i-a plăcut* instead of *i-a plăcut vocea lui Nick*.

An important subcategory concerns the handling of NEs and linguistic conventions, including capitalization (*primul război mondial* instead of *Primul Război Mondial*), adaptation of proper names and titles, and localization of acronyms and institutional names (*Regional Internet Registry* translated as *Registru Regional al Internetului*).

Overall, the essential difference lies in the fact that MT typically delivers translations that are semantically correct, but structurally dependent on the source language, whereas

human translators intervene along lexical, syntactic, and stylistic dimensions to produce coherent, natural texts fully aligned with the norms of the target language.

3.2 Tokenisation

Automatic tokenisation was carried out with UDPipe[3] (Straka, Hajic, and Straková 2016) using a Romanian model[4] based on the Romanian Reference Treebank[5] (RRT) (Barbu Mititelu and Irimia 2016). The resulting tokenised sentences were then distributed for manual correction in a shared Google Sheets document, accompanied by a concise set of guidelines outlining annotation instructions common to all participating languages.

The guidelines emphasized the importance of accurate tokenisation as a foundation for all subsequent layers of annotation, including lemmatization, POS tagging, UD parsing, MWE annotation, NE annotation, and WSD. At this stage, each language team must decide how to treat multiword compounds at the WSD annotation level, more specifically, whether to WSD-annotate the individual tokens within a compound. If disambiguation of the elements is decided, compounds should be treated as multiword tokens and tokenised accordingly. In the case of Romanian, we identified only one situation requiring such token-level disambiguation, namely ad-hoc compounds (see Table 1).

Tokenisation correction is performed by adding or deleting lines in the Google Spreadsheet file, with due attention to the "SpaceAfter=No" attribute in the last column (see last column of Table 1), which indicates that a token is not followed by a whitespace in the original sentence. This attribute is essential, as it allows the reconstruction of the original, untokenized text by signaling where no space should be inserted.

Guidelines specific to the Romanian language and based on tokenization conventions used in previous Romanian linguistic resources developed in the UD framework were designed. As mentioned, we decided that the only compounds treated as multi-token words and to be disambiguated at token level are the ad-hoc, context specific compounds, which are not established dictionary entries (and therefore do not have an individual WSD label), but are formed productively as needed and typically exhibit compositional meaning: e.g., *sud-vest* "south-west": a directional compound formed as needed; *sud-dunărean* "south-Danubian": created to specify a geographic relation; *ruso-turc* "Russo-Turkish": combines two national/ethnic adjectives for a specific context. In such cases, the ELEXIS guidelines recommend splitting the compound while preserving compound information by specifying its span (see tokens 19–21 in Table 1).

Other hyphenated sequences are treated case by case, as follows:

- Compounds recorded as words (not locutions) in Romanian dictionaries are treated as one single token: e.g. *prim-ministru* "prime minister", *nou-născut* "newly born", *fast-food*, *mass-media*, *an-lumină* "light-year", *auto-raportat* "self-reported", *așa-numitul* 'so called' "the so called", etc.;

3 https://github.com/ufal/udpipe
4 romanian-rrt-ud-2.12-230717
5 https://universaldependencies.org/treebanks/ro_rrt/index.html

19-21	sud-vest	_	_	_	_	_	_	_	_
19	sud	_	_	_	_	_	_	_	SpaceAfter=No
20	-	_	_	_	_	_	_	_	SpaceAfter=No
21	vest	_	_	_	_	_	_	_	_

Table 1
Multi-token word treatment in ELEXIS

- Names of places are also treated as one token: e.g. *Nagorno-Karabah*, *Cluj-Napoca*;
- Semi-contextual compounds (whose tokens have independent meaning but the compound acquires a specific meaning of its own in specific contexts) are split: e.g., *Dunning-Kruger* → Dunning, -, Kruger , *unu-la-mai (mulți)* 'one-to-more' → unu, -, la, -, mai, (, mulți,);
- Opaque uses of hyphens: whenever it is not clear what the role of the hyphen is, we treat the elements it links as different tokens: e.g., *F-35B Lightning II* → F-35B, Lightning, II;
- Hyphen as inflection marker: split the inflexion from the root: *live-ului* "to the live" → live, -ului;
- Intervals marked by hyphen are split: e.g., *1521–26* → 1521, –, 26.
- Foreign compound words are also split: *Ready-to-wear* → ready, -, to, -, wear.

In Romanian, the hyphen plays an important role in marking contractions, resulting from cliticization and vowel elision. In such structures, the hyphen signals the phonological dependency of reduced functional elements (e.g., pronouns, auxiliaries) on adjacent lexical hosts. All contractions are split in the tokenisation process.

In structures involving pronouns, the spliting attaches the hyphen to the pronouns, which is the one affected by the vowel ellision:

- Clitic pronouns attached to verbs: *dă-mi cartea* "give me the book" -> dă, -mi;
- Auxiliary + pronoun contractions: *l-am văzut* 'him-have seen' "(I) have seen him";
- Clitic combinations: *mi-l dai* 'me-him give' "you give it to me" ⟶ mi, -l, dai; *ți-o dau* 'you-her give' "I give it to you" ⟶ ți,-o, dau.

Vowel ellision contractions can involve other functional elements like prepositions (*printr-o cunoștință* 'through-a acquaintance' "through a known person"), negation (*n-a fost* 'not-has been' "It hasn't been"), conjunctions (*c-ai venit* 'that-have come' "that you have come"). In all cases of vowel elision, the hyphen stays with the word from which the respective vowel dropped.

There are contractions when the hyphen does not replace any missing sound, like the gerund+clitic constructions, in which case the hyphen stays with the word that is not found in that form outside the respective construction: *Adaptându-se* → adaptându-, se (*adaptându* cannot occur but with the hyphen, it is not an independent form; its corresponding independent form is *adaptând*.

There are contextual special cases with specific rules when the principle of consistency is applied:

- In *văzându-și* 'seeing-Refl.', none of the words can occur independently, but the hyphen is consistently attached to the reflexive pronoun *și* and we attach it similarly in such cases for consistency.
- In *de-a lungul* → de-, a, lungul: *de-* and *a* are both prepositions and both can occur independently, but the hyphen is consistently attached at the end of a preposition (as opposed to the front) and we attach it similarly for consistency.

The apostrophe can occur in the following situation:

- in consonant elision, when it remains attached to the word (*domnu' Ionescu* "mister Ionescu" the apostrophe marks the elision of "l").
- in foreign words or sequences: all the token in the sequence are split, except the ones attached by apostrophe: S*gt. Pepper's Lonely Hearts Club Band* → Sgt., Pepper's, Lonely, Hearts, Club, Band; *don't* → don't.

Other specific rules were designed for:

- Symbols and formulas:
  - All terms of mathematical formulas are split: e.g., 76% → 76, %; *-100* → -, 100; *(3+5)*5/8=5* → (, 3, +, 5,), *, 5, /, 8, =, 5.
  - chemical compounds and formulas are kept as one token: e.g. *-NO2* (keep the polarity +/- together with the formula), *HLA-DRB1*11*.
  - Currency symbols are split from values: e.g., *3$* → 3, $
- Units of measure: the units are split from the values but kept as one token: e.g., *100 km/h* → 100, km/h.
- Slashes are always treated as individual tokens: e.g., *NOS/NTR* → NOS, /, NTR; *și/sau* "and/or" → și, /, sau
- Domain specific terms are split: e.g. *Haemophilus influenzae tip B* → Haemophilus, influenzae, tip, B; *F-35B Lightning II* → F-35B, Lightning, II.
- NEs are split: e.g. *NGC 205* → NGC, 205; *Apollo 12* → Apollo, 12.
- Abbreviations are treated as one token: e.g., *FC* → FC; *d.Hr.* → d.Hr.
- IT terms (written without blanks) are kept as one token: e.g. *W32.MyDoom@mm*, *www.voluntim.go.ro*.

3.3 Lemmatization and morpho-syntactic annotation

Annotation with parts of speech, with their features and the syntactic (i.e. dependency) relations between them was also done automatically, with UDPipe (Straka 2018), trained on the model romanian-rrt-ud-2.12-230717 based on the Romanian Reference Treebank (Barbu Mititelu 2018). The resulting CONLL-U files were converted to TSV3 format compatible with INCEpTION (Klie et al. 2018) (version 23.1). Based on a user account, the morphologic annotation and lemmatization were manually checked and corrected.

A number of recurrent annotation errors were identified during the processing of the Romanian data, many of which were caused by homonymy across PoSes. Such ambiguities frequently affected automatic tagging and lemmatization, requiring subsequent manual correction. One common source of error involved the ambiguity between adjectives and participles, particularly in forms that are morphologically identical but functionally distinct depending on context. Examples include *ridicată* "lifted/raised" and *utilizate* "used", which may function either as adjectival modifiers or as past participles within verbal constructions. Without sufficient contextual disambiguation, automatic tools often assigned inconsistent or incorrect tags.

Another frequent issue involved homonymy between proper nouns and common nouns, especially when common nouns appear in geographical or astronomical contexts and require capitalization. For instance, *pământ* may refer either to "ground" or to the planet "Earth", *lună* may denote "month" or "Moon", and *damasc* may refer either to the textile "damask" or to the proper noun "Damascus" (a metonymy). Automatic annotation often fail to capture these distinctions, particularly when capitalization conventions or contextual cues are subtle.

Ambiguity between adverbs and nouns also generated annotation inconsistencies. Forms such as *seara* "in the evening"/"the evening" or *vara* "in the summer"/"summer" may function either as temporal adverbs or as nominal expressions. Automatic tagging frequently misclassified these items, especially in contexts where syntactic cues were limited.

Homonymy within the same part of speech but involving different grammatical properties also posed challenges. For example, Romanian verbs may display identical forms in both present and imperfect tenses, such as *bea* "he drinks" / "he was drinking" and *beau* "they drink" / "they were drinking". Similarly, certain clitic pronouns share identical forms across grammatical cases, including accusative and dative, as in *le*, *îi*, *i-*, and *-le*, which complicated automatic morphological disambiguation. Another source of ambiguity involved nouns with identical forms in singular oblique case and plural direct case, such as *companii* "companies" and *soluții* "solutions", leading to inconsistent morphological analysis.

Abbreviations represented an additional challenge, as most were initially identified as proper nouns by the automatic annotation tools and subsequently corrected manually. Examples include *IT*, *MTV*, *TBC*, and *EBBA*, which required contextual interpretation to assign appropriate part-of-speech and entity labels.

Lemmatisation also revealed several systematic difficulties. One recurrent issue concerned neuter nouns appearing in plural form, which were incorrectly lemmatised with a feminine singular ending *-ă*. Examples include *tarifă* instead of *tarif*, *crateră* instead of

crater, and *metală* instead of *metal*. Another frequent error involved plural forms ending in *-i*, which were mistakenly interpreted as plural forms in *-uri* and truncated during lemmatisation. For instance, *armuri* was lemmatised as *arm* instead of *armură*, and *naturi* as *nat* instead of *natură*. These errors highlight the limitations of automatic lemmatisation for Romanian, particularly in the presence of morphological ambiguity and irregular inflectional patterns, and underscore the importance of manual validation in ensuring annotation accuracy.

3.4 Semantic annotations: MWEs

A further annotation level of the corpus is the semantic one. At the moment of this writing the corpus was automatically annotated with MWEs. These (i) are word combinations made up of at least two lexicalized components, (ii) have a neutral form which represents a weakly connected graph, and (iii) display some orthographic, morphological, syntactic and/or semantic idiosyncrasy with respect to what is considered general grammar rules of the respective language[6].

The types of MWEs currently represented in the Romanian corpus follow the PARSEME guidelines 2.0 and are as follows:

- Verbal MWEs:
  - Verbal Idioms (VID): *avea în vizor* (‘have in sight’ “keep in view, target”), *da nas în nas* (‘give nose in nose’ “run into, bump into”)
  - Light Verb Constructions (LVC)
    full: *avea întâlnire* (‘have meeting’ “have a meeting”), *da citire* (‘give reading’ “read”)
    cause: *da asigurare* (‘give assurance’ “assure”), *pune în aplicare* (‘put in application’ “implement, carry out”)
  - Reflexive Verbs (IRV): *-și permite* (‘3SG.DAT.REFL allow’ “afford”), *se abține* (“refrain”)
  - Inherently Adpositional Verbs (IAV): *se baza pe* (‘3SG.ACC.REFL base on’ “rely on, be based on”), *depinde de* (“depend on”)
- Nominal MWEs:
  - Nominal Idioms (NID): *act de identitate* (‘act of identity’ “identity card”), *an de grație* (‘year of grace’ “year of Our Lord, A.D.”)
  - Pronominal Idioms (PronID): *câte ceva* (“something”), *Domnia Sa* (‘His/Her Lordship’ “His Excellency”)
  - Deverbal Nominal MWEs (NV): *punere în funcțiune* (‘putting in function’ “commissioning, putting into operation”), *băgare de seamă* (‘putting of notice’ “attention, observation”)
- Modifier MWEs:

6 We use the notion of MWE as defined within PARSEME guidelines: https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/?page=010_Definitions_and_scope/020_Multiword_expressions, last accessed April 20, 2026.

  - Adjectival Idioms (AdjID): *cu o falcă în cer și cu una în pământ* (‘with one jaw in sky and one in earth’ “very furious”), *cu cântec* (‘with song’ “with flair, extravagantly”)
  - Adverbial Idioms (AdvID): *așa cum se cuvine* (‘so as 3SG.ACC.REFL befi’ “properly, as it should be”), *ca oamenii* (‘like people.THE.PL’ “like decent people, properly”)
  - Deverbal adjectival / adverbial MWEs (AV): *cu luare aminte* (‘with taking notice’ “attentively”), *luat în seamă* (‘taken in notice’ “taken into account, considered”)
- Functional MWEs:
  - Determiner Idioms (DetID): *picior de* (‘foot of’ “not a single”), *ca atare* (“as such”)
  - Adposition Idioms (AdpID): *de dragul* (‘of dear.DEF.SG.M’ “for the sake of”), *cu excepția* (‘with the exception’ “except for”)
  - Conjunction Idioms (ConjID): *astfel încât* (“so that”), *pe motiv că* (‘on reason that’ “because, on the grounds that”)
  - Interjection Idioms (IntjID): *ce folos* (‘what use’ “what’s the point?”), *nici vorbă* (‘no word’ “no way, not a chance”)

Automatic MWE annotation was carried out using a lexicon of MWEs extracted from the previously annotated and manually validated PARSEME-Ro corpus (Barbu Mititelu et al. 2025). A total of 7,268 unique surface-form occurrences of MWEs, together with their corresponding PARSEME labels, were automatically retrieved from the corpus and subsequently matched against ELEXIS-Ro. The matching procedure allowed for intervening tokens between the components of a MWE in order to account for discontinuous realizations in context. To balance the trade-off between permitting gaps (which generate a large number of candidates requiring subsequent manual filtering) and prohibiting them (which results in lower recall and increases the need for manual annotation), we opted to allow up to two intervening tokens – a compromise intended to capture most long-distance dependencies while keeping noise at a manageable level. A total of 3,018 MWEs were automatically identified in ELEXIS-Ro, and their distribution according to PARSEME labels is presented in Table 3.

4. Statistics on the ELEXIS-Ro corpus

The corpus has 2024 sentences containing 35,316 words and the average sentence length is 17.45 words.

The distribution of tokens according to their part of speech can be seen in Table 2. This show the dominance of nominal structures (see the number of nouns, proper nouns, adjectives and pronouns in the corpus). Sentences are not too complex, as the average number of verbs per sentence is 1.5, while the number of subordinate conjunctions is also low (234).

PoS	Number	PoS	Number
NOUN	8413	PRON	1187
ADP	4924	NUM	1178
PUNCT	4110	CCONJ	1019
VERB	3109	PART	484
ADJ	2653	SCONJ	234
AUX	2372	X	222
DET	2322	INTJ	1
PROPN	1750	SYM	1
ADV	1337	**TOTAL**	**35316**

Table 2
POS distribution in ELEXIS-Ro. The meaning of the abbreviation is as follows: ADP = prepositions, PUNCT = punctuation, ADJ = adjective, AUX = auxiliary verb, DET = determiner, PROPN = proper noun, ADV = adverb, PRON = pronoun, NUM = numeral, CCONJ = coordinating conjunction, PART = particle, SCONJ = subordinating conjunction, X = other, a residual class, INTJ = interjection, SYM = symbol

MWE type	Number	MWE type	Number
VID	79	NV.LVC.cause	4
LVC.full	18	AdjID	347
LVC.cause	2	AdvID	845
IRV	177	AV.VID	6
IAV	252	AV.IAV	30
NID	93	DetID	17
PronID	73	AdpID	863
NV.VID	1	ConjID	206
NV.LVC.full	1	IntjID	4
	TOTAL	**3018**	

Table 3
Number of MWEs by label in ELEXIS-Ro

5. Conclusions and future work

This paper has presented the extension of the ELEXIS multilingual corpus through the addition of Romanian as a new language component. The integration process was carried out primarily through automatic methods, followed by systematic manual validation to ensure annotation quality and cross-linguistic consistency.

At the current stage, manual revision has been completed for the translation, tokenisation, and morphological annotation layers. Observations regarding the types of errors encountered, the challenges specific to Romanian, and the strategies adopted for correction have been discussed in detail. These findings contribute to improving both the Romanian

component and the overall multilingual consistency of the ELEXIS corpus, while also providing insights into the interaction between automatic processing and manual validation in multilingual corpus development.

A key direction for future work concerns the manual validation and refinement of the automatically annotated MWEs in the Romanian component of the ELEXIS corpus. Although the current annotation follows the PARSEME 2.0 guidelines and relies on automatic identification procedures, previous studies have shown that such approaches inevitably introduce both false positives and false negatives, due to the structural variability and context-dependent interpretation of MWEs (Mititelu et al. 2026). In line with established methodologies, we plan to perform a systematic validation of the automatically extracted and annotated MWEs, focusing on verifying their boundaries, eliminating spurious candidates, and ensuring the correct assignment of MWE types. Particular attention will be paid to ambiguous cases, homonymous expressions, and constructions whose MWE status depends on contextual interpretation.

The validation process will involve multiple trained annotators and will follow a cross-validation protocol inspired by previous annotation campaigns. Each candidate MWE will be independently assessed by at least two annotators, applying the PARSEME decision tests and annotation guidelines. Inter-annotator agreement will be measured in order to evaluate annotation consistency, and disagreement cases will be further analysed through adjudication procedures and expert discussion until consensus is reached. In addition, the validation stage will include manual correction of the automatic annotation at corpus level, such as removing incorrect annotations, adjusting MWE boundaries, and identifying missing expressions not captured automatically. This step is expected to substantially improve the quality and reliability of the Romanian dataset and to provide a solid basis for subsequent layers of annotation and cross-linguistic comparison within the ELEXIS framework.

The syntactic annotation of the corpus will also undergo manual validation by trained linguists familiar with the UD annotation principles and dependency relations inventory.

Given the importance of the semantic aspect in the design and development of the ELEXIS corpus, annotation of words and MWEs with senses will also be done. We envisage transfer of senses via aligned wordnets from languages that use wordnets as sense repositories for their sense annotation. This will naturally also be followed by manual validation.

The Romanian component of the ELEXIS corpus is part of the release of the new ELEXIS version 2.0 and is available with a permissive license.[7]

Acknowledgements

This work was supported by a grant of the Ministry of Education and Research, CCCDI – UEFISCDI, project number PN-IV-PCB-RO-MD-2024-0142, within PNCDI IV. This work also received support from the CA21167 COST action UniDive, funded by the European Union via COST (European Cooperation in Science and Technology).

7 The corpus is available at https://vlo.clarin.eu/record/https_58__47__47_hdl.handle.net_47_11356_47_2101_64_format_61_cmdi?0.

References

Baisa, Vít, Jan Michelfeit, Marek Medveď, and Miloš Jakubíček. 2016. European Union language resources in Sketch Engine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia*, pages 2799–2803, European Language Resources Association (ELRA).

Barbu Mititelu, Verginica. 2018. Modern syntactic analysis of Romanian. In *Clasic şi modern în cercetarea filologică românească actuală*, pages 67–78, Iaşi, Romania.

Barbu Mititelu, Verginica, Ioana-Maria Biolan, Ioana Buhnila, Mihaela Cristescu, Elena Irimia, Catalin Mihaila, Isabella Sinca, Amalia Todirascu, and Carmen Vasile. 2025. Enhancing the PARSEME-Ro corpus with nominal, modifier and functional multiword expressions. In *Proceedings of the 20th International Conference on Linguistic Resources and Tools for Natural Language Processing*, pages 113–125.

Barbu Mititelu, Verginica and Elena Irimia. 2016. Linguistic data retrievable from a treebank. In *Proceedings of the Second International Conference on Computational Linguistics in Bulgaria (CLIB 2016), Sofia, Bulgaria*, pages 19–27, Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

Barbu Mititelu, Verginica, Dan Tufiş, and Elena Irimia. 2018. The reference corpus of the contemporary Romanian language (CoRoLa). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan*, European Language Resources Association (ELRA).

Ceauşu, Alexandru. 2008. Colectarea şi procesarea documentelor româneşti ale corpusului jrc-acquis (collecting and processing the Romanian documents of the JRC-Acquis corpus). In *Lucrările atelierului Resurse Lingvistice şi Instrumente pentru Prelucrarea Limbii Române (Proceedings of the Workshop Linguistic Resources and Tools for Processing the Romanian Language), Iaşi, Romania*.

Diewald, Nils, Michael Hanl, Eliza Margaretha, Joachim Bingel, Marc Kupietz, Piotr Bański, and Andreas Witt. 2016. KorAP architecture — diving in the deep sea of corpus data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia*, pages 3586–3591, European Language Resources Association (ELRA).

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. MIT Press.

Ion, Radu, Elena Irimia, Dan Ştefănescu, and Dan Tufiş. 2012. ROMBAC: The Romanian balanced annotated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey*, pages 339–344, European Language Resources Association (ELRA).

Irimia, Elena, Verginica Barbu Mititelu, Radu Ion, Vasile Păiş, Maria Mitrofan, and Dan Tufiş. 2026. CoRoLa version 2.0: Corpus Enrichment and a New Annotation Level. In *Proceedings of the 12th Workshop on the Challenges in the Management of Large Corpora*, Language Resources Association (ELRA).

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations (COLING 2018)*, pages 5–9, Association for Computational Linguistics.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X, Phuket, Thailand: Papers*, pages 79–86.

Krstev, Cvetana, Gordana Pavlović-Lažetić, Duško Vitas, and Ivan Obradović. 2004. Using textual and lexical resources in developing Serbian WordNet. *Romanian Journal of Information Science and Technology*, 7:147–161.

Krstev, Cvetana, Ranka Stanković, Aleksandra M. Marković, and Teodora Sofija Mihajlov. 2024. Towards the semantic annotation of SR-ELEXIS corpus: Insights into multiword expressions and named entities. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024, Torino, Italy*, pages 106–114, ELRA and ICCL.

Macoveiciuc, Monica and Adam Kilgarriff. 2010. The RoWaC Corpus and Romanian WordSketches. In *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian*. Romanian

Academy Publishing House, Bucharest, Romania, pages 151–168.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Martelli, Federico, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafael-J. Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, Simon László, Valeria Quochi, Monica Monachini, Francesca Frontini, Carole Tiberius, Rob Tempelaars, Rute Costa, Ana Salgado, Jaka Čibej, and Tina Munda. 2021. Designing the ELEXIS parallel sense-annotated dataset in 10 European languages. In *Proceedings of Electronic Lexicography in the 21st Century Conference*, pages 377–395, Lexical Computing CZ s.r.o.

Miller, George A. 1995. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41.

Mititelu, Verginica, Mihaela Cristescu, Elena Irimia, and Carmen Mîrzea Vasile. 2026. Two birds with one stone: Annotating Romanian multiword expressions with an eye to the PARSEME 2.0 guidelines applicability. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), Rabat, Marocco*, pages 66–74, Association for Computational Linguistics.

Pedersen, Bolette, Sanni Nimb, Sussi Olsen, Thomas Troelsgård, Ida Flörke, Jonas Jensen, and Henrik Lorentzen. 2023. The DA-ELEXIS corpus - a sense-annotated corpus for Danish with parallel annotations for nine European languages. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023), Tórshavn, the Faroe Islands*, pages 11–18, Association for Computational Linguistics.

Savary, Agata, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markantonatou, Nurit Melnik, Joakim Nivre, Atul Kr. Ojha, Carlos Ramisch, Abigail Walsh, Beata Wójtowicz, and Alina Wróblewska. 2024. UniDive: A COST action on universality, diversity and idiosyncrasy in language technology. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024, Torino, Italy*, pages 372–382, ELRA and ICCL.

Straka, Milan. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Brussels, Belgium*, pages 197–207, Association for Computational Linguistics.

Straka, Milan, Jan Hajic, and Jana Straková. 2016. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey*, pages 2214–2218, European Language Resources Association (ELRA).

Tufiş, Dan and Elena Irimia. 2006. RoCo-news: A hand validated journalistic corpus of Romanian. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy*, European Language Resources Association (ELRA).

Tatarstan Toponyms: A Bilingual Dataset and Hybrid RAG System for Geospatial Question Answering

Mullosharaf K. Arabov
Department of Data Analysis and Programming Technologies, Institute of Computational Mathematics and Information Technologies, Kazan (Volga Region) Federal University, Kazan, Russia
MKArabov@kpfu.ru

Svetlana S. Khaybullina
Department of Data Analysis and Programming Technologies, Institute of Computational Mathematics and Information Technologies, Kazan (Volga Region) Federal University, Kazan, Russia
khaybulinas@mail.ru

Dinara M. Naumetova
Department of Data Analysis and Programming Technologies, Institute of Computational Mathematics and Information Technologies, Kazan (Volga Region) Federal University, Kazan, Russia
ndinara03@gmail.com

This paper addresses the end-to-end task of automatic geospatial question answering over multilingual toponymic data. An original bilingual (Russian–Tatar) dataset of toponyms of the Republic of Tatarstan is introduced, comprising 9,688 structured records with detailed linguistic, etymological, administrative, and coordinate information (93.1% of objects are georeferenced). Based on this dataset, a specialized question-answering corpus of approximately 39,000 "question–context–extractable answer" triples is constructed, with guaranteed answer localization within the text. To solve the task, an architecture combining two key components is proposed: a hybrid retriever that integrates dense semantic indexing using multilingual-e5-large with a geospatial filter and ranking (KD-trees, haversine distance), and an extractive reader based on fine-tuned transformer models. On 500 test queries, the hybrid search achieves Recall@1 = 0.988, Recall@5 = 1.000, and MRR = 0.994, statistically significantly outperforming both BM25 and the purely spatial method. Among the tested reader architectures (RuBERT, XLM-RoBERTa-large, T5-RUS), the best answer extraction quality is attained by the multilingual XLM-RoBERTa-large model: EM = 0.992, F1 = 0.994. A contrasting effect is observed: on raw outputs, RuBERT-based models fail to answer coordinate-related questions (F1 = 0), whereas XLM-RoBERTa-large achieves F1 = 0.984; however, simple post-processing completely eliminates the gaps in numerical values and restores RuBERT accuracy to 100%. This discrepancy is attributed to tokenization specifics and the composition of pre-training corpora. The created resources (dataset, QA corpus, trained model weights, and web demonstrator) are

https://doi.org/10.47810/JCLIB.2.2026.03

openly published on the Hugging Face platform. The obtained results can be directly applied in the development of geospatial question-answering services, geocoding systems, and digital humanities projects that benefit from etymological and multilingual place-name data.

Keywords: Tatarstan toponyms, question answering, hybrid search, geospatial analysis, bilingual dataset, low-resource NLP

1. Introduction

Geographical names constitute a fundamental component of spatial data infrastructure and play a key role in cartography, navigation, regional governance, and digital humanities research (Suleymanov et al. 2021; Galimov, Burnashev, and Gatiatullin 2023). In multilingual regions such as the Republic of Tatarstan, where Russian and Tatar are both official state languages, toponyms function simultaneously in two linguistic codes and accumulate rich etymological, historical, and dialectological information. A notable feature of the local toponymy is the coexistence of several administrative categories of rural settlements: *selo* (село) and *derevnya* (деревня) are distinct types, the former historically larger and usually possessing a church, while the latter is a smaller settlement. Both are conventionally rendered as "village" in English, yet they represent different object types in the source data and are retained as separate classes in our dataset.

Traditional geospatial resources (GeoNames, OpenStreetMap) provide basic coordinate and classification data; however, they lack deep linguistic and etymological details and are not designed for semantic search that accommodates synonymy of geographical terms (e.g., "reka" – "pritok" [river – tributary]) and multilingual spelling variants of the same object (Rehurek and Sojka 2006). At the same time, etymological information, which is present in about 63% of our records, enables a qualitatively different kind of query: "Why is Lake Kaban so named?" or "What is the origin of the name Sarmanovo?" Such explanatory question answering is increasingly demanded by educational platforms, digital tourism services, and cultural heritage projects, yet it remains absent from conventional geospatial tools.

In parallel, the past decade has witnessed notable progress in the computational processing of the Tatar language; representative text corpora, morphological analyzers, tokenizers, and semantic annotation tools have been created (see Section 2.2 for a concise overview). Nevertheless, specialized question-answering datasets that account for toponymic specificity have been absent until now. Existing QA collections (SQuAD, SberQuAD, and their Russian-language counterparts) are oriented toward news and encyclopedic texts and do not cover structured geographic information that includes coordinates, object types, and etymological descriptions.

Modern generative models underlying Retrieval-Augmented Generation (RAG) require reliable retrieval components capable of supplying relevant context. In the domain of geographic search, the combination of dense semantic embeddings obtained from multilingual transformers (multilingual E5 (Wang et al. 2024), XLM-RoBERTa (Conneau et al. 2020)) with geospatial ranking and filtering is of particular interest. However, the integration of such approaches for multilingual toponymic data, especially coupled with subsequent extractive reading, remains underexplored. The identified gap – the simultaneous absence of both

a dataset and an end-to-end question-answering architecture for geographic queries in a bilingual setting – determines the relevance of the present work.

2. Related Work

2.1 Toponymic Data and Resources for the Tatar Language

International projects such as GeoNames and OpenStreetMap provide open coordinate and attribute information; however, they do not contain the detailed etymological and linguistic data necessary for scholarly onomastics (GeoNames 2025). Fundamental academic works on the toponymy of Tatarstan include the hydronym dictionaries and monographs by Garipova (Garipova 1984, 1990, 1991, 2010), Sattarov's comprehensive study of Tatar toponymy (Sattarov 1998), the two-volume etymological dictionary of the Tatar language by Äkhmätÿanov (Äkhmätÿanov 2015), the multi-volume Tatar Encyclopedia (Khasanov 2002–2014), and the large dialectological dictionary (Bayazitova et al. 2009). These sources, which have been partially digitized on the "Toponyms of Tatarstan" portal, provide the authoritative foundation for the dataset created in the present work (see Section 4.1).

2.2 Tools for Automatic Processing of the Tatar Language

For the Tatar language, general-purpose corpora (Saykhunov, Khusainov, and Ibragimov 2025; Sketch Engine 2024; IPSAN 2024), morphological analysis systems (Gilmullin and Gataullin 2017), tokenizers (Arabov 2026b), as well as methods for semantic annotation based on knowledge graphs (Mukhamedshin et al. 2025) and machine learning (Mindubaev and Gatiatullin 2024), have been developed. Software solutions for constructing vector representations of words and documents have been created (Arabov 2026a; Gafarov, Gafarova, and Ayupov 2025). All of the listed resources are consolidated on the Hugging Face platform within the TatarNLPWorld organization, ensuring reproducibility of experiments (TatarNLPWorld 2025c).

2.3 Semantic and Geospatial Search

Classical lexical methods such as BM25 (Rehurek and Sojka 2006) struggle with cross-lingual and synonymic queries. The development of Transformer architectures has led to the emergence of dense retrieval models. The E5 family (Wang et al. 2024) is trained with a contrastive loss function on multilingual collections and is capable of encoding documents and queries into a unified semantic space. The multilingual-e5-large model demonstrates high effectiveness for low-resource and multilingual scenarios (Conneau et al. 2020; Schweter 2020). In parallel, geoinformatics employs spatial indices (KD-trees (Bentley 1975), R-trees) and metrics that account for the Earth's curvature (haversine distance (Sinnott 1984)). Works (Galimov, Burnashev, and Gatiatullin 2023; Burnashev, Gatiatullin, and Khamidullin 2025) proposed geolinguistic systems combining fuzzy logic with spatial data for dialect analysis; however, they did not utilize dense embeddings. Hybrid approaches that combine

textual and geographic relevance in a weighted manner have been successfully applied in recommendation services and point-of-interest search, but such solutions have not been previously investigated for multilingual toponyms with detailed etymological attribution.

2.4 Question-Answering Systems and Extractive Reading

For the extractive QA task, SQuAD (Rajpurkar et al. 2016) and its multilingual versions have become the de facto standard datasets. Russian-language counterparts, such as SberQuAD (Efimov et al. 2020), are oriented primarily toward news and encyclopedic texts and do not contain structured geographic data with coordinates. Broader diagnostic benchmarks for the Russian language, in particular Russian SuperGLUE (Fenogenova et al. 2021), enable the evaluation of multiple aspects of language understanding but also do not include tasks requiring the extraction of numerical coordinates and etymological information. Among models, BERT-based architectures lead the field: RuBERT (Devlin et al. 2019) (further pre-trained on Russian texts) and the multilingual XLM-RoBERTa (Conneau et al. 2020), trained on one hundred languages. The latter has demonstrated excellent results in recognizing nested named entities, including geographical ones, as confirmed, for example, within the Russian-language competition RuNNE-2022 (Artemova et al. 2022). Generative T5 models (Raffel et al. 2020) are also applied, but typically in the answer generation setting. A significant limitation of monolingual models, revealed in recent experiments (Choure, Adhao, and Pachghare 2022), is their inability to process numerical coordinates due to WordPiece tokenization specifics and a shortage of relevant examples in training sets. Multilingual models employing SentencePiece, by contrast, successfully extract numerical sequences. To date, no specialized QA benchmarks combining structured geographic information with etymological and coordinate components have been proposed.

Thus, the literature review attests to the existence of disparate components (language resources for Tatar, dense retrieval models, geospatial filtering methods, extractive reading models); however, their integration into a unified system capable of finding the required toponym and extracting the precise answer for an arbitrary geographic query in Russian or Tatar is lacking. The present work is aimed at filling this gap.

3. Methodology

3.1 Hybrid Retrieval: Architecture and Method

The purpose of the retrieval component is to take a textual query q, optionally accompanied by a geographic point (lat_q, lon_q) and a radius R, and produce a ranked list of the k most relevant toponyms from a set of documents D, each of which is equipped with coordinates. The architecture comprises two parallel indexing stages – semantic and spatial – which interact during query processing in a hybrid ranking procedure.

Semantic Indexing. To represent the content portion of a document, the multilingual encoder model `intfloat/multilingual-e5-large` (Wang et al. 2024) is employed, which transforms

the contextual field `context` into a 1024-dimensional dense vector. The model was chosen as it demonstrates state-of-the-art results in cross-lingual search and supports both Russian and Tatar languages. All document embeddings are normalized to unit L_2-norm, allowing cosine proximity to be assessed via the dot product. Computation is performed in batches of 32 documents using a GPU. For efficient nearest-neighbor search, a flat inner product index (`IndexFlatIP`) is constructed using the FAISS library (Johnson, Douze, and Jégou 2019); when a GPU accelerator is available, the index is placed in GPU memory. Query processing reduces to encoding its text with the same model, L_2-normalizing the query vector, and extracting the top-k documents with the maximum dot product. The resulting quantity

$$sem\_score(i) = \langle e_q, e_i \rangle$$

is interpreted as the semantic relevance of document i.

Geospatial Filter and Ranking. Spatial selection is implemented in two passes. In the first stage, a bounding box is constructed around the query point:

$$\Delta lat = R/111320, \quad \Delta lon = R/(111320 \cdot \cos(\phi)),$$

where ϕ is the latitude in radians and R is the search radius. Objects whose coordinates fall within this region are declared candidates. Then, for each candidate, the exact haversine distance (Sinnott 1984) is computed, and those objects whose distance exceeds R are filtered out. The spatial score of the remaining candidates is determined by an exponentially decaying function of distance:

$$geo\_score(i) = e^{-\frac{d_i}{R}}$$

To accelerate the filtering stage, a KD-tree (Bentley 1975) is built over the coordinates of all documents during corpus preparation, enabling logarithmic-time retrieval of the subset of objects within a given neighborhood.

Combined Relevance. The final relevance is formed as a weighted sum of normalized semantic and spatial scores. Semantic scores are normalized using min-max scaling among the candidates that passed the geofilter:

$$sem\_norm(i) = \frac{sem\_score(i) - \min_{sem}}{\max_{sem} - \min_{sem}},$$

and in degenerate cases (when the difference is close to zero), all values are set to 1. Spatial scores are normalized by dividing by the maximum value among the candidates:

$$geo\_norm(i) = \frac{geo\_score(i)}{\max_{geo}}.$$

The combined score is computed as

$$\mathit{score}(i) = \alpha \cdot \mathit{sem\_norm}(i) + (1 - \alpha) \cdot \mathit{geo\_norm}(i),$$

where $\alpha \in [0, 1]$ is a hyperparameter governing the contribution of the semantic component. Objects are ranked by descending $\mathit{score}(i)$, and the top-k are returned. The value of α is tuned empirically on a validation set (200 queries) via grid search over $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ by maximizing the average Recall@5. In the experiments, $\alpha = 0.1$ was found to be optimal, with a fixed radius of $R = 50$ km, chosen based on the typical scale of local search within Tatarstan.

3.2 Extractive Reading: QA Corpus Generation and Model Training

For the extractive reading component, the original toponym dataset is transformed into a set of extractive question-answer pairs with precise answer spans. The procedure, described in full in Section 4.2, guarantees that every answer appears verbatim within the context and covers all major information categories (object type, location, etymology, coordinates, etc.). In total, 38,696 pairs were generated and split into training (90 %) and validation (10 %) sets, stratified by question type.

Models and Training. Two groups of transformer architectures were fine-tuned for the reading task: monolingual Russian-language RuBERT models (base and large versions, initialized from SberQuAD checkpoints (Efimov et al. 2020; Devlin et al. 2019)) and the multilingual XLM-RoBERTa-large model (Conneau et al. 2020) pre-trained on SQuAD 2.0. The generative model T5-RUS (Raffel et al. 2020) was used without fine-tuning as a reference point due to technical incompatibilities. All models were trained on the generated corpus for three epochs with the AdamW optimizer, a learning rate of 3×10^{-5}, batch size 4, and linear warm-up over 500 steps. Extractive models predict start and end positions of the answer span; T5-RUS was prompted in the format "question: ... context: ... → answer". A simple heuristic baseline that extracts the answer based on question keywords and field prefixes was also implemented.

The trained models, together with the QA corpus, are published on Hugging Face (TatarNLPWorld 2025b) to ensure reproducibility.

4. Data

4.1 Source Dataset of Tatarstan Toponyms

The empirical foundation of the study is a specialized bilingual (Russian–Tatar) dataset of Tatarstan toponyms, collected, structured, and published by the authors in open access on the Hugging Face platform (TatarNLPWorld 2025a). The necessity of this resource stems from the absence of machine-readable datasets that simultaneously contain coordinate, linguistic, and etymological information about the region's geographical objects.

The dataset draws on fundamental academic works on the toponymy of Tatarstan: the hydronym dictionaries and monographs by Garipova (Garipova 1984, 1990, 1991, 2010), Sattarov's comprehensive study (Sattarov 1998), the two-volume etymological dictionary by Äkhmätÿanov (Äkhmätÿanov 2015), the multi-volume Tatar Encyclopedia (Khasanov 2002–2014), the large dialectological dictionary (Bayazitova et al. 2009), and materials from the digital portal "Toponyms of Tatarstan" (toponym.antat.ru). Reliance on peer-reviewed academic sources guarantees high reliability and completeness of the data.

The total volume of the dataset amounts to 9,688 records, each corresponding to a single geographical object within the Republic of Tatarstan and adjacent regions of compact Tatar settlement. The record structure includes the following fields: a unique identifier, source URL, toponym type (toponym or microtoponym), toponym subtype (oikonym, hydronym, oronym, or no type), geographical object (76 unique categories: village, selo,[1] river, lake, field, mountain, meadow, river tributary, glade, etc.), name in Russian, name in Tatar, federal subject, physiographic details, geographical location, name etymology, bibliographic sources, latitude and longitude coordinates (where available), and a map availability flag. Etymological information, of particular value for onomastic and historical-geographic research, is present in approximately 63 % of records.

A key characteristic of the dataset is the presence of coordinate referencing: 9,023 records (93.1 %) are furnished with latitude and longitude values, enabling their use in geospatial analysis and distance-based filtering. The remaining 665 records (6.9 %) without coordinates were excluded from the hybrid search experiments but were still used for generating QA pairs that do not require spatial information (etymology, sources).

The distribution of records by toponym subtype is presented in Table 1. The predominant share consists of oikonyms – names of populated places (villages, selos, settlements), which mirrors the settlement structure of the region and the thematic coverage of the sources used. Hydronyms (names of water bodies: rivers, lakes, streams) and oronyms (names of landforms: mountains, hills, ravines) together constitute about one quarter of the corpus. The "no type" category unites records for which the toponym subtype was not explicitly indicated in the sources, although they retain information about the geographical object and etymology.

Granularity by geographical object type (Table 2) reveals the most frequent categories. Populated places dominate: villages (36.1 %) and selos (20.6 %), which corresponds to the share of oikonyms in Table 1. Among natural objects, the most represented are meadows (8.3 %), rivers (7.2 %), fields (5.2 %), and mountains (4.1 %). About 10.7 % of records belong to other, less frequent categories, such as lakes, river tributaries, glades, springs, natural tracts, swamps, and ravines. The presence of 76 unique geographical object types ensures high entity diversity and allows testing search algorithms under a heterogeneous taxonomy.

The regional distribution reflects the geographic focus of the dataset: approximately 93 % of records concern objects within the Republic of Tatarstan, about 4 % belong to Tyumen Oblast (areas of compact settlement of Siberian Tatars), and the remaining 3 % are distributed

1 In Russian administrative classification, "village" (деревня) and "selo" (село) are distinct types of rural settlements; the latter is historically larger and usually possessed a church. Both are translated as "village" in English but are kept separate in the dataset to preserve the original distinction.

Toponym Subtype	Number of Records	Share, %
Oikonym	7000	72.2
Hydronym	1500	15.5
Oronym	800	8.3
No type	388	4.0
Total	**9688**	**100.0**

Table 1
Distribution of dataset records by toponym subtype

Geographical Object	Share, %
Village	36.1
Selo	20.6
Meadow	8.3
River	7.2
Field	5.2
Mountain	4.1
Lake	3.1
River tributary	2.6
Glade	2.1
Other	10.7
Total	**100.0**

Table 2
Top-10 most frequent geographical objects in the dataset

among adjacent subjects of the Russian Federation (Bashkortostan, Ulyanovsk, Samara, Orenburg Oblasts, and others).

To facilitate semantic search, all textual fields of each record, excluding coordinates, were merged into a single contextual field `context` using English-language key prefixes. The context format is as follows:

```
Name (rus): <name_rus> | Name (tat): <name_tat> | Type: <toponym_type> |
Subtype: <toponym_subtype> | Object: <geographical_object> | Etymology:
<etymology> | Details: <physiographic_details> | Location:
<geographical_location> | Sources: <bibliographic_sources>
```

Fields containing no information were excluded from the context. The choice of English-language prefixes is motivated by the use of the multilingual model `multilingual-e5-large`, for which English-language keys provide more robust cross-lingual matching of semantically similar fields. This procedure yielded 9,023 documents with non-empty

contextual fields and coordinate referencing, which formed the corpus for indexing and testing the retrieval component.

4.2 Generation of the Question-Answering Corpus

Based on the structured toponym dataset, a synthetic question-answering corpus was automatically generated for training and evaluating extractive QA models. The central requirement was that every answer be guaranteed to appear verbatim within the context string, enabling precise character-level annotation of the answer span.

For each information type in a source record (name, object type, location, etymology, coordinates, region, physiographic characteristics, sources), the corresponding field was transformed into a string with a Russian-language prefix that unambiguously identifies the category: `Name (Rus):`, `Name (Tat):`, `Object:`, `Etymology:`, `Location:`, `Coordinates:`, `Region:`, `Physiographic Details:`, `Sources:`. All prefixed strings were concatenated via the "|" separator into a single context. If the resulting string exceeded 2048 characters, each field was proportionally truncated while preserving its prefix. This design guarantees that any potential answer appears together with its identifying prefix, so the answer start position can be unambiguously computed as the sum of the lengths of the preceding fields and separators.

Russian-language question templates with a `{name}` placeholder were developed for seven information categories; the complete list is available in the dataset card on Hugging Face.

During generation, one template was randomly selected for each record, and the Russian name (or, if unavailable, the Tatar name) was inserted into the placeholder. The answer was taken directly from the corresponding field (for coordinates, a comma-separated latitude–longitude string). The answer start position was computed algorithmically as described above. Each resulting pair thus carries exact positional labels required for extractive model training.

Question Type	Number of Pairs	Share, %
Coordinates	12,344	31.9
Object type	8,032	20.8
Location	5,724	14.8
Etymology	5,032	13.0
Region	3,868	10.0
Sources	3,052	7.9
Physiographic characteristics	644	1.6
Total	**38,696**	**100**

Table 3
Distribution of the synthetic QA corpus by question type

The maximum number of QA pairs per record was capped at ten to avoid dominance of objects with many populated fields and to ensure uniform coverage. The total volume of the generated corpus is 38,696 pairs. Their distribution by question type is shown in Table 3.

Coordinate questions form the largest share (31.9 %), reflecting the primary importance of spatial information. Questions about object type (20.8 %) and location (14.8 %) together account for more than one third of the corpus. Etymological questions (13.0 %) exploit the dataset's unique feature – the availability of name-origin data. The smallest category, physiographic characteristics (1.6 %), corresponds to the fragmentary population of that field in the source records.

The average context length is 1250 characters; answers vary from 2 to 150 characters. Coordinate answers are the shortest (around 20 characters on average), while answers containing bibliographic sources can reach up to 500 characters. An illustrative span-annotated example is:

```
{
"id": "1530_coordinates_0",
"context": "Name (rus): Rantamak | Name (tat): Rantamak | Object: Selo |
Etymology: The toponym originates from the oikonym ``Rangazar-Tamak''. |
Location: Located on the Melya River, 21 km east of the village of Sarmanovo.
| Sources: R.G. Äkhmätÿanov, Etymological Dictionary of the Tatar Language...
| Coordinates: 55.205461, 52.881862",
"question": "What are the coordinates of Rantamak?",
"answers": [{"text": "55.205461, 52.881862", "answer_start": 312}]
}
```

The corpus was randomly split into training (90 %, 34,826 pairs) and validation (10 %, 3,870 pairs) sets while preserving the stratification by question type. Both parts, together with the source toponym dataset, are openly published on the Hugging Face platform (TatarNLPWorld 2025a,b) under the CC BY-SA 4.0 license, and are available for scholarly and applied use.

5. Experimental Study

The experimental validation of the proposed question-answering system was conducted in two stages, corresponding to its architectural components. In the first stage, the effectiveness of the hybrid retrieval module, combining semantic indexing with geospatial filtering, was evaluated; in the second, the ability of the extractive reading component to accurately localize the answer within the context provided by the retrieval mechanism was assessed. Such a two-part evaluation protocol made it possible, on the one hand, to characterize each subsystem in isolation, and on the other, to demonstrate their compatibility within a unified pipeline. All experiments were carried out on the corpus of 9,023 coordinate-referenced documents described above and on the synthetic question-answering set of 38,696 examples generated according to the procedure in Section 4.2.

To evaluate the retrieval component, an independent test set was prepared by randomly selecting 500 records from the source dataset and generating natural-language queries for

them using five templates (“What is {name}?”, “Where is {name}?”, “Tell about {name}”, etc.), in which the object name was inserted either in Russian or in Tatar with probabilities of 0.7 and 0.3, respectively. This approach guaranteed that for each query there exists exactly one relevant document – the source record from which the query was generated. Additionally, a validation set of 200 queries, disjoint from the test set, was allocated; it was used exclusively for tuning the hybrid search hyperparameters – the weighting coefficient α and the spatial filter radius R.

Retrieval quality was measured by two main metrics: Recall@k ($k = 1, 3, 5$) and Mean Reciprocal Rank (MRR). The former indicates the proportion of queries for which the relevant document appears among the top k retrieval results; the latter averages the reciprocal of the rank of the first relevant document and is thus sensitive to early hits. To obtain statistically sound conclusions, all metrics were accompanied by 95% confidence intervals constructed via bootstrap with 1,000 resamples with replacement. This enabled not only comparison of average values but also assessment of the significance of differences between methods.

Four ranking strategies were compared: classical lexical BM25 search, purely spatial search (top-k nearest by haversine distance), semantic search based on `multilingual-e5-large` with a FAISS index, and the proposed hybrid method. The hybrid search was performed with a fixed radius of $R = 50$ km and $\alpha = 0.1$, selected on the validation set as yielding the maximum Recall@5 $= 1.0$. The comparison results are summarized in Table 4.

Method	Recall@1	Recall@3	Recall@5	MRR
BM25	0.438 [0.394, 0.484]	0.574 [0.528, 0.620]	0.618 [0.572, 0.662]	0.508 [0.469, 0.545]
Spatial only	0.536 [0.492, 0.576]	0.708 [0.668, 0.748]	0.796 [0.760, 0.830]	0.634 [0.598, 0.673]
Semantic only	0.774 [0.736, 0.810]	0.904 [0.878, 0.928]	0.940 [0.918, 0.960]	0.840 [0.813, 0.866]
Hybrid ($\alpha = 0.1$, $R = 50$ km)	0.988 [0.978, 0.996]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.994 [0.988, 0.998]

Table 4
Comparison of search methods (95% bootstrap confidence intervals)

The figures in Table 4 indicate a substantial superiority of the hybrid scheme. First of all, the value Recall@5 $= 1.000$ is striking, meaning that for each of the 500 test queries, the relevant object is guaranteed to be present among the first five results. Such high recall practically eliminates the risk of losing the necessary information before the reading stage, which is a critical requirement for RAG systems. The Recall@1 $= 0.988$ score indicates that only in six cases out of five hundred did the target document fail to occupy the first position; moreover, the lower bound of the confidence interval (0.978) lies substantially above the upper bounds of Recall@1 for all competing methods. This arrangement of intervals confirms the statistical significance of the hybrid’s superiority over the alternatives.

Semantic search by itself demonstrates fairly high quality (Recall@1 $= 0.774$), confirming the ability of the `multilingual-e5-large` model to adequately encode multilin-

gual toponymic contexts and capture semantic proximity between different name variants. However, the absence of spatial filtering allows objects with similar names but located far apart to appear in top positions, which reduces accuracy by approximately 0.2 compared to the hybrid. Spatial search, which ignores the query text, shows Recall@1 $= 0.536$. This is expected: within a 50 km radius, several objects are often found, and the geographically nearest one does not always coincide with the sought one. Nevertheless, even this simple strategy outperforms BM25 (Recall@1 $= 0.438$), underscoring the fundamental role of geographic proximity for toponymic queries. BM25's most modest results are explained not only by the lexical gap between Tatar and Russian names but also by the synonymy of geographical terms: a classical inverted index is unable to equate "selo" and "village", "river" and "tributary".

To visually consolidate these comparative findings, Figure 1 presents a bar chart of Recall@1 and Recall@5 for each method, supplemented with 95% confidence intervals. This graphical representation not only confirms the numerical superiority of the hybrid method but also reveals a striking pattern: the hybrid method exhibits minimal variability in its error bars, indicating high stability and reproducibility of performance across different query instances. In contrast, the spatial and lexical methods show substantially wider confidence intervals, reflecting greater variance in their retrieval effectiveness. The clear separation between the hybrid confidence intervals and those of all competing methods provides visual confirmation of the statistical significance established earlier.

While the aggregate metrics presented in Table 4 and Figure 1 unequivocally establish the superiority of the hybrid approach over its constituents, the structure of the dataset allows for a more granular analysis. To understand whether the system performs uniformly across different types of geographical features, we next dissected the retrieval results by toponym category. This is important because toponyms vary greatly in their spatial density and linguistic representation. For instance, large populated places (oikonyms) appear frequently in the dataset with rich contexts, while microtoponyms such as meadows or springs have fewer surrounding landmarks and often rely on precise coordinate information. We report the Recall@1 breakdown directly: for toponyms (349 queries) it is 0.986, and for microtoponyms (151 queries) it is 0.993.

Both categories demonstrate near-ceiling values; however, microtoponyms show a slightly higher result. This difference is explained by their local nature: for small objects, the density of toponyms within a 50 km radius is generally lower, meaning the spatial filter leaves fewer candidates, and the semantic component finds it easier to identify the sole correct one. In contrast, for larger toponyms, the higher density of historical records and alternative names within the same radius increases the likelihood of semantic confusion, despite the overall high performance. This pattern suggests that the hybrid method scales gracefully to objects of varying granularity, maintaining near-perfect recall even in densely populated regions. Figure 2 visualizes this comparison, with the bar chart clearly illustrating the slight but systematic superiority of microtoponym retrieval.

Having established the near-perfect performance across both toponym categories, it is equally instructive to examine the rare instances where the retrieval system falls short. These borderline cases offer critical insights into the system's limitations and, more importantly, reveal concrete pathways for future data-centric improvements. Understanding

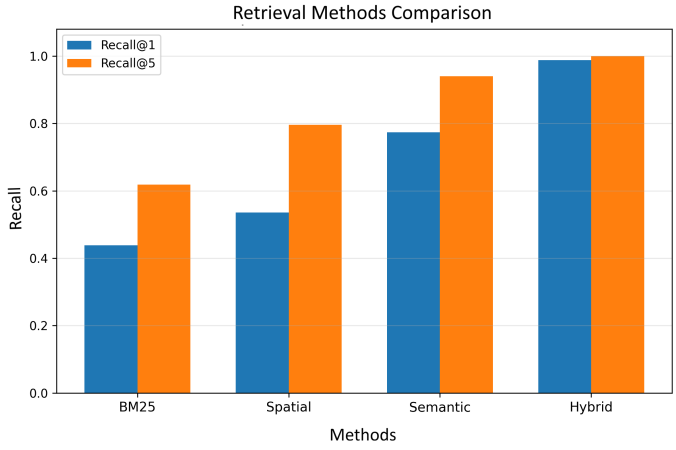


Figure 1
Comparison of Recall@1 and Recall@5 with 95% confidence intervals for BM25, purely spatial, semantic, and hybrid methods.

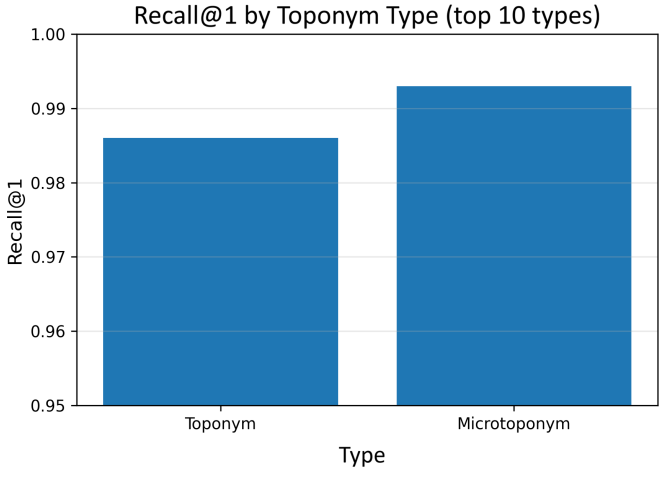


Figure 2
Recall@1 of the hybrid method for the “Toponym” and “Microtoponym” categories.

the nature of these failures allows us to separate algorithmic deficiencies from data quality issues.

Despite the impressive aggregate metrics, the hybrid retrieval made six errors in which the relevant document did not appear in the first position. Manual analysis of these cases, undertaken to identify root causes, showed that all of them are related not to shortcomings of the combination algorithm but to the quality of the data themselves. In two cases, com-

plete namesakes were present within a 50 km radius – two different objects with identical names (e.g., two natural tracts named “Krasnaya Gorka”). Their contexts proved practically identical, and the semantic component could not give preference to one over the other. In three cases, the coordinates of the target object were offset by 10–15 kilometres from the true position, due to which it did not fall within the search zone and was eliminated at the geofiltering stage. In one further case, the `context` field contained a minimum of information (only name and type), which yielded a predictably low semantic score and allowed another, more informative candidate to take the lead. Consequently, further improvement of the retrieval component lies primarily in enhancing data quality: verifying coordinates through cross-validation with OpenStreetMap or satellite imagery, enriching contextual fields, and developing a component for resolving toponymic homonymy.

The retrieval analysis confirms that the hybrid algorithm is highly robust and achieves near-ideal recall, with its remaining errors attributable to data quality rather than algorithmic shortcomings. With the search module validated, the next logical step is to evaluate the downstream component responsible for extracting precise answers from the retrieved contexts.

Turning to the second stage, extractive reading, we note that the task here was to precisely locate the answer within the context retrieved by the retrieval module. Three architectures, representing both monolingual and multilingual approaches, were selected for training and evaluation: RuBERT base, RuBERT large, and XLM-RoBERTa large, each fine-tuned on the synthetic QA corpus of 34,826 pairs. The generative model T5-RUS, unfortunately, could not be trained due to tokenizer version incompatibility and was used only in its base pre-trained variant as an additional reference point. Additionally, as a simple baseline, a heuristic rule was implemented that extracts the answer based on question keywords and the corresponding prefixes in the context. The training hyperparameters, common to all models, are presented in Table 5.

Parameter	**Value**
Maximum sequence length	384 tokens
Stride	128 tokens
Training batch size	4
Evaluation batch size	8
Learning rate	3×10^{-5}
Number of epochs	3
Weight decay	0.01
Warm-up steps	500
Optimizer	AdamW

Table 5
Hyperparameters for QA model fine-tuning

Following the fine-tuning protocol outlined in Table 5, all models were evaluated on the held-out validation set of 3,870 QA pairs. The results, summarized in Table 6, revealed a critical nuance regarding tokenizer-induced artifacts that necessitates a careful distinc-

tion between raw and normalized performance metrics. It turned out that the RuBERT models, operating with the WordPiece tokenizer, tend to insert extraneous spaces inside numerical coordinates (“55. 175195” instead of “55.175195”) and hyphenated constructions (“north - west”). This is a purely tokenization artifact unrelated to the semantic capability of the model. By applying elementary post-processing (removal of breaks within floating-point numbers and unification of hyphens and brackets), we ensured that RuBERT answers became identical to the reference ones. Table 6 presents both raw Exact Match and F1 metrics and the values obtained after normalization.

Model	EM (raw)	F1 (raw)	EM (norm)	F1 (norm)	Time (ms)
xlm_roberta_large	0.992	0.994	0.992	0.994	22.4
rubert_base	0.402	0.684	1.000	1.000	6.6
rubert_large	0.398	0.679	1.000	1.000	6.5
xlm_roberta_base (w/o fine-tun.)	0.440	0.574	–	–	22.4
rubert_base (w/o fine-tun.)	0.068	0.187	–	–	6.6
rule_based	0.492	0.492	–	–	≈0
t5_rus_base (w/o fine-tun.)	0.176	0.220	–	–	27.2

Table 6
QA model results on the validation set (raw and normalized metrics)

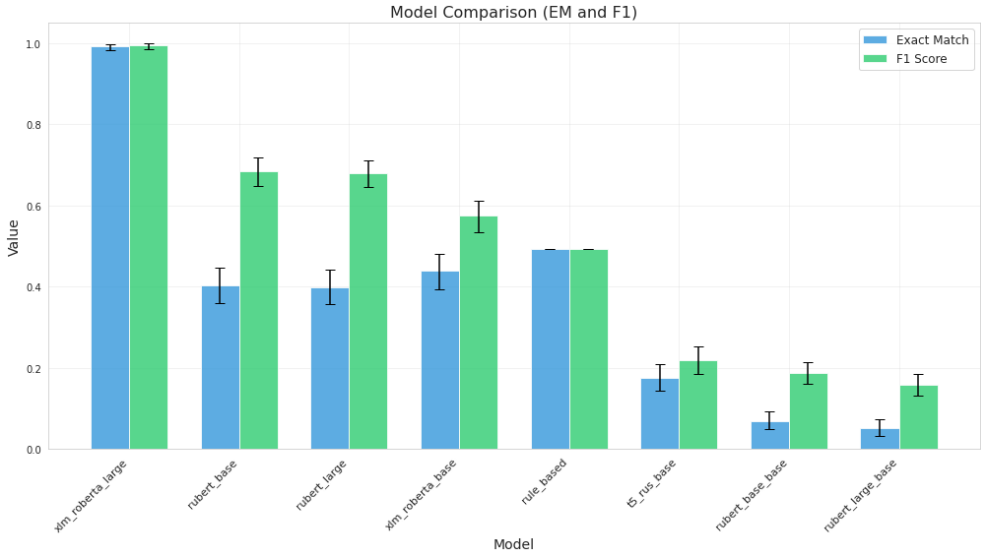


Figure 3
Comparison of models by Exact Match and F1 score after normalization (error bars represent 95% confidence intervals).

The multilingual XLM-RoBERTa large initially delivers virtually perfect answers (EM = 0.992) and requires no post-processing. Both RuBERT models, after minimal normalization,

achieve exact match with the reference (EM $= 1.000$), while operating 3.5 times faster – 6.5 ms versus 22.4 ms per query. This speed advantage makes RuBERT the preferred choice for production systems, provided a thin post-processing layer is added. Figure 3 presents a comparison of models in terms of EM and F1 with 95% bootstrap confidence intervals, providing a visual complement to the tabular data.

While the overall metrics presented in Figure 3 are encouraging, a nuanced understanding of the models' capabilities requires a deeper dive into their performance across different question categories. To this end, we conducted a fine-grained per-category quality analysis, presented in Table 7. After normalization, all three fine-tuned models achieve 100% F1 on questions about etymology, location, region, and object type. The only noticeable deviation is observed for XLM-RoBERTa large on coordinates (F1 $= 0.984$), which is associated with rare cases of inaccurate copying of the last digit of a decimal fraction. This minor degradation does not affect the RuBERT models, which achieve perfect scores across all categories after post-processing.

Model	Coordinates	Etymology	Location	Region	Sources	Object Type
xlm_roberta_large	0.984	1.000	1.000	1.000	0.988	1.000
rubert_base	1.000	1.000	1.000	1.000	1.000	1.000
rubert_large	1.000	1.000	1.000	1.000	1.000	1.000

Table 7
F1 score by question type after normalization

To provide an intuitive and comprehensive visualization of this per-category analysis, we present a heatmap in Figure 4. This representation immediately reveals the overall structure of model performance: all cells are coloured in the warmest tones, corresponding to an F1 of unity, with the single exception of the coordinates category for XLM-RoBERTa-large. The lighter cell visually confirms the minor but consistent deviation reported in Table 7. The heatmap not only facilitates rapid cross-model and cross-category comparison but also highlights the structural similarity of performance across RuBERT-based models, which exhibit uniformly perfect scores. The stark contrast between the uniformly dark rows for RuBERT and the slightly lighter cell for XLM-RoBERTa-large underscores the effectiveness of the post-processing strategy when applied to monolingual architectures.

The detailed per-category breakdown and its heatmap visualization confirm that, following the appropriate normalization, the RuBERT models deliver flawless performance across every question type. While these results validate the model's theoretical capability, they do not fully address the practical constraints of real-world deployment. Two additional dimensions are critical for assessing the readiness of the system for production use: the structural consistency of generated outputs and the computational efficiency of inference. The first dimension ensures that the model produces answers in a format that can be seamlessly integrated into downstream applications, while the second determines the feasibility of deploying the system in latency-sensitive environments such as real-time geospatial services or interactive digital archives. The final part of the experimental study addresses these

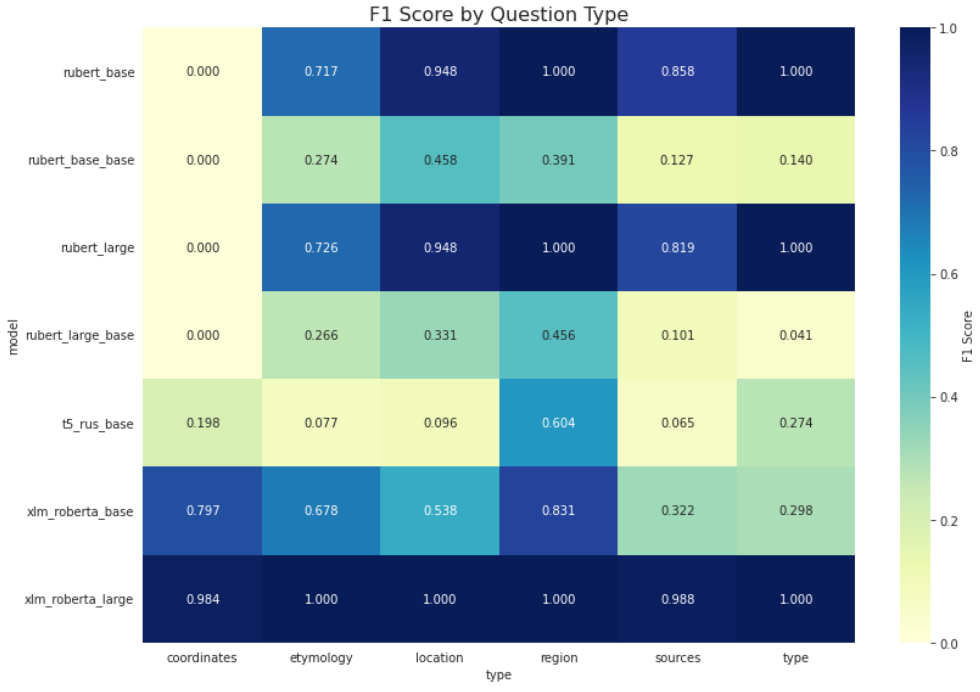


Figure 4
Heatmap of F1 score by question type for fine-tuned models (after normalization).

aspects through a systematic analysis of answer length distributions and computational latency, providing a comprehensive assessment of the system's practical deployability.

The distribution of predicted answer lengths for the XLM-RoBERTa large model is practically identical to the reference, as shown in Figure 5. The bulk of answers is concentrated in the 20–40 character range, corresponding to coordinates and short names; there is also a small peak around 200 characters, corresponding to bibliographic sources. For RuBERT after normalization, the distribution shape coincides with the reference without significant deviations, indicating that the post-processing step does not distort the natural answer length distribution.

While the high-quality answer extraction demonstrated in Figure 5 is encouraging, the practical applicability of the proposed system in real-world scenarios depends not only on accuracy but also on computational efficiency. To fully assess the trade-offs between precision and latency, we performed a detailed benchmark of inference times across all evaluated models. Speed measurements (Figure 6) show that RuBERT spends only about 6.5 ms per query, XLM-RoBERTa large spends 22.4 ms, and the generative T5 spends 27.2 ms. Thus, for tasks where latency is critical, RuBERT with post-processing appears to be the optimal alternative, whereas XLM-RoBERTa large may be preferable when maximum quality of numerical coordinate extraction is required without additional programming.

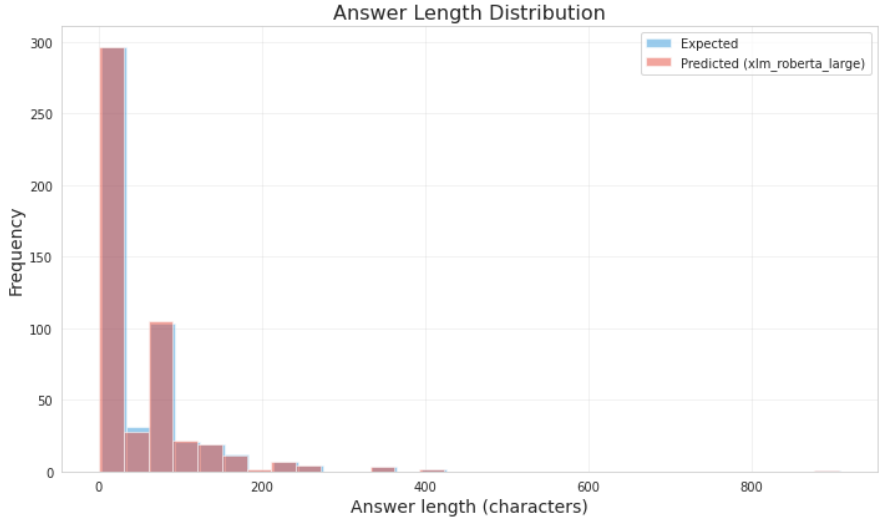


Figure 5
Distribution of answer length (in characters) for reference values and predictions of the xlm_roberta_large model.

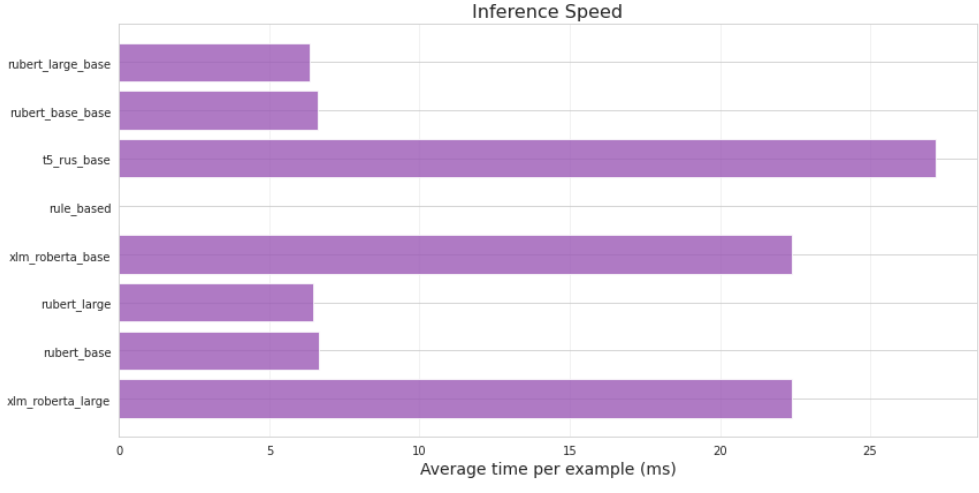


Figure 6
Average processing time per example (ms) for the considered QA models.

The contrast between the two figures is particularly informative. Figure 5 confirms that all models correctly capture the structural characteristics of the target answers, regardless of their length. This means that the accuracy gains obtained through fine-tuning are not achieved at the expense of generating malformed or structurally inconsistent outputs. The

models learn not only the content but also the natural distribution of answer lengths. In contrast, Figure 6 reveals a substantial divergence in computational overhead across architectures. The 3.5× speed advantage of RuBERT over XLM-RoBERTa-large, combined with its perfect accuracy after normalization, makes it a compelling choice for high-throughput applications. This finding challenges the common assumption that multilingual models are universally superior for multilingual tasks. In scenarios where post-processing is permissible, a carefully fine-tuned monolingual model may offer both superior speed and comparable accuracy.

In summary, these inference speed measurements provide a clear practical trade-off for system designers. The near-instantaneous inference of RuBERT, combined with simple post-processing, offers a highly efficient solution for large-scale or real-time applications, whereas the robust native output of XLM-RoBERTa-large is advantageous for tasks demanding maximum precision with minimal post-processing overhead.

Thus, the totality of the experimental data convincingly demonstrates that the proposed two-component architecture achieves near-ideal performance at all stages – from retrieval of the relevant document to precise extraction of the answer – and the identified tokenization peculiarities of RuBERT not only explain previous observations but also provide a simple recipe for their complete elimination.

6. Discussion

The experimental results presented above allow us to draw several interconnected conclusions about the developed system and, more broadly, about building question-answering pipelines over structured bilingual toponymic data. The most significant finding is that a lightweight hybrid architecture, coupling a state-of-the-art multilingual dense retriever with a straightforward spatial filter, achieves practically ideal recall for geographic queries at the scale of a single region. The fact that the optimal contribution of the semantic component turned out to be as low as $\alpha = 0.1$ quantitatively confirms an intuitive but previously unmeasured insight: for toponym search, geographic proximity is the dominant relevance signal. Once the search space is narrowed by a 50 km radius, even a modest semantic signal suffices to pinpoint the correct object.

This observation has direct consequences for the design of geospatial retrieval systems in low-resource multilingual settings. Rather than investing heavily in complex neural rankers, one can achieve near-perfect performance by concentrating on data quality—accurate coordinates, rich contextual descriptions, and a mechanism for resolving homonymy. The manual inspection of the six retrieval errors, which were all traced to coordinate inaccuracies, namesakes, or sparse contexts (see Section 5), reinforces this message: the algorithm itself is robust; further gains lie primarily in data curation.

The extractive reading experiments revealed an equally instructive, and somewhat counterintuitive, pattern. The apparent failure of monolingual RuBERT models on coordinate questions was not due to any semantic deficiency but solely to WordPiece tokenization artifacts that inserted spurious spaces into numerical strings. Trivial post-processing completely eliminated the issue, allowing both RuBERT variants to reach perfect accuracy while being 3.5 times faster than the multilingual XLM-RoBERTa-large. This result challenges the

widespread assumption that multilingual SentencePiece models are inherently necessary for handling numerical data. In practical deployments where a thin post-processing layer is acceptable, monolingual models can offer a superior speed–accuracy trade-off—a valuable lesson for NLP engineering in low-resource language contexts.

The high performance of the individual components suggests that an end-to-end RAG pipeline can be built with confidence, as the retrieval module guarantees that the correct context is supplied to the reader in virtually every case. Although a full end-to-end evaluation remains to be conducted, the near-ceiling metrics at both stages provide a strong upper-bound estimate of the overall system quality.

Several limitations must be kept in mind. The QA corpus and the test queries were generated from templates, which, while guaranteeing answer localization, do not fully reflect the variability of real user queries. Expanding the evaluation to include naturalistic questions is an important next step. The geospatial filter uses a fixed 50 km radius, which may be suboptimal for objects of very different spatial scales (e.g., a mountain range vs. a spring). Furthermore, physiographic characteristics constituted only 1.6% of the QA corpus, so conclusions about this category remain preliminary.

The open release of all resources—the bilingual toponymic dataset, the QA corpus, the fine-tuned models, and the web demonstrator—ensures reproducibility and invites follow-up work. The inclusion of etymological information, questioned by one reviewer, proves valuable for explanatory question answering, which is increasingly demanded by educational platforms, digital tourism services, and cultural heritage projects. In such applications, the ability to answer not only *where* but also *why* a place is named a certain way adds a qualitatively new dimension to geospatial exploration.

Future work will pursue adaptive radius selection, coordinate cross-validation against external sources (e.g., OpenStreetMap), and a full end-to-end RAG evaluation including generative models. The proposed methodology is portable to other multilingual regions, provided that toponymic datasets of comparable structure and coordinate coverage can be constructed.

7. Conclusion

This work has demonstrated that a carefully designed, yet conceptually simple, hybrid question-answering system can achieve near-perfect performance on bilingual geospatial queries over structured toponymic data. By combining multilingual dense embeddings with lightweight spatial filtering, and by pairing the resulting retriever with fine-tuned extractive readers, we obtained an end-to-end pipeline that locates the correct document and pinpoints the exact answer in almost every case. The key practical insights—the dominance of geographic proximity for toponym search and the speed advantage of monolingual models after tokenization-aware post-processing—extend beyond the immediate setting and can inform the design of similar systems for other low-resource languages.

The main scientific and practical contributions are as follows:

1. An openly available bilingual (Russian–Tatar) dataset of 9,688 toponyms with rich linguistic, etymological, and coordinate information, together with a

specialized extractive QA corpus of 38,696 span-annotated pairs, was created and published (TatarNLPWorld 2025a,b).

2. A hybrid retrieval method integrating dense semantic indexing with geospatial ranking was shown to achieve Recall@5 $= 1.000$ and MRR $= 0.994$, significantly outperforming BM25, purely semantic, and purely spatial alternatives.

3. A multi-architecture benchmark of extractive readers demonstrated that XLM-RoBERTa-large reaches EM $= 0.992$ without post-processing, while RuBERT models, after trivial normalization that compensates for WordPiece tokenization artifacts, achieve perfect accuracy with 3.5 times faster inference.

4. All resources—dataset, QA corpus, model weights, and an interactive web demo—are released on Hugging Face, guaranteeing full reproducibility and immediate applicability.

The developed system is ready for integration into geoinformation services, digital archives, educational platforms, and cultural heritage preservation projects. The near-perfect retrieval recall ensures that downstream generative models, when incorporated into a full RAG pipeline, will receive relevant context in virtually all cases, effectively eliminating hallucination risks originating from the retrieval stage.

Further research will address adaptive spatial filtering, coordinate verification against external sources, end-to-end RAG evaluation, and the expansion of the QA corpus with open-ended questions. The approach can be scaled to other bilingual regions, provided that toponymic datasets with comparable structure and coordinate referencing are available.

References

Arabov, Mullosharaf K. 2026a. Tatar2vec. Certificate of State Registration of Computer Program No. 2026610619 dated 14.01.2026. (In Russian).

Arabov, Mullosharaf K. 2026b. Tatartokenizers. Certificate of State Registration of Computer Program No. 2026611049 dated 16.01.2026. (In Russian).

Artemova, Ekaterina, Maxim Zmeev, Natalia Loukachevitch, Igor Rozhkov, Tatiana Batura, Vladimir Ivanov, and Elena Tutubalina. 2022. RuNNE-2022 Shared Task: Recognizing nested named entities. In *Proceedings of the Dialogue 2022 Conference, Moscow, Russia*, RSUH.

Bayazitova, Flera S., Dariya B. Ramazanova, Zida R. Sadykova, and Tanzilya Kh. Khairetdinova. 2009. *Tatar telenen zur dialektologik süzlege.* Tatarstan kitap näşriyatı, Kazan. (In Tatar).

Bentley, Jon Louis. 1975. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517.

Burnashev, Rustam, Ayrat Gatiatullin, and Mansur Khamidullin. 2025. Geolinguistic system for dialect similarity analysis based on associative rules and fuzzy logic. In *Proceedings of the 10th International Conference on Computer Science and Engineering (UBMK), Istanbul, Turkiye*, pages 1782–1785.

Choure, Aditya A., Rahul B. Adhao, and Vinod K. Pachghare. 2022. NER in Hindi language using transformer model: XLM-Roberta. In *2022 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS), Pune, India*, pages 1–5.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsuper-

vised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, Minnesota*, pages 4171–4186.

Efimov, Pavel, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. 2020. SberQuAD – Russian reading comprehension dataset: Description and analysis. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020)*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 3–15, Springer Science and Business Media Deutschland GmbH.

Fenogenova, Alena, Maria Tikhonova, Vladislav Mikhailov, Tatiana Shavrina, Anton Emelyanov, Denis Shevelev, Alexandr Kukushkin, Valentin Malykh, and Ekaterina Artemova. 2021. Russian SuperGLUE 1.1: Revising the lessons not learned by Russian NLP models. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue 2021", Moscow, Russia.*

Gafarov, Fail M., Viliuza R. Gafarova, and Madehur M. Ayupov. 2025. Explainable artificial intelligence methods in text classification of machine learning models for Tatar language. In *Proceedings of the 10th International Conference on Computer Science and Engineering (UBMK), Istanbul, Turkiye*, pages 1796–1800.

Galimov, Mansur, Rustam Burnashev, and Ayrat Gatiatullin. 2023. Designing a prototype of a fuzzy expert system for a dialectologist using geographic information systems and technologies. In *Proceedings of the 8th International Conference on Computer Science and Engineering (UBMK)*, pages 382–386, Burdur, Turkiye.

Garipova, Firdaus G. 1984. *Tatarstan gidronimnarı süzlege. Berenche kitap.* Tatarstan kitap näşriyatı, Kazan. (In Tatar).

Garipova, Firdaus G. 1990. *Tatarstan gidronimnarı süzlege. Ikenche kitap.* Tatarstan kitap näşriyatı, Kazan. (In Tatar).

Garipova, Firdaus G. 1991. *Issledovaniya po gidronimii Tatarstana.* Nauka, Moscow. (In Russian).

Garipova, Firdaus G. 2010. *Tatar toponimnarı süzlege.* Tatarstan Academy of Sciences, Kazan. (In Tatar).

GeoNames. 2025. Geonames. [Electronic resource]. Accessed: 28.02.2026.

Gilmullin, Rinat A. and Ramil R. Gataullin. 2017. Morphological analysis system of the Tatar language. In *Proceedings of the 9th International Conference on Computational Collective Intelligence (ICCCI)*, pages 519–528, Springer, Cham.

IPSAN. 2024. tat_monocorpus_v2. Hugging Face. [Electronic resource]. Accessed: 28.02.2026.

Johnson, Jeff, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Khasanov, Mansur Kh., editor. 2002–2014. *Tatar ensiklopediyase: 6 tomda.* Institute of the Tatar Encyclopedia, Academy of Sciences of the Republic of Tatarstan, Kazan. (In Russian).

Mindubaev, Artur and Ayrat R. Gatiatullin. 2024. Problems of semantic relation extraction from Tatar Text Corpus 'Tugan Tel'. In *Proceedings of the 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE), Novosibirsk, Russian Federation*, pages 1690–1694.

Mukhamedshin, Damir R., Ayrat R. Gatiatullin, Nikolai A. Prokopyev, and Rinat A. Gilmullin. 2025. Semantic annotation in Electronic Corpus of Tatar Language 'Tugan Tel' based on knowledge graph. In *Proceedings of the 10th International Conference on Computer Science and Engineering (UBMK), Istanbul, Turkiye*, pages 1792–1795.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), Austin, Texas, pages 2383–2392.

Rehurek, Radim and Petr Sojka. 2006. Gensim – python framework for vector space modelling. In *Proceedings of the 9th International Conference on Text, Speech and Dialogue (TSD), Brno, Czech Republic*, pages 45–50.

Sattarov, Gumar F. 1998. *Tatar toponimiyase.* Kazan University Press, Kazan. (In Tatar).

Saykhunov, Mansur R., Rustem R. Khusainov, and Tavzikh I. Ibragimov. 2025. Written Corpus of the Tatar Language. [Electronic resource]. Accessed: 03.01.2026.

Schweter, Stefan. 2020. BERTurk – BERT models for Turkish. DOI: 10.5281/zenodo.3770924.

Sinnott, Roger W. 1984. Virtues of the Haversine. *Sky and Telescope*, 68(2):159.

Sketch Engine. 2024. Tatar mixed corpus – Tatar corpus from the web. Sketch Engine. [Electronic resource]. Accessed: 03.01.2026.

Suleymanov, Dzavdet Sh., Alexander Ya. Fridman, Rinat A. Gilmullin, and Boris A. Kulik. 2021. System analysis of the problem of natural language modeling. *Transactions of the Kola Science Centre. Information Technologies*, 12(5):57–66. (In Russian).

TatarNLPWorld. 2025a. Tatarstan toponyms dataset. Hugging Face. [Electronic resource]. Accessed: 28.02.2026.

TatarNLPWorld. 2025b. Tatarstan toponyms QA dataset. Hugging Face. [Electronic resource]. Accessed: 28.02.2026.

TatarNLPWorld. 2025c. Turkic NLP & low resource languages research hub. Hugging Face. [Electronic resource]. Accessed: 28.02.2026.

Wang, Liang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. Text embeddings by weakly-supervised contrastive pre-training. ArXiv, abs/2212.03533.

Äkhmätÿanov, Rifkat G. 2015. *Tatar telenen etimologik süzlege: Ike tomda.* Mägarif – Vakıt, Kazan. (In Tatar).

Automatic Literary Adaptation? A Survey of Related Tasks, Trends and Challenges

Iglika Nikolova-Stoupak
Sens Texte Informatique Histoire
Sorbonne Université, Paris, France
iglika.nikolova-stoupak@etu.sorbonne-universite.fr

Gaël Lejeune
Sens Texte Informatique Histoire
Sorbonne Université, Paris, France
gael.lejeune@sorbonne-universite.fr

Eva Schaeffer-Lacroix
Sens Texte Informatique Histoire
Sorbonne Université, Paris, France
eva.lacroix@inspe-paris.fr

This paper serves as a survey of the contemporary Natural Language Processing (NLP) landscape with regard to its potential to perform or assist in literary adaptation - the modification of literature in order to meet the needs and preferences of specific reader audiences, such as children, foreign language learners and people with learning disabilities. Firstly, we present the need for and nature of literary adaptation. Then, we provide overviews of the current contexts in relation to existing NLP tasks that are of direct relevance to literary adaptation: simplification, summarisation, style transfer, and creative text generation. Throughout the survey, we focus on large language models (LLMs) as the current state of the art, as well as on existing methods of evaluation of automatically generated text and, where possible, on multilingual applications. We wrap up with a discussion of future directions for literary adaptation as the intersection of the discussed tasks. In our view, contemporary NLP has the potential to engage in this highly novel task, given an established multidimensional evaluation framework and ongoing engagement on the side of professionals in fields such as education and creative writing.

Keywords: simplification, summarisation, style transfer, creative generation, survey, literary adaptation

1. Introduction

To date, Natural Language Processing (NLP) has not consistently engaged in the generation of adapted literary text that aims to meet the needs of specific readerships (such as children of a defined age, language learners of a defined level, or people with learning disabilities). Whilst related tasks such as text simplification and endeavours like easy-read (Freyer, Kempt, and Klöser 2024) come close in terms of definition and purpose and do not intrinsically exclude application to literature, the current focus remains on information texts. Therefore, no unified framework and practices are established regarding the distinguishing

https://doi.org/10.47810/JCLIB.2.2026.04

features of literature, such as large textual length, narrative voice, style, and figurative language.

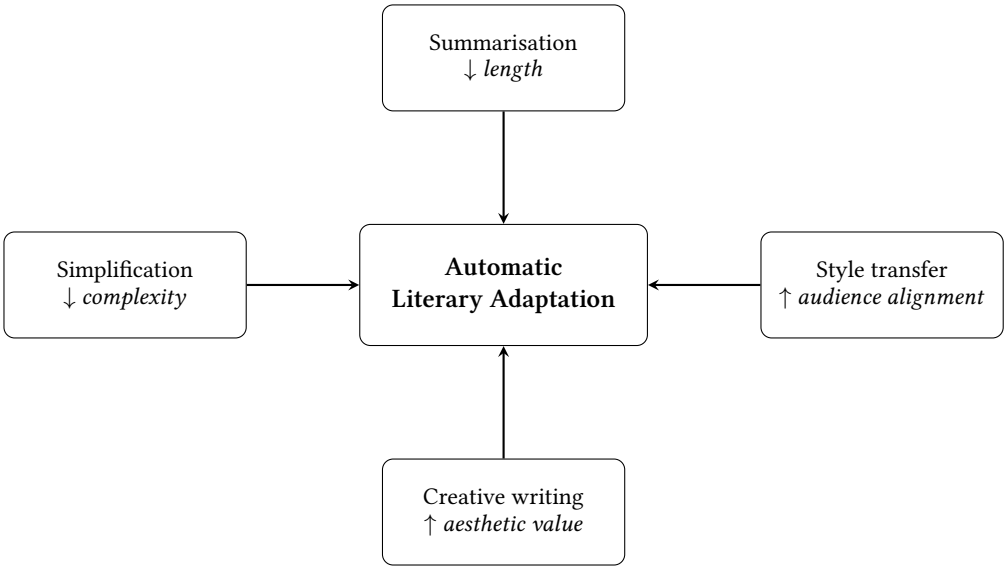


Figure 1
Automatic literary adaptation at the crossroads of relevant NLP tasks.

In this survey-style paper, we lay out the current landscapes in relation to NLP (more specifically, "natural language generation", or "NLG") tasks that are of direct relevance to the concept of automatic literary adaptation: simplification, summarisation, style transfer, and creative generation. See Fig. 1 for an overview of these tasks' significance to the topic at hand. We performed a bibliometric analysis in order to summarise and visualise a timeline of the interest in the mentioned tasks through the proxy of the related articles in the platform Semantic Scholar[1]. We utilised Semantic Scholar's official API to determine the number of title matches per year (1950 to 2026) of search terms elaborated to represent each task[2]. We also limited the results by the domain "Computer Studies". As Fig. 2 shows, interest (and, coincidentally, advancement) in all four tasks has been growing through the years. Summarisation is represented by the highest number of articles overall as well as by significant growth since the early 2000s i.e. the time when statistical models became largely used. The other three tasks, for which the aspect of semantics has a stronger role, are associated with peaks in the late 2010s and early 2020s, when LLMs were introduced.

1 https://www.semanticscholar.org/
2 simplification: SIMPLIFICATION OR SIMPLIFY OR SIMPLIFYING
summarisation: SUMMARIZATION OR SUMMARISATION OR SUMMARIZING OR SUMMARISING
style transfer: STYLE TRANSFER OR STYLE ADAPTATION
creative generation: CREATIVE WRITING OR CREATIVE GENERATION OR STORY GENERATION OR STORY WRITING OR NARRATIVE GENERATION OR NARRATIVE WRITING OR POETRY GENERATION OR POETRY WRITING

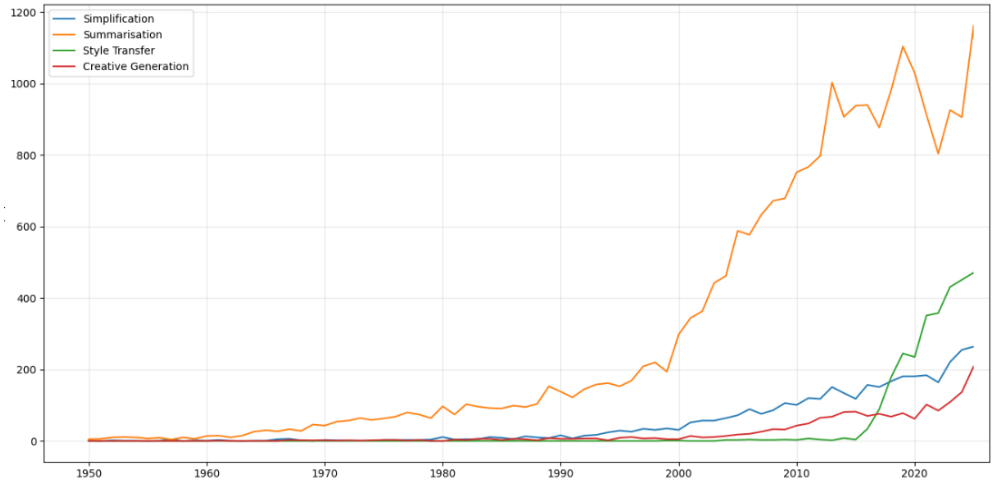


Figure 2
Interest in the four investigated NLP tasks, represented by the number of associated articles per year in Semantic Scholar.

In particular, simplification and creative generation demonstrate very similar patterns. A recent phenomenon, style transfer emerged around this same time and as of today has surpassed both simplification and creative generation in terms of article presence.

We will commence the analysis by presenting the need for automatic literary adaptation and introducing its traditional, human-led counterpart. Then, in sections 4-7, we will focus on each of the four mentioned related tasks (simplification, summarisation, style transfer, and creative generation). Each of these sections will comprise: 1) a definition of the task, 2) a general timeline of its development (typically moving from rule-based methods through statistical and neural methods to reach LLMs as the current state of the art), 3) associated evaluation methods and 4) takeaways in view of automatic literary adaptation. Our focal points will include evaluation practices, multilingualism and the use of LLMs. We will end with a discussion that sums up the practices, strengths and limitations of the analysed NLP subfields and sets general guidelines concerning the steps that should be taken in order for automatic literary adaptation to be defined as a task in its own right. Throughout the study, we will use the term "literary adaptation" rather than "literary simplification" in order to underline the complex and multifaceted nature of altering text to address a large variety of readerships.

2. Motivation

The reading of literature has been shown to aid the development of key psychological and social abilities ranging from critical thinking and imagination to emotional understanding and empathy (Nussbaum 2010; Oatley 2011; Mar et al. 2006). With their experimental study

conducted with over 33,000 participants, Mar et al. (2021) found that narrative text is better understood and memorised than essay-like text. However, there is also evidence that when a reader does not sufficiently understand a literary text, the benefits are at best reduced and, at worst, there may be harmful consequences for the reader, such as a decrease in confidence and motivation (Krashen 1982; Nation 2001; Day and Bamford 1998). In contrast, the matching of text to one's current abilities has been shown to result in better learning outcomes (FitzGerald et al. 2018). The achieved higher reader interest reduces "mind wandering", leading to higher engagement and even understanding across ages, languages, and text types (Bonifacci et al. 2023). Accordingly, the past few decades have seen significant production of specially conceived literary works, which in turn more or less narrowly address the needs of defined audiences. These works are often adaptations of already existing and popular (a.k.a. "canonical") texts. Specific book series and editions have been developed that are for children, learners of foreign languages, and people with learning disabilities. The most established ones follow clear frameworks that separate their audience in a fine-grained manner in accordance with their needs, such by specific ages (e.g. Oxford Reading Tree, Penguin Young Readers) or specific proficiency levels (e.g. Oxford Bookworms for English, Lectures CLE en français facile for French).

Still, the availability of adapted literature is limited by a series of factors, including the predominance of English and other popular languages at the expense of most others, concerns of physical availability and affordability, the multiplicity of differentiated audiences and even the implied workload required from professionals in order to compose the texts. Given the current technological context, it is therefore natural to envision the assistance of automated tools, whose extent remains to be determined based on technological potential and ethical norms. For the purpose of gently addressing both of these concerns, in the current paper we firstly raise the question of adaptation of already existing literary texts rather than the composition of new ones.

3. Background: Literary Adaptation

Hutcheon (2006) points out that the word "adaptation" denotes both a result or artefact and the process of its creation. Ewers (2009, p. 179) goes on to say that the practice signals "any modification of a work as it moves across boundaries of media or genres". Occasionally, more than one such movement is accomplished at a time; for instance, in its versions for children, the novel *Gulliver's Travels* typically shifts from satire to adventure story. Other elements that may change in the process of adaptation include points of view, ontologies and philosophies (Hutcheon 2006). Artistic adaptation has a long history that predates modernity; for instance, in Victorian times, stories, poems, operas, paintings and other media would often allude to one another.

Specifically addressing literary adaptation, Lefebvre (2021, p. 54) defines it as "texte rémanent d'un texte littéraire source, socialement reconnu dont il a pour fonction de tenir lieu" ("a text deriving from a socially recognised source literary text, whose function is to stand in its place"[3]). The extent of the undergone changes may vary significantly in size

3 Translations are the first author's.

and nature. Nières-Chevrel (2009, p. 190) ironically notes that children's literature "se donne toute liberté de ramener à vingt pages un roman qui en faisait deux cents, de maintenir ou de taire le nom de l'auteur de l'œuvre originale pour n'en garder que le seul titre" ("allows itself all freedom to bring down a two-hundred-page novel to twenty pages, to retain the name of the original author or to omit it, keeping only the title"). Occasionally, it may be the original author that adapts their own text, such as in the case of Lewis Carroll's *Alices' Adventures in Wonderland* and its two later versions, including one markedly meant for children (entitled *The Nursery Alice*). One may notice that it is often recurring canonical literary works (e.g. *Pinocchio*, *Don Quixote*, *Les Miserables*) that are subject to adaptations into films, children's stories and other media and genres. Ellis (1982, p. 190) points out that such works have been so recycled through formal adaptation and other allusions that the audience is now familiar with "a generally circulated cultural memory" rather than the original texts. Considerations in addition to (but often related to) canonical status that may influence the choice of literary works to undergo adaptation include financial planning and the demands of educational systems.

The canonical texts that are subject to literary adaptation often carry cultural significance for their target audience e.g. young children that share the text's underlying culture or language learners that may be interested in discovering not only the target language but its associated culture. The status of cultural heritage may cause adverse reactions to the process of adaptation and equate it to a loss in authenticity. An example of the phenomenon occurred in relation to author Russi Chanev's adaptation of the famous nineteenth-century Bulgarian novel *Under the Yoke*, which highly altered the original archaic language of the work in order to increase its comprehensibility for children of Bulgarian origins who live outside of the country (Vazov 2024; Intrigi.bg 2024). Addressing potential issues of the type, Hutcheon (2006, p. 70) claims that "[t]here always is a trade-off in adaptation". Thiel (2013) qualifies said trade-off in the following way: whilst an adaptation might decrease an original text's historical value, it offers contemplation on the relationship between the past and the present in relation to literature and beyond. Postmodern and poststructuralist views of adaptation include a shift away from its secondary status and towards its importance as an independent artefact (Müller 2013).

We will now discuss several key audiences that tend to be addressed in the process of literary adaptation. Children constitute the most largest one. Adapted texts for children typically have higher availability compared to those for other audiences, notably when it comes to low-resourced languages. The practice of adaptation for children has a century-long history. Müller (2013) cites Joachim Heinrich Campe's *Robinson the Younger* of 1780 (an adaptation of *Robinson Crusoe*) as the first formal literary adaptation for a young audience. Other early adaptions include Thomas Bowdler's *Family Shakespeare* (1807), a children-friendly version of Shakespeare's plays. The need for adaptations for children has historically stemmed from a general focus on education as well as on the idea that children need to be protected from explicit or traumatic material. In the current landscape, the link between literature and education is sometimes rethought, and there is a vast variety of books to choose from in accordance with a child's specific interests (Shavit 1999). In addition to being adapted, children's classics often reach their audience as translations, thereby introducing

another layer of "adaptation" (one to a different language and culture). Also, an older source text may be seen as simultaneously adapted to another time period i.e. modernised.

There are existing attempts to methodologically define and even quantify the practice of literary adaptation for children. Van Lierop-Debrauwer (2000) categorise the adjustable aspects of a literary work as content, style, structure and design. Providing a microscopic view on ideational grammatical metaphors (the abstract in nature expression as a noun of a notion that would typically require a verb or adjective) in literary adaptation for a younger audience, Liu (2021) defines the following five strategies: addition, maintenance, revision, unpacking and demetaphorisation. In a curious study that compares texts written in English and Dutch for children, teenagers and adults by the same authors (referred to as "crosswriters"), Haverals, Geybels, and Joosen (2022) apply stylometric features, clustering and PCA analysis to ultimately conclude that literature for children shares more similarities than works for different audiences by the same author. This conclusion suggests that audience-driven variation is a measurable and systematic phenomenon.

In the past few decades, practices such as comprehensive input and extensive reading have been dominant in foreign language education (Krashen 1982). They are directly related to observations such as the easier acquisition of vocabulary items when presented within natural context (Godwin-Jones 2018; Restrepo Ramos 2015). Literary series, known as "graded readers", have emerged in an attempt to maximise the learning process. Used in the context of both independent and in-class learning, these materials have proved to help reaffirm vocabulary knowledge as well as increase student motivation and sense of community (Wan-a rom 2008; Hill 2013; Dickinson 2017). Graded readers often consist in adaptations of famous literary texts, typically related to the target language's culture. Their difficulty is classified in terms of proficiency levels, such as per the Common European Framework of Reference for Languages (CEFR). The expected number of "headwords" i.e. already acquired vocabulary knowledge, may also be indicated. The literary text may be accompanied by glossaries, exercises, and cultural information. Today, graded reader series in English, such as Oxford Bookworms, Penguin Readers, and Cambridge English Readers are an established phenomenon. Albeit to a much lesser extent and proportionately to their general popularity/resourcedness, other languages also benefit from such materials; for instance, Lectures CLE en français facile (French), Leichtzulesen (German), Sistema Kalinka (Russian), Mandarin Companion (Chinese), and Taishukan (Japanese).

Specific learning disabilities and other medical conditions also influence a person's understanding of written text (and literary text in particular) in different ways, which are more or less narrowly related to the condition at hand. Autistic people tend to have difficulty in processing figurative language (Happé 1993; Kalandadze et al. 2018) that may be improved given relevant intervention such as the inclusion of images or explicit explanations (Melogno, Pinto, and Orsolini 2017; Rutherford et al. 2020; Mashal and Kasirer 2012). Dyslexia is also often associated with reduced understanding of abstract language as well as long words and complex syntax (Snowling, Hulme, and Nation 2000; Cersosimo, Domaneschi, and Cancer 2025). While not intrinsically related to language acquisition and understanding, ADHD similarly impedes the processing of long segments and complex syntax, and patients benefit from the inclusion of explicit signalling, such as headings and keywords (Martinussen et al. 2005). In turn, aphasia has been associated with difficulties

in relation to infrequent words and ambiguous syntactic structures (Caplan 2012; Dickey and Thompson 2009). While to our knowledge, there are no established literary editions or series that target specific disabilities, it is worth mentioning examples such as the UK-based publisher Every Cherry, which defines its audience via the wider term "special needs" as well as Spanish publisher Almadraba's collection Kalafate, which implements practices of easy-read, thereby addressing conditions such as dyslexia and ADHD.

4. Related NLP Tasks: Simplification

4.1 Definition

The purpose of automatic simplification is to provide a simpler version of a text that retains efficiently its original information. The output should also be coherent and grammatical. The new text may be addressed at a reader audience or used as a pre-processing step in the context of additional NLP tasks. A common division is that between lexical and syntactic simplification. Utilised techniques may include syntactic operations (e.g. sentence splitting), lexical operations (substitution of infrequent or technical words), content reduction (through summarisation and paraphrases) and clarification (added explanations or definitions) (Štajner and Saggion 2013). Recent large-scale initiatives aiming at text simplification (such as the US Government's Plain Language Action and Information Network[4] and Inclusion Europe's Easy-to-Read Guidelines[5]) have placed a focus and a sense of urgency on the need for efficient automation of the process.

4.2 Timeline

Rule-based text simplification typically includes the application of manually-established syntax rules and the replacement of vocabulary items based on pre-established lists of synonyms. Devlin and Tait composed parse trees of sentences, based on which rules of splitting, rearrangement and deletion were applied. Devlin (1999) and Devlin and Unthank (2006) specifically developed rules targeted at the increase of text understandability by aphasia patients.

The launch of Simple Wikipedia in 2003 became a turning point in statistical text simplification, as it provided numerous ready examples of paired original and simplified text. Using aligned sentences from the original and simple Wikipedia, Coster and Kauchak (2011a) defined the simplification task as an English-to-English translation task. The translation method was to also be applied to additional languages and dialects, such as German, Chinese, Swedish and Basque (Klaper, Ebling, and Volk 2013; Chen et al. 2012; Stymne et al. 2013; Aranzabe, Ilarraza, and Gonzalez-Dios 2012). Biran, Brody, and Elhadad (2011)'s unsupervised model learned simplification rules from a comparable corpus. In contrast, Glavaš and Štajner (2015) used word embeddings and cosine similarity in the task of lexical

4 https://www.plainlanguage.gov
5 https://www.inclusion-europe.eu/easy-to-read/

simplification in addition to WordNet lists, frequency and a classifier model that seeks words with highest simplicity. Coming first at the relevant SemEval 2016 Complex Word Identification task, Paetzold and Specia (2016) combined different lexicon-based, threshold-based and machine learning approaches. Stajner and Saggion (2013)'s work on automatic text simplification was focused on the Spanish language and addressed discrete reader audiences, such as people with Down's syndrome and autism spectrum disorders. They used a corpus of professionally adapted articles on different topics and applied the operations of sentence deletion and splitting.

The introduction of neural NLP offered new state-of-the-art simplification techniques. Nisioi et al. (2017)'s sequence-to-sequence simplification model achieved higher grammaticality and meaning preservation as well as a higher degree of simplification compared to earlier methods. Zhang and Lapata (2017)'s reinforcement-learning-based DRESS (Deep REinforcement Sentence Simplification) model was trained to encourage output that preserves the meaning of original input while maintaining simplicity and fluency. Cripwell, Legrand, and Gardent (2023) proposed a high-performing transformer model that is based on document-level simplification (in contrast to the common sentence-level). The model marked sentences to be copied, rephrased, split or deleted based on both context and internal structure.

The LLM era has been marked with a variety of projects linked to text simplification. Kew et al. (2023) proposed Benchmarking Large Language Models on Sentence Simplification (BLESS), a comparison of 44 LLMs ranging from 60 million to 176 billion parameters in relation to the task. They used a few-shot setting and applied both qualitative and automatic analysis (including SARI, BERTScore and LENS). Closed-weight models, in particular Davinci-003 and GPT-3.5-Turbo, performed particularly well, and instruction-tuning was shown to improve models' performance. Feng et al. (2023) concluded that GPT3.5 and ChatGPT perform the task as well as humans in English, Portuguese and Spanish. Evaluation was based on SARI and FKGL[6] (and the latter's Spanish counterpart, FRES); in addition, 100 sentences were human-evaluated. Yang (2024) compared fine-tuning and few-shot prompting of GPT models on the simplification of medical-related sentences using a professional dataset and human evaluation of correctness and quality, reaching the conclusion that the two methods were comparable in performance. Ormaechea and Tsourakis (2024) fine-tuned FLANT5-based language models for the simplification of French text, using the WiViCo dataset (which contains over 40k pairs of complex and simpler sentences) and evaluated them using human evaluation, BLEU and SARI scores. The BEA 2024 shared task approached multilingual lexical simplification as covering a variety of textual genres and target audiences (Shardlow et al. 2024). The highest-performing models used a variety of techniques ranging from handcrafted rules to LLM models such as GPT, proving the former's ongoing relevance.

Recent advancements in automatic text simplification have also led to finer-grained reader audience distinctions, such as simplification to a particular CEFR level. Farajidizaji, Raina, and Gales (2024) tested the ability of the popular models ChatGPT and Llama-2

6 Flesch–Kincaid Grade Level, a readability measure, see (DuBay 2004)

to bring text of any source level to any target level in terms of Flesch Reading Ease[7]. Indeed, Spearman correlation proved that the models succeeded in adjusting the readability of documents. However, the issuing levels did not map unambiguously to Flesch values; instead, they were dependent on the source texts' readability. Similarly, Benedetto and Buttery (2025) noted that a selection of LLMs, including Llama-3 and GPT, perform CEFR-targeted simplification efficiently yet imperfectly, typically oversimplifying texts. Jamet, Shrestha, and Vlachos (2024) approached French-language simplification based on CEFR levels, focusing on current models' ability to reduce a source sentence by a single level at a time. They used labelled sentences in a few-shot setting and classified the derived text's level using the CamemBERT model. GPT-4 achieved best results as compared to Mistral and Davinci models.

4.3 Evaluation

Alva-Manchego et al. (2020) sum up that the following main criteria are considered at evaluation of automatic textual simplification models: output fluency, meaning preservation and simplicity in comparison to the input (unsimplified) text. Many authors opt for automatic scores conceived for the task at hand or for related NLG tasks.

Translation Edit Rate-plus (TERp) provides the number of edits required in order to transform the source text into the simplified output (Snover et al. 2009). Designed specifically for sentence simplification, SARI (Xu et al. 2016) measures accuracy and recall in relation to words that have been added, deleted or retained, making use of several reference simplified sentences. Originally used for machine translation, the term-overlap metric BLEU measures n-gram overlap between the reference and evaluated texts, while penalising output deemed too short (Papineni et al. 2002). Adjustements of BLEU introduced in view of the task of simplification include FKBLEU (Xu et al. 2016), which compares the output to both a reference and the input. BLEU has been noted to perform better in terms of meaning preservation, whilst SARI is a better estimate for simplicity; however, both metrics show overall low correlation with human judgement (Xu et al. 2016; Al-Thanyyan and Azmi 2021). Also used primarily in machine translation tasks, METEOR improves on BLEU by calculating matches to the reference based on unigrams' surface forms, stemmed forms and meanings, while also taking into consideration the order of matched elements (Banerjee and Lavie 2005). Another metric derived from BLEU is NIST, which assigns different weights to n-grams based on their informativeness (Doddington 2002). Originally used for machine translation and widely applied to text simplification, COMET, state of the art on the WMT 2019 Metrics shared task, has an XLM-RoBERTa encoder and makes use of the source text and a multilingual embedding space (Rei et al. 2020). ROUGE (Lin 2004) is another relevant n-gram-based metric, which was conceived for the task of summarisation (see Section 5.3 for more detail). It has, however, been noted to correlate poorly to human judgement at the task of simplification (Scialom et al. 2021a). Demonstrating better correlation with human judgement, SAMSA may be seen as an improvement on SARI that provides a better overview

7 See DuBay (2004)

of text simplification results by addressing semantic structure (Sulem, Abend, and Rappoport 2018). It is a reference-less metric that focuses on predicate-argument relations while considering "scenes" rather than sentences. BERTScore changes the evaluation landscape by making use of a BERT model to calculate semantic rather than n-gram-based similarity between an automatic output and a reference text (Zhang et al. 2020). It is applied to a variety of NLG tasks, including simplification.

QuestEval, formulated for summarisation tasks, was later updated so as to be relevant to simplification evaluation (Scialom et al. 2021b). The metric evaluates the factual consistency of an output based on the associated source document. For the purpose, a list of questions is generated based on the former, and answers are then retrieved from both texts, whilst their similarity is calculated. In its simplification-friendly version, the similarity is calculated using BERTScore instead of the original F1-score, thereby accounting for the use of synonyms and reformulations. Finally, LENS is a metric specifically crafted for simplification that uses a pre-trained model, based on human judgements pertaining to meaning preservation, simplicity and fluency (Maddela et al. 2022). The EASSE framework (Alva-Manchego et al. 2019) standardises the evaluation of text simplification through commonly used automatic metrics, offering a unified toolkit. In turn, Stodden (2024) propose EASSE-multi, a related framework that extends to non-English languages, and exemplify its use for German simplification.

Many of the mentioned automatic metrics used in text simplification are dependent on the presence of high-quality datasets of associated simple and complex sentences. Popular datasets of the type include Wikipedia-Simple Wikipedia for English (Coster and Kauchak 2011b), Newsela for English and Spanish (Xu, Callison-Burch, and Napoles 2015), PorSimples for Brazilian Portuguese (Aluísio and Gasperin 2010), Alector for French (Gala et al. 2020) and 20 Minuten for German (Rios et al. 2021). OneStopEnglish is a corpus for readability assessment and automatic text simplification that consists of 189 original texts, each accompanied with three graded simplifications for learners of English as a second language(Vajjala and Lučić 2018). It is worth mentioning that datasets focused on literary simplification, although limited in number and size, have also been produced. Washburne and Vogel (1926) published the *Winnetka Graded Book List*, which includes 700 children's books, manually annotated for reading difficulty. MULTISIM is a multilingual simplification benchmark that includes 1.7 million pairs of complex and simplified sentences over 12 languages(Ryan, Naous, and Xu 2023). It covers eight domains, including literature (novels). Brunato et al. (2015) offer a corpus of 32 short Italian novels for children and their manually simplified versions for "poor comprehenders''. They go on to annotate it for a selection of simplification procedures[8]. Dmitrieva and Tiedemann (2021) compile a simplification dataset of 19 books for Russian-language learners at different CEFR levels; then, they go on to train an LSTM-based classification model and evaluate it using the EASSE framework.

Constituting a differing approach, evaluation through measures of readability is often evoked in relation to automatic simplification. The term "readability", in use since the 1950s, denotes measures and formulas that are meant to estimate a text's level of difficulty, typically

8 split, merge, reordering, insert, delete and transformation

so as to link it to a particular reader audience, such as a school grade level (DuBay 2004). Early readability formulas focus on length-based measures such as word and sentence length (e.g. Flesch Reading Ease) and/or on lexical mapping (e.g. Dale-Chall). Originally conceived with the English language in mind, readability formulas based on superficial text characteristics have been shown to correlate across languages (van Oosten, Tanghe, and Hoste 2010). Also, adaptations of some readability formulas for other languages have been elaborated, such as Flesch-Douma (an adaptation of Flesch Reading Ease for Dutch) (Douma 1960) and Spaulding's Spanish readability formula (an adaptation of Dale-Chall) (Spaulding 1956). The main advantage of readability-based features and formulas is that they make up an universal convention that is not dependent on a limited reference corpus. They can therefore be applied independently in evaluating the difficulty of both an original text and a simplified text derived from it. However, the method is also associated with significant limitations, such as the inability to account for errors and to address textual semantics. Less "shallow" atomic textual features not typically present in readability formulas have also been noted to have high relevance to textual simplification, such as the use of complex sentences, passive and conditional constructions, and figurative expressions (Štajner and Saggion 2013; Evans, Orăsan, and Dornescu 2014).

Human-based evaluation has also been widely applied in automatic simplification tasks, with a common focus on fluency, grammaticality, simplicity, and meaning preservation. Al-Thanyyan and Azmi (2021) use a Likert scale (1 to 5 or 1 to 3) to estimate the mentioned features. Maddela et al. (2022) introduce the human evaluation framework RANK & RATE, in which candidate simplifications are rated through an interactive interface.

4.4 Takeaways

Automatic text simplification is possibly the discussed task with highest relevance to automatic literary adaptation, and many of its developments and conclusions are readily applicable to the latter. Initially, the task tended to involve no clear definition of "simple" and "complex" text, and the focus was on the difference in simplicity between source and target text rather than a more universally applicable distinction. Gradually, narrow reader audiences such as people with different disabilities and foreign language learners at different levels came to be more frequently addressed in simplification research. Recent advancements, such as document-level granularity and application to various languages, although still developing, are cause for optimism. They go hand-in-hand with general NLP trends, such as language models' increasing semantic knowledge and proficiency in low-resource languages.

The evaluation of simplified text via readability-based and other predefined characteristics has the advantage of alleviating the need for a large number of reference datasets associated with different textual genres, audiences or languages. Established automatic measures, while imperfect individually, have the potential of covering different textual aspects (e.g. SARI has proven to be a good proxy for lexical simplicity) and to therefore work efficiently when used collectively. Finally, human evaluation has been used consistently with text simplification, including in the era of LLMs and is likely to continue being a necessity.

Future advances, including in literary adaptation, can benefit from a systematisation of efficient human evaluation frameworks.

5. Related NLP Tasks: Summarisation

5.1 Definition

As Fig. 2 demonstrates, interest in automatic summarisation has been both early and widespread. The need for reduction in textual size goes hand in hand with the increase in quantity of online data and people's related difficulty in consuming it in terms of time and effort. Summarisation is often an element of simplification, although the two tasks can also exist independently of one another (as shown, the simplification of a text may sometimes require its increase in textual length, such as through the addition of explanations). Although the decrease in size involved in summarisation may vary significantly based on the associated documents and tasks at hand, a general guideline is that reduction by at least half is to be aimed at (Radev, Hovy, and McKeown 2002). The most widely accepted categorisation is that between the extractive and abstractive approaches to summarisation. In the case of the former, the parts of a text that are of highest importance are taken verbatim to constitute its summarised version. In contrast, the latter approach implies reformulation of the text. Earlier practices of automatic summarisation typically followed the extractive approach and involved preprocessing the source text and ranking its tokens (usually, sentences) in order of importance, such as through TF-IDF-based methods. In turn, as abstractive summarisation involves novel text generation, it became widespread during the pre-neural and neural periods. Still, this is not to say that the need for extractive summarisation disappeared, nor that abstractive summarisation was not attempted earlier on. A distinction also exists between generic and query-based summarisation. Generic summarisation seeks to provide a general overview of a text's content, whilst a query-based one focuses on information that is relevant to a specific search query at hand. In addition, summarisation may be of one or multiple documents. Like simplification, summarisation may also be used as an intermediary step in other NLP tasks, such as information retrieval and question answering.

5.2 Timeline

Luhn (1958)'s work on automatic generation of articles' abstracts has been cited as the first application of automatic text summarisation. He made use of word frequency and distribution in order to rank the significance of both words and sentences. Hu and Liu (2004) provided summaries of the totality of customer reviews of a given product based on the mentioned features of the product as well as a review's overall positivity or negativity. Ko and Seo (2008) combined pairs of consecutive sentences prior to ranking them for significance so as to account for context. Addressing a literary audience, Kazantseva and Szpakowicz (2010) extracted summaries of short stories with the purpose of aiding readers in deciding whether to read the full text. Heuristics were used to select descriptive content that pertains to a story's setting and main entities but does not reveal the plot. Genest and

Lapalme (2012) combined rule-based methods of information extraction with NLG in the production of summaries.

During the pre-neural NLP period, extractive summarisation methods still dominated the field. Ouyang et al. (2011) used Support Vector Regression (SVR) to estimate the importance of sentences in the context of query-focused multi-document summarisation. Wang and Ma (2013) represented texts as matrices and inferred underlying topics via dimensionality reduction, achieving high ROUGE-score results. In turn, Shetty and Kallimani (2017) applied matrix decomposition and k-means clustering of sentences in order to extract diverse and redundancy-free summaries. Kobayashi, Noguchi, and Yatsuka (2015) used word embeddings to represent documents and candidate summaries and ranked different extractive summaries based on distances between embeddings.

Due to its simplicity and efficiency, extractive summarisation did not become obsolete even with the rise of neural NLP. Cheng and Lapata (2016) approached the task using a hierarchical encoder that models document structure combined with an attention mechanism that determines tokens' importance for a given document. The model was trained on a large dataset and did not imply textual preprocessing. Abdel-Salam and Rafea (2022) evaluated BERT models on the task of extractive summarisation and proposed the fine-tuned easily trainable "SqueezeBERTSum" model.

Attention-based sequence-to-sequence models proved to be particularly fitting for the task of automatic summarisation, particulary in relation to long documents. Chopra, Auli, and Rush (2016) trained a RNN model with an attention-based encoder to aid the decoder's focus at the generation of abstractive summaries. Chen et al. (2019) proposed RC-Transformer (RCT), a model that works efficiently with long-term dependencies and outputs better-ordered text. The model was extended with an additional RNN-based encoder and a convolution module that filters text with local importance.

The LLM era brought about a variety of summarisation-related work and an exploration of methods such as prompt engineering, fine-tuning and knowledge distillation (Zhang et al. 2026). The primary focus remained on informativeness and the reduction of large amounts of data for practical purposes. Xiao and Chen (2023) proposed an LLM-based method of evolutionary fine-tuning of news articles, focusing on informativeness. Human evaluation ranked LLM output as better than one produced by humans and fine-tuned neural models, specifically emphasising its qualities of fluency and coherence. Zhang et al. (2024) compared performance at the task between humans and ten discrete LLM models, concluding that the former's summaries are more abstractive in nature. Pu, Gao, and Wan (2023) boldly claimed in their article's very title that (human-based) "summarization is (almost) dead", basing themselves on the results of five summarisation tasks, including crosslingual summarisation. In contrast, Wang et al. (2023a) had significant criticism regarding current LLMs' crosslingual summarisation abilities. More specifically, they noted that GPT-4 demonstrates competitive performance based on ROUGE and BERTScore (but still performs worse compared to a fine-tuned BART model), whilst Vicuna-13B outright lacks zero-shot abilities for the task. They pointed out that the use of chain-of-thought, such as instructions to first translate and then summarise a document, helps improve models' performance.

The potential of current models to summarise long text is particularly relevant in view of literary adaptation. Wu et al. (2024) incrementally extracted and concatenated key

sentences from a long document, stopping the process when the derived summary resulted in the highest ROUGE score. Chang et al. (2023) divided a long document of over 100k tokens into chunks, and the chunks were then merged hierarchically and incrementally, while textual coherence was used to evaluate the resulting summaries. The highest scoring models were GPT-4 and Claude 2, and hierarchical merging achieved better results.

5.3 Evaluation

The main qualities brought forward in relation to automatically summarised text include its information coverage, information significance (general or user-/task-defined), coherence and lack of redundancy (Huang et al. 2010). Human-based evaluation has focused on these criteria, typically as organised in ranking scales (El-Kassas et al. 2021). Many of the automatic metrics used for the task, including term-overlap based ROUGE, BLEU and METEOR and semantic-similarity-based BERTScore, were already mentioned in relation to automatic simplification (see Section 4.3).

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric was conceived for the task of summarisation (Lin 2004). In its framework, an automatic summary is compared to one or more references in terms of overlapping units. Variations of ROUGE include ROUGE-N, which is based on n-gram units; ROUGE-L, which denotes the longest common subsequence; and ROUGE-S, which offers skip-bigram statistics, thus allowing for gaps between tokens. The ROUGE metric focuses on recall, whilst BLEU (typically used in the evaluation of machine translation), focuses on precision. Metrics that have been introduced in order to improve on ROUGE's performance include Basic Elements, where sentences are broken down into small meaning units, against which summaries are compared (Hovy et al. 2006) and BEwT-E (extended Basic Elements), which also accounts for textual transformations, such as paraphrases and syntactic variations (Tratz and Hovy 2008). In turn, DEPEVAL makes use of dependency parsing and compares summaries based on syntactic structures (Owczarzak 2009).

Metrics such as BERTScore, which take into account textual semantics, are often seen as a better evaluation alternative of summarisation output, especially in cases of abstractive summarisation. Yet, BERTScore has been noted not to offer significant benefits when compared to ROUGE (Scialom et al. 2021a). BLEURT is another similarity score based on BERT, which models human judgement. It can be used in cases of limited reference data; however, it requires training on existing human ratings (Sellam, Das, and Parikh 2020). BARTScore is a comprehensive framework that makes use of a pre-trained BART sequence-to-sequence model (Yuan, Neubig, and Liu 2021). Its main idea is that higher-quality summaries receive better scores based on how likely they are to be generated from, or to generate, a reference text. The score has been noted to correlate well with fluency and accuracy. The Pyramid method is a human-centric evaluation framework that weighs Summary Content Units (units that represent key information) across multiple reference summaries (Nenkova, Passonneau, and McKeown 2007).

Scialom et al. (2021a) proposed QuestEval, a question-answer-based method that does not require reference summaries or ratings at evaluating whether a summary contains all relevant information from an associated source document. QuestEval correlates highly with

human judgement in terms of consistency, coherence, fluency, and relevance. Recently, LLMs have also been made use of as evaluators of automatic summaries. Jain et al. (2023) explored ChatGPT's ability to replicate evaluation methods such as Pyramid and pairwise comparison, achieving high results. Chang et al. (2023)'s BOOOOKSCORE uses GPT-4 prompts to identify problems within automatic output, including causal omissions, salience, and duplication.

Large reference datasets used in the evaluation of automatic summarisation include Gigaword (Napoles, Gormley, and Van Durme 2012) and DUC (National Institute of Standards and Technology 2004)[9], LCSTS (Hu, Chen, and Zhu 2015)[10] and Wikilingua (Ladhak et al. 2020)[11].

5.4 Takeaways

Literary adaptation typically involves an element of summarisation i.e. a significant reduction in size. However, a key distinguishing feature in literary adaptation is that unlike in the case of most examples of automatic summarisation, the focus therein lies in the reading experience itself rather than the efficient presentation and consumption of information. As a result, additional questions arise that do not tend to be addressed within the generation and evaluation of automatic summaries, such as those of general aesthetics, narrative voice, and anaphora resolution.

Abstractive summarisation is the method that holds greater relevance in relation to the literary domain and, as with simplification, recent advancements, such as increasing efficient work with long context and a focus on semantics, speak of possible "readiness" for NLP to engage in the task. Current evaluation methods for automatic summarisation that may be applicable to literary adaptation include those that involve a multitude of features and rely on a common semantic space, thereby being largely language-independent.

6. Related NLP Tasks: Style Transfer

6.1 Definition

The term "style transfer" or "style adaptation" may be used in relation to different modalities, such as image, audio and textual data. For the purpose of this survey, we will be addressing solely textual style transfer i.e. an NLP task that consists in altering the style of a text whilst preserving its semantic content. The definition of textual style is largely open to interpretation, and the task may so much as be taken as an umbrella term that includes the aforementioned tasks of simplification and summarisation. Tikhonov and Yamshchikov (2018) note that style needs to be "orthogonal" in relation to semantics in the sense that any semantic quality of a text can be expressed in any defined style. While admitting subjectivity,

9 news summaries in English
10 blog articles in Chinese
11 a multilingual knowledge base

Xu (2017) refers to the following seven key styles: "simple and short", "instructional and robotic", "historical and evolving", "colloquial and internet", "gendered and personalised", "pervasive and framing" and "polite and abusive". The most commonly approached dichotomies within traditional style transfer tasks are those between formal and informal text and positive and negative sentiment. Additional focus may fall on authorial voice, register and audience adaptation. Applications of style transfer include the generation of personalised texts and consistent dialogue systems. Style transfer may also be used within other NLP tasks, such as data augmentation. In their extensive survey on the topic of textual style transfer, Jin et al. (2022) note that the rising interest in style within NLP is of high significance, as it implies unprecedented user-centredness.

6.2 Timeline

As noticeable in Fig.2, the practice and interest in style transfer were very scarce during the pre-neural period. The limited number of rule-based studies typically involved the application of hand-crafted grammars and templates. Reddy and Knight (2016) experimented with the alternation of authorial gender through lexical substitution for the purpose of concealing identity in social media communication. Working with the Japanese language, Mizukami et al. (2015, p. 129) defined style transfer as a statistical machine translation task that seeks to "express the individuality of the writer or speaker".

The emergence of neural NLP opened new potential in relation to the examined task. Inspired by recent advances in style transfer within computer vision, Zhang, Zhao, and LeCun (2015) used ConvNets (temporal convolution networks) to deconstruct English and Chinese text at different granularities ranging from character-level to abstract textual features. Addressing sentiment transfer, Li et al. (2018) were among the first to seek to disentangle stylistic from content-related textual attributes in an unsupervised way. They then deleted and added from the original text phrases determined to be distinctive to positive or negative sentiment. Aiming to increase the politeness of chatbot replies, Mukherjee, Hudeček, and Dušek (2023) addressed the unavailability of training data by generating synthetic data via the intermediary of a politeness transfer model.

An example of the widespread application of adversarial networks in style transfer is the work of Zhao et al. (2018). They made use of an adversarially regularized autoencoder (ARAE) that jointly trains a rich discrete-space encoder (e.g. an RNN) and a continuous space generator function, constraining the resulting distributions to be similar. In turn, Fu et al. (2018) defined style as a set of measurable categorial and continuous parameters and used adversarial networks to learn separate content and style representations. They evaluated output based on transfer strength (the extent to which the target style is reflected) and content preservation (the extent to which the original content is retained).

Key training data assembled and/or utilised efficiently within the task of style transfer includes Jhamtani et al. (2017)'s parallel corpus of Shakespearean versus modern language, used to train a "translator" between the two language styles. They used a bidirectional LSTM as an encoder and a mixed RNN and pointer network module that enables copy action as a decoder. Simultaneously undertaking the tasks of automatic translation and style transfer, García et al. (2021) defined both sentiment and language as textual attributes that may be

modified jointly. Surya et al. (2019) also tackled several NLP tasks at applying style-transfer techniques for the purpose of content reduction and lexical simplification. They used an adversarial setup and several auxiliary losses. Addressing this landscape of numerous and varying neural style transfer methods and projects, Tikhonov and Yamshchikov (2018) published an opinion paper that served as a quest for formalisation and standardisation of the task in terms of definition, approaches and evaluation.

Naturally, LLMs brought an array of new approaches to the practice of textual style adaptation. Reif et al. (2022, p. 837) focused on models' zero-shot abilities at a sentence rewriting task, achieving good results in relation to sentiment transfer as well as "transformations such as 'make this melodramatic' or 'insert a metaphor.' " In particular, ACL's 2024 edition was centred around controllable generation and introduced a variety of works that tackle the examined task. Liu et al. (2024) addressed the issues of errors within output and lack of user control in their prompt-based approach for specific-region editing. LLMs were instructed to modify limited portions of text within a defined editing region. Region selection was based on prompt engineering as well as frequency-based strategies. Employing LLMs and chain-of-thought prompting, Han et al. (2024) generated synthetic parallel data for the task of style transfer, on which they then trained a model that performs style-content disentanglement, introducing two losses in order to focus on attribute-related features while constraining a text's semantic space. In their recent work, Huynh and McNamara (2025) interestingly evoked the term "textual personalisation" rather than "style transfer", thus offering an emphasis on the reader's characteristics. They prompted several LLMs to adapt ten scientific texts in accordance with several reader profiles that were defined in terms of age, educational background, reading skills, interests and learning goals. Prompts were augmented using chain-of-thought, personification and RAG techniques. The authors noted that evaluation of the achieved modifications was a rather challenging task. Through automatic metrics, Llama was concluded to produce the most complex and academic-like texts, Gemini - the most diverse ones in terms of both vocabulary and syntax, whilst Claude - the least cohesive ones.

6.3 Evaluation

Evaluation of style transfer output is challenging due to the markedly non-quantitative nature of "style". Jin et al. (2022) suggest that, despite the current technologically advanced landscape, the task necessitates both human-based and automatic evaluation.

The following qualities are typically evoked in relation to style transfer evaluation: fluency, transfer strength, and content preservation (Fu et al. 2018). The last measure has typically been evaluated with established n-gram-based and semantic-based similarity metrics, such as BLEU, ROUGE, METEOR, and BERTScore. However, there is a general lack of associated reference datasets for the task. Popular parallel style-transfer datasets include GYAFC (Grammarly Yahoo Answers Formality Corpus) for formality transfer (Rao and Tetreault 2018), the Yelp Reviews dataset for sentiment transfer (Shen et al. 2017), the Shakespeare dataset of modern versus Shakespearean style (Xu et al. 2012), and CDS (Corpus of Diverse Styles), which involves a plurality of defined styles (e.g. poetry, Biblical text)

(Krishna, Wieting, and Iyyer 2020). XFORMAL, a formality-based dataset, includes text in several languages, including Portuguese, French, and Italian (Briakou et al. 2021).

Reference-free frameworks that focus on various textual features (such as readability-based ones) have also been applied to style transfer evaluation, an example being Coh-Metrix (McNamara et al. 2014). Perplexity based on large language models has also been utilised as a measure of fluency (Jin et al. 2022).

6.4 Takeaways

Unlike the previously discussed tasks of simplification and summarisation, style transfer implies change of a more subjective and qualitative nature, thus providing a strong link with literary adaptation. In addition, the newly-found focus on the reader as well as many of the textual features evoked as representative of style are directly related to the adaptation of literary texts in view of different audience needs and preferences: reader profiles, authorship, modernisation, and even sentiment polarity (e.g. a more "positive" text may be directed at a younger audience). Extending García et al. (2021)'s reasoning, we could unify a large number of both text-centred and reader-centred textual variables, including even language, under a common term tentatively referred to as "style".

As demonstrated throughout the current section, isolated evaluation metrics and limited in size and breadth datasets cannot easily capture the nature of style transfer as a multifaceted framework. Prompt engineering of state-of-the-art LLMs constitutes a promising direction; however, the technique is associated with low explainability and replicability as well as a lack of clarity in terms of the definition of style. A larger yet carefully categorised selection of independent evaluation metrics would likely be of benefit to the evaluation of literary adaptation as a crossroads of textual styles. Ideally, human evaluation including comprehensive studies of reader experience, would also be included.

7. Related NLP Tasks: Creative Generation

7.1 Definition

Creativity is typically associated with human intelligence and with notions such as novelty, personality, uniqueness, emotion and aesthetics. In her work on Artificial Intelligence and creativity, Boden (1998) differentiates between psychological and historical creativity, where the former is limited to a single person and the latter extends to humanity as a whole. She also defines combinational, exploratory, and transformational creativity, which move incrementally from reuse to actual novelty. Although the notion of creative generation (and any creativity for that matter) might strike as an antithesis of technology, NLP methods have been applied at different steps of the process, and current generative models are capable of producing full texts of different genres and sizes, such as short stories and poems, that largely mimic human creative writing.

7.2 Timeline

In 1950, the Turing Test, whose purpose was to evaluate machines by proxy of their resemblance to humans in linguistic interaction, sparked interest in the potential of technological tools to perform creative tasks, such as storytelling (Turing 1950). The "rule-based" NLP period saw the introduction of simple creative tasks, such as the production of pun-like jokes by JAPE (Joke Analysis and Production Engine) (Binsted 1996). The underlying mechanism included schemas that define the structure of jokes, templates of realisation, lexical resources and estimation of phonological similarity.

Real breakthrough came with Transformer models and sequence-to-sequence text generation in the late 2010s. A state-of-the-art convolutional model of story writing was proposed by Fan, Lewis, and Dauphin (2018). It made use of a hierarchical framework: firstly, a single sentence (referred to as "prompt") that describes a story was generated, based on which further generation was conditioned, guaranteeing consistency and staying on-topic. A dataset of human-made stories and related prompts was used in training the model. A year later, a GPT-2 model outperformed Fan, Lewis, and Dauphin (2018)'s model in the following aspects: order of events, lexical variety and reliance on the provided prompt. Both models were noted to exhibit the shortcoming of excessive repetitiveness. Also in 2019, Trăuşan-Matu claimed that despite recent advances, stories generated by language models still lacked in empathy and human-like creativity. In his survey of NLG and creative generation that covers the years 2016 to 2021, Alsharhan (2022) identified 16 high-quality papers, 75% of which evaluated positively the impact of current technology on creative writing. Shanahan and Clarke (2023) used prompting strategies and temperature fine-tuning of GPT-4 in a task of story writing and evaluated the output qualitatively, reaching the conclusion that LLMs can competitively engage in creative writing, while one should keep in mind their currently high dependence on user prompts. Gómez-Rodríguez and Williams (2023) compared the creative generation abilities of several LLMs to each other as well as to a human benchmark for the same task, applying human evaluation to the derived stories, also coming to mixed conclusions. Despite performing very highly, commercial LLMs were judged not to match human writers in terms of originality. Specifically, humour was concluded to be an emerging ability of LLMs.

The implication of LLMs in creative writing gave rise to questions of ethical and philosophical nature. Referring to Boden's notions of value, novelty and surprise, Franceschelli and Musolesi (2024) claimed that the creativity of LLMs is, by definition and insurmountably, limited in both nature and scope. What about the fair use of LLMs? Should one apply limitations to their ability to impersonate an author's style? Can an LLM own authorship? Do the authors of texts that the model was trained on share in its authorship? Currently, there are publishers and literary contest organisers that ask authors to fill in declarations about potential use of AI in their work, although no global frameworks on the topic have been established (Coeckelbergh 2020).

The use of contemporary AI tools in discrete modular tasks related to creative writing (such as character development and plot outlining) is typically seen as an ethically safe practice and as an example of effective human-machine cooperation. Kreminski and Martens (2022) offer an overview of LLMs' potential to support creative writers in cases of "writer's

block". The tool Story Centaur makes use of LLMs' few-shot abilities at discrete writing tasks, such as the generation of a sentence based on the previous one and a "magic word" provided by the user (Swanson et al. 2021).

Additional NLP tasks that imply an element of creative generation have been considered challenging in terms of implementation and evaluation. Toral and Way (2018, p. 263) referred to the translation of literary texts as "the greatest challenge for [pre-LLM] M[achine] T[ranslation]". At a 2023 shared task on discourse-level literary translation organised by Tencent AI Lab and China Literature Ltd, LLMs were the clear winner based on both human and automatic evaluation (Wang et al. 2023b).

7.3 Evaluation

Fan, Lewis, and Dauphin (2018)'s sequence-to-sequence model made use of measures of perplexity on the test set as well as of prompt ranking accuracy in its evaluation of specific aspects of creative writing, such as "staying on topic". Given the task of creative writing translation, Wang et al. (2023b) opted for document-level SacreBLEU[12] (d-BLEU). The limitations of quantitative evaluation of automatic creative output become more prominent with the technology's advancement and a movement away from lower-level concerns such as those of grammaticality, fluency, and cohesion. Boden (1998) notes that evaluation of high-level creativity is largely impossible, as the implied method would, by definition, need to be external to previously established frameworks in order to be applicable to true novelty. The following are examples of human-evaluation-based measures that have been associated with automatic creative generation: characterisation, imagery, humour, and use of idioms. Their quantification and systematisation, however, remains challenging.

7.4 Takeaways

The current quality of automatic tools, notably LLMs, in relation to creative writing, give reasons to be optimistic about NLP's readiness for literary adaptation. However, whilst language proficiency and fluency are unlikely to come as a significant concern, the situation might be different in relation to the knowledge and application of diverse artistic traditions as associated with different cultures.

Key limitations to automatic creative generation, such as ethical concerns related to authorship and the lack of objectivity at evaluation, apply to a lesser extent to the practice of literary adaptation due to the presence of an original text that the output may be compared against. At the same time, additional questions may emerge that require careful contemplation and possibly regulations. For instance: should there be a limit of the extent to which the original text may be reused? Does the original author's copyright come into play?

Due to the presence of an original text, comparison-based evaluation measures such as BLEU score may be relevant to literary adaptation, while likely insufficient on their

12 a standardised implementation of BLEU

own. Elements of creative writing that have been applied in human evaluation of machine-produced text, such as humour, imagery, and idioms, are also to be incorporated. It is worth noting that as semantics is steadily making its way into NLP tools and methods, textual features that were once seen as strictly qualitative, such as the presence of metaphors in text, are currently identifiable and measurable by neural models (Wachowiak, Gromann, and Xu 2022).

8. Discussion

8.1 General Trends

The perceived lagging behind of literary adaptation within NLP in comparison to other tasks may be understandable given its less practical nature and lower sense of urgency compared to, for instance, work such as Segura-Bedmar and Martínez (2017)'s simplification of drug package texts. However, in terms of technological readiness, automated literary adaptation can be accommodated as of today. Common NLP trends of relevance include the strong linguistic, including multilingual, performance of LLMs, a growing focus on style and qualitative textual characteristics, and reader-centredness. One may also notice that as technology advances, user groups tend to be more narrowly defined, an example being readers with different language proficiency levels. Also, associations of two or more discrete tasks (such as translation and style transfer) become more common, and literary adaptation in itself is, as discussed, a meeting point of several practices that are prone to automation, such as simplification, summarisation, style change and literary production. Finally, although the current success of creative text generation is ambiguous, literary adaptation largely mitigates its limitations by being based on a source text.

In terms of the exact generation methods that literary adaptation is most likely to benefit from, LLMs are the first candidate given the current context in NLG, and their performance may be efficiently complemented with established techniques such as prompt engineering and RAG. Additional quantitative and qualitative development within NLP in the near future is likely to continue paving the way for literary adaptation; in particular, in relation to its concerns with textual length, coherence, plot, and figurative language.

8.2 Evaluation Framework

One of the main challenges of automatic literary adaptation is the need to elaborate a relevant evaluation framework. See Appendix A for a table presenting the automatic metrics mentioned within the current survey, along with their broad types, applications and characteristics/benefits. These metrics may be roughly categorised into surface-based, semantic, learned, and reference-free. Surface-level metrics such as BLEU and readability scores are still widely used due to their simplicity combined with the ability to capture key textual characteristics. Although they typically come with significant limitations (such as BLEU's inability to account for paraphrased text), these can be alleviated when multiple metrics are used jointly in carefully selected combinations.

The combination of a number of different automatic metrics can help construct a reusable and adjustable evaluation framework for automatically generated text, in particular in relation to the task of literary adaptation. Additional metrics that we propose to be included in the framework by merit of their relevance to literary adaptation as well as additional tasks addressed in this survey include: Biber (1988)'s comprehensive selection of textual features (including, but not limited to, morphology- and syntax-related ones), literature-specific features, such as figurative language, humour and surprise, as currently measurable by neural tools (e.g. Wachowiak, Gromann, and Xu 2022; Fan et al. 2020) and different aggregators for applicable features, such as average, minimal and maximal values, which have also proved relevant in the description of a text's linguistic qualities (Wilkens et al. 2022).

Many of the mentioned automatic metrics require reference texts. However, reliance on multiple limited datasets in relation to different audience types and languages would be highly inefficient. Instead, we propose the composition of a large literary corpus and the establishment of gold standard values for the various utilised metrics in relation to it. Ideally, the corpus should contain a large number of paired original and professionally adapted versions of the same text, possibly along with additional original texts that account for lower-resource settings, thereby helping include a variety of genres, languages and literary traditions. Based on estimation of feature relevance with respect to different audience types and languages, as well as on correlation analyses that can help efficiently reduce the number of utilised features, different evaluation formulas may be defined for different adaptation scenarios. Finally, the gathered insight about feature relevance may be used not only in the context of evaluation but also in automatic editing of output text or as a protocol integrated into LLM prompts.

8.3 Human Involvement

Currently, despite being costly in terms of time, effort and money, human-based evaluation is the preferred standard in all discussed NLP tasks. The need for substantial human participation is reinforced by concerns of an ethical and philosophical nature in relation to the role of professionals, such as when it comes to creative output and to pedagogical applications.

It is best practice for automated metrics to be validated through their correlation with human evaluation, as demonstrated by Vajjala and Lucic (2019). In the specific case of literary adaptation, a selected reader audience may be asked both to answer comprehension questions, a practice shown to directly measure meaning preservation (Agrawal and Carpuat 2024), and to subjectively evaluate their reading experience and appreciation of the text.

Also, the task of literary adaptation task can realistically be seen as partially rather than fully automatable. Involvement in the process of professionals in fields such as foreign language teaching would be crucial and extend from prompt composition and use to manual editing of automatically produced text.

9. Conclusion

Whilst the independent generation of high-quality creative text by NLP tools is not a current reality, concepts relevant to literature, such as humour, constitute emergent abilities of contemporary models. Given the advanced landscapes in relation to the relevant tasks of automatic text simplification and summarisation as well as style transfer and creative generation, it is not far-fetched to envision the beginning of systematic automated literary adaptation. This survey lays out methods and challenges that the task is likely to be met with and that are to be kept in mind in view of a process of preparation as well as the initial implementation steps of the emerging task.

The immediate steps that we recommend be taken in view of the proposed task include the compilation of a dedicated literary adaptation corpus of a substantial size and the definition of reference-free formulas (similar to readability ones) to be used in the evaluation of adapted literary texts. The corpus should consist of pairs of original and adapted literary texts as annotated according to fine-grained audience categories (e.g. age ranges for children and proficiency levels for foreign language learners). Importantly, a variety of languages should be represented within the corpus. A statistical analysis should then be applied to the corpus, based on a variety of quantitative metrics as laid out in Section 8.2. The most relevant (yet not strongly correlated) metrics should be determined by audience and by language, and corresponding formulas should be formulated for future use. The described framework would pave the way for standardised adaptation effort that extends to a variety of audiences and accounts for a diversity of literary traditions.

10. Limitations

Relevance to literary adaptation is not reserved to the four NLP tasks that we chose to focus on. For instance, machine translation would be heavily implied in cases of different source and target languages. Another task, paraphrase, is also relevant in its transformation of textual formulation whilst meaning is preserved. However, due to the paraphrase's under-specified nature, we have opted for the inclusion of tasks that are associated with concrete communicative objectives that are typically shared with those of literary adaptation: reduction in size (in summarisation), reduction in complexity (in simplification) and change in style to appeal aesthetically to an alternative e.g. more modern audience (style transfer).

Also, we should note that the current survey draws conclusions about the potential of a yet undeveloped NLP task, namely that of literary adaptation, based on related tasks. The nature of the intersection of the described tasks may be hypothesised but is bound to lack certainty. It is therefore important that the proposed guidelines for literary adaptation be factually experimented with. Finally, the mentioned guidelines are of a technical and scientific rather than ethical nature. The ethical aspects and associated potential limitations of automatic literary adaptation, which include for instance questions of access and authorship, have only slightly been touched upon and go beyond the scope of this study.

References

Abdel-Salam, Shehab and Ahmed Rafea. 2022. Performance study on extractive text summarization using BERT models. *Information*, 13(2).

Agrawal, Sweta and Marine Carpuat. 2024. Do text simplification systems preserve meaning? A human evaluation via reading comprehension. *Transactions of the Association for Computational Linguistics*, 12:432–448.

Al-Thanyyan, Suha S. and Aqil M. Azmi. 2021. Automated text simplification: A survey. *ACM Computing Surveys*, 54(2):Article 43, 1–36.

Alsharhan, Abdulla. 2022. Natural language generation and creative writing: A systematic review. *International Journal of Advances in Applied Computational Intelligence*, 1(1):69–90.

Aluísio, Sandra Maria and Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: the PorSimples project for simplification of Portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas (YIWCALA '10), Los Angeles, CA, USA*, page 46–53, Association for Computational Linguistics.

Alva-Manchego, Fernando, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Association for Computational Linguistics.

Alva-Manchego, Fernando, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China: System Demonstrations*, pages 49–54, Association for Computational Linguistics.

Aranzabe, María, Arantza Ilarraza, and Itziar Gonzalez-Dios. 2012. First approach to automatic text simplification in Basque. In *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) Workshop*, pages 1–8.

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, Michigan*, pages 65–72, Association for Computational Linguistics.

Benedetto, Luca and Paula Buttery. 2025. Towards CEFR-targeted text simplification for question adaptation. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing – Natural Language Processing in the Generative AI Era, Varna, Bulgaria*, pages 150–157, INCOMA Ltd., Shoumen, Bulgaria.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge. Reprinted/online publication: 2012.

Binsted, Kim. 1996. *Machine Humour: An Implemented Model of Puns*. Ph.D. thesis, University of Edinburgh.

Biran, Or, Samuel Brody, and Noémie Elhadad. 2011. Putting it simply: A context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501, Association for Computational Linguistics.

Boden, Margaret A. 1998. Creativity and artificial intelligence. *Artificial Intelligence*, 103(1-2):347–356.

Bonifacci, Paola, Cinzia Viroli, Chiara Vassura, Elisa Colombini, and Lorenzo Desideri. 2023. The relationship between mind wandering and reading comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, 30(1):40–59.

Briakou, Eleftheria, Di Lu, Ximing Lu, and Joel Tetreault. 2021. Olá, Bonjour, Salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Association for Computational Linguistics.

Brunato, Dominique, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first Italian corpus for text simplification. In *Proceedings of the 9th Linguistic Annotation Workshop, Denver, Colorado, USA*, pages 31–41, Association for Computational Linguistics.

Caplan, David. 2012. *Language: Structure, Processing, and Disorders.* MIT Press, Cambridge, MA.

Cersosimo, Rita, Filippo Domaneschi, and Alice Cancer. 2025. The impact of metaphors on academic text comprehension: The case of students with dyslexia. *Dyslexia*, 31(1).

Chang, Yapei, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. BooookScore: A systematic exploration of book-length summarization in the era of LLMs. ArXiv, abs/2310.00785.

Chen, Han-Bin, Hen-Hsen Huang, Hsin-Hsi Chen, and Ching-Ting Tan. 2012. A simplification-translation-restoration framework for cross-domain SMT applications. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), Mumbai, India*, pages 545–560.

Chen, Kai, Rui Wang, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2019. Improving transformer-based neural machine translation with recurrent and convolutional components. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 195–206.

Cheng, Jianpeng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany (Volume 1: Long Papers)*, pages 484–494, Association for Computational Linguistics.

Chopra, Sumit, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, Association for Computational Linguistics.

Coeckelbergh, Mark. 2020. *AI Ethics.* MIT Press, Cambridge, MA.

Coster, William and David Kauchak. 2011a. Learning to simplify sentences using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation, Portland, Oregon, USA*, pages 1–9, Association for Computational Linguistics.

Coster, William and David Kauchak. 2011b. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Association for Computational Linguistics.

Cripwell, Liam, Joël Legrand, and Claire Gardent. 2023. Document-level planning for text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia*, pages 993–1006, Association for Computational Linguistics.

Day, Richard R. and Julian Bamford. 1998. *Extensive Reading in the Second Language Classroom.* Cambridge University Press, Cambridge.

Devlin, Siobhan. 1999. *Simplifying Natural Language Text for Aphasic Readers.* Ph.D. thesis, University of Sunderland, Sunderland, United Kingdom.

Devlin, Siobhan and John Tait. The use of a psycholinguistic database in the simplification of text for aphasic readers. In J. Nerbonne, editor, *Linguistic Databases.* CSLI Publications, Stanford, California, pages 161–173.

Devlin, Siobhan and Gary Unthank. 2006. Helping aphasic people process online information. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '06), Portland, Oregon, USA*, pages 225–226, ACM.

Dickey, Michael Walsh and Cynthia K. Thompson. 2009. Automatic processing of wh- and NP-movement in agrammatic aphasia: Evidence from eyetracking. *Journal of Neurolinguistics*, 22(5):563–584.

Dickinson, Paul. 2017. Effects of extensive reading on EFL learner reading attitudes. In *Proceedings of the 21st Conference of the Pan-Pacific Association of Applied Linguistics, Tamkang University, Taiwan*, pages 28–35.

Dmitrieva, Anna and Jörg Tiedemann. 2021. Creating an aligned Russian text simplification dataset from language learner data. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing, Kiyv, Ukraine*, pages 73–79, Association for Computational Linguistics.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the 2nd International Conference on Human Language Technology Research (HLT 2002)*, pages 138–145, Morgan Kaufmann.

Douma, Willem Hendrik. 1960. *De Leesbaarheid van Landbouwbladen: Een Onderzoek naar en een Toepassing van Leesbaarheidsformules.* Veenman & Zonen, Wageningen, Netherlands.

DuBay, William H. 2004. *The Principles of Readability.* Impact Information, Costa Mesa, CA, USA.

El-Kassas, Wafaa S., Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Ellis, John. 1982. The literary adaptation. *Screen*, 23(3):3–5.

Evans, Richard, Constantin Orăsan, and Iustin Dornescu. 2014. An evaluation of syntactic simplification rules for people with autism. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014) Workshop on Language Technology for Closely Related Languages and Language Variants*, pages 131–140, Association for Computational Linguistics.

Ewers, Hans-Heino. 2009. *Fundamental Concepts of Children's Literature Research: Literary and Sociological Approaches.* Routledge, London.

Fan, Angela, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia (Volume 1: Long Papers)*, pages 889–898, Association for Computational Linguistics.

Fan, Xiaochao, Hongfei Lin, Liang Yang, Yufeng Diao, Chen Shen, Yonghe Chu, and Yanbo Zou. 2020. Humor detection via an internal and external neural network. *Neurocomputing*, 394:105–111.

Farajidizaji, Asma, Vatsal Raina, and Mark Gales. 2024. Is it possible to modify text to a target readability level? An initial investigation using zero-shot large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italia*, pages 9325–9339, ELRA and ICCL.

Feng, Yutao, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. Sentence simplification via large language models. ArXiv, abs/2302.11957.

FitzGerald, Elizabeth, Ann Jones, Natalia Kucirkova, and Eileen Scanlon. 2018. A literature synthesis of personalised technology-enhanced learning: What works and why. *Research in Learning Technology*, 26:2095.

Franceschelli, Giorgio and Mirco Musolesi. 2024. On the creativity of large language models. *AI & Society*, 40(5):3785–3795.

Freyer, Nils, Hendrik Kempt, and Lars Klöser. 2024. Easy-Read and large language models: On the ethical dimensions of LLM-based text simplification. *Ethics and Information Technology*, 26(3):1–10.

Fu, Zhenxin, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: exploration and evaluation. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18)*, AAAI Press.

Gala, Núria, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C. Ziegler. 2020. Alector: A parallel corpus of simplified French texts with alignments of misreadings by poor and dyslexic readers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France*, pages 1353–1361, European Language Resources Association.

García, Xavier, Noah Constant, Mandy Guo, and Orhan Firat. 2021. Towards universality in multilingual text rewriting. ArXiv, abs/2107.14749.

Genest, Pierre-Etienne and Guy Lapalme. 2012. Fully abstractive approach to guided summarization. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju Island, Korea (Volume 2: Short Papers)*, pages 354–358, Association for Computational Linguistics.

Glavaš, Goran and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China (Volume 2: Short Papers)*, pages 63–68, Association for Computational Linguistics.

Godwin-Jones, Robert. 2018. Contextualized vocabulary learning. *Language Learning & Technology*, 22(3):1–19.

Gómez-Rodríguez, Carlos and Paul Williams. 2023. A confederacy of models: A comprehensive evaluation of LLMs on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore*, pages 14504–14528, Association for Computational Linguistics.

Han, Jingxuan, Quan Wang, Zikang Guo, Benfeng Xu, Licheng Zhang, and Zhendong Mao. 2024. Disentangled learning with synthetic parallel data for text style transfer. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand (Volume 1: Long Papers)*, pages 15187–15201, Association for Computational Linguistics.

Happé, Francesca G. E. 1993. Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition*, 48(2):101–119.

Haverals, Wouter, Lindsey Geybels, and Vanessa Joosen. 2022. A style for every age: A stylometric inquiry into crosswriters for children, adolescents and adults. *Language and Literature*, 31(1):62–84.

Hill, David R. 2013. Graded readers. *ELT Journal*, 67(1):85–125.

Hovy, Eduard, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. 2006. Automated summarization evaluation with basic elements. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy*, European Language Resources Association (ELRA).

Hu, Baotian, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale Chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1967–1972.

Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, Association for Computing Machinery, New York, NY, USA.

Huang, Lei, Yanxiang He, Furu Wei, and Wenjie Li. 2010. Modeling document summarization as multi-objective optimization. In *Proceedings of the Third International Symposium on Intelligent Information Technology and Security Informatics*, pages 382–386.

Hutcheon, Linda. 2006. *A Theory of Adaptation*. Routledge, New York.

Huynh, Linh and Danielle S. McNamara. 2025. GenAI-powered text personalization: Natural language processing validation of adaptation capabilities. *Applied Sciences*, 15(12).

Intrigi.bg. 2024. Adaptatsiyata na Rusi Chanev na "Pod igoto" predizvika skandal. Published: December 7, 2024.

Jain, Sameer, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-dimensional evaluation of text summarization with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada*, pages 8487–8495, Association for Computational Linguistics.

Jamet, Henri, Yash Raj Shrestha, and Michalis Vlachos. 2024. Difficulty estimation and simplification of French text using LLMs. In *Generative Intelligence and Intelligent Tutoring Systems. Proceedings of the 20th International Conference ITS 2024, Thessaloniki, Greece*, pages 395–404, Springer Nature Switzerland.

Jhamtani, Harsh, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation, Copenhagen, Denmark*, pages 10–19, Association for Computational Linguistics.

Jin, Di, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Kalandadze, Tamar, Courtenay Frazier Norbury, Terje Nærland, and Kari-Anne B. Næss. 2018. Figurative language comprehension in individuals with autism spectrum disorder: A meta-analytic review. *Autism*, 22(2):99–117.

Kazantseva, Anna and Stan Szpakowicz. 2010. Summarizing short stories. *Computational Linguistics*, 36(1):71–109.

Kew, Tannon, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore*, pages 13291–13309, Association for Computational Linguistics.

Klaper, David, Sarah Ebling, and Martin Volk. 2013. Building a German/Simple German parallel corpus for automatic text simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2013), Sofia, Bulgaria*, pages 11–19, Association for Computational Linguistics.

Ko, Youngjoong and Jungyun Seo. 2008. An effective sentence-extraction technique using contextual information and statistical approaches for text summarization. *Pattern Recognition Letters*, 29(9):1366–1371.

Kobayashi, Hayato, Masaki Noguchi, and Taichi Yatsuka. 2015. Summarization based on embedding distributions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language*

Processing, Lisbon, Portugal, pages 1984–1989, Association for Computational Linguistics.

Krashen, Stephen D. 1982. *Principles and Practice in Second Language Acquisition*. Pergamon Press, Oxford.

Kreminski, Max and Chris Martens. 2022. Unmet creativity support needs in computationally supported creative writing. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022), Dublin, Ireland*, pages 74–82, Association for Computational Linguistics.

Krishna, Kalpesh, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Ladhak, Faisal, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048.

Lefebvre, Julie. 2021. "Adaptation" ou "texte intégral" ? Représentations éditoriales de l'enfant lecteur. *Le Français Aujourd'hui*, 213(2):53–64.

Li, Juncen, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874, Association for Computational Linguistics.

Lin, Chin-Yew. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Association for Computational Linguistics.

Liu, Pusheng, Lianwei Wu, Linyong Wang, Sensen Guo, and Yang Liu. 2024. Step-by-step: Controlling arbitrary style in text with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italia*, pages 15285–15295, ELRA and ICCL.

Liu, Yan. 2021. Readability and adaptation of children's literary works from the perspective of ideational grammatical metaphor. *Journal of World Languages*, 7(2):334–354.

Luhn, Hans Peter. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Maddela, Mounica, Yao Dou, David Heineman, and Wei Xu. 2022. LENS: A learnable evaluation metric for text simplification. ArXiv, abs/2212.09739.

Mar, Raymond A., Jingyuan Li, Anh T. P. Nguyen, and Cindy P. Ta. 2021. Memory and comprehension of narrative versus expository texts: A meta-analysis. *Psychonomic Bulletin & Review*, 28(3):732–749.

Mar, Raymond A., Keith Oatley, Jacob B. Hirsh, Jennifer dela Paz, and Jordan B. Peterson. 2006. Bookworms versus nerds: Exposure to fiction versus non-fiction, divergent associations with social ability, and the simulation of fictional social worlds. *Journal of Research in Personality*, 40(5):694–712.

Martinussen, Rhonda, Josephine Hayden, Sheilah Hogg-Johnson, and Rosemary Tannock. 2005. A meta-analysis of working memory impairments in children with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, 44(4):377–384.

Mashal, Nira and Anat Kasirer. 2012. Principal component analysis study of visual and verbal metaphoric comprehension in children with autism and learning disabilities. *Research in Developmental Disabilities*, 33(1):274–282.

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

Melogno, Sergio, Maria A. Pinto, and Margherita Orsolini. 2017. Novel metaphors comprehension in a child with high-functioning autism spectrum disorder: A study on assessment and treatment. *Frontiers in Psychology*, 7:2004.

Mizukami, Masahiro, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2015. Linguistic individuality transformation for spoken language. In G.G. Lee, H.K. Kim, M. Jeong, and J.-H. Kim, editors, *Natural Language Dialog Systems and Intelligent Assistants*. Springer International Publishing, Cham, pages 129–143.

Mukherjee, Sourabrata, Vojtěch Hudeček, and Ondřej Dušek. 2023. Polite chatbot: A text style transfer application. In *Proceedings of the 17th Conference of the European Chapter of the Association for*

Computational Linguistics, Dubrovnik, Croatia: Student Research Workshop, pages 87–93, Association for Computational Linguistics.

Müller, Anja, editor. 2013. *Adapting Canonical Texts in Children's Literature.* Bloomsbury Academic, London.

Napoles, Courtney, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)@NAACL-HLT 2012, Montrèal, Canada*, pages 95–100.

Nation, I. S. Paul. 2001. *Learning Vocabulary in Another Language.* Cambridge University Press, Cambridge.

National Institute of Standards and Technology. 2004. Document Understanding Conference (DUC). https://duc.nist.gov.

Nenkova, Ani, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, 4(2):4–es.

Nières-Chevrel, Isabelle. 2009. *Introduction à la littérature de jeunesse.* Didier Jeunesse, Paris.

Nisioi, Sergiu, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada (Volume 2: Short Papers)*, pages 85–91, Association for Computational Linguistics.

Nussbaum, Martha Craven. 2010. *Not for Profit: Why Democracy Needs the Humanities.* Princeton University Press, Princeton, NJ.

Oatley, Keith. 2011. *Such Stuff as Dreams: The Psychology of Fiction.* Wiley-Blackwell, Chichester, UK.

van Oosten, Philip, Dieter Tanghe, and Véronique Hoste. 2010. Towards an improved methodology for automated readability prediction. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta*, European Language Resources Association (ELRA).

Ormaechea, Lucía and Nikos Tsourakis. 2024. Automatic text simplification for French: Model fine-tuning for simplicity assessment and simpler text generation. *International Journal of Speech Technology*, 27:957–976.

Ouyang, You, Wenjie Li, Sujian Li, and Qin Lu. 2011. Applying regression models to query-focused multi-document summarization. *Information Processing & Management*, 47(2):227–237.

Owczarzak, Karolina. 2009. DEPEVAL(summ): Dependency-based evaluation for automatic summaries. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Suntec, Singapore*, pages 190–198.

Paetzold, Gustavo and Lucia Specia. 2016. SV000gg at SemEval-2016 task 11: Heavy gauge complex word identification with system voting. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California*, pages 969–974, Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA*, pages 311–318, Association for Computational Linguistics.

Pu, Xiao, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. ArXiv, abs/2309.09558.

Radev, Dragomir R., Eduard Hovy, and Kathleen McKeown. 2002. Introduction to the special issue on summarization. *Computational Linguistics*, 28(4):399–408.

Rao, Sudha and Joel Tetreault. 2018. Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 129–140, Association for Computational Linguistics.

Reddy, Sravana and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science, Austin, Texas*, pages 17–26, Association for Computational Linguistics.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Association for Computational Linguistics.

Reif, Emily, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland (Volume 2: Short Papers)*, pages 837–848, Association for Computational Linguistics.

Restrepo Ramos, Fabian Dario. 2015. Incidental vocabulary learning in second language acquisition: A literature review. *PROFILE Issues in Teachers' Professional Development*, 17(1):157–166.

Rios, Annette, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A new dataset and efficient baselines for document-level text simplification in German. In *Proceedings of the Third Workshop on New Frontiers in Summarization, Dominican Republic*, pages 152–161, Association for Computational Linguistics.

Wan-a rom, Uthai. 2008. Comparing the vocabulary of different graded-reading schemes. *Reading in a Foreign Language*, 20(1):43–69.

Rutherford, Marion, Julie Baxter, Zoe Grayson, Lorna Johnston, and Anne O'Hare. 2020. Visual supports at home and in the community for individuals with autism spectrum disorder. *Autism*, 24(2):447–457.

Ryan, Michael J., Tarek Naous, and Wei Xu. 2023. Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada (Volume 1: Long Papers)*, pages 4898–4927, Association for Computational Linguistics.

Scialom, Thomas, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021a. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic*, pages 6594–6604, Association for Computational Linguistics.

Scialom, Thomas, Louis Martin, Jacopo Staiano, Éric Villemonte de la Clergerie, and Benoît Sagot. 2021b. Rethinking automatic evaluation in sentence simplification. ArXiv, abs/2104.07560.

Segura-Bedmar, Isabel and Paloma Martínez. 2017. Simplifying drug package leaflets written in spanish by using word embedding. *Journal of Biomedical Semantics*, 8(1):45.

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Association for Computational Linguistics.

Shanahan, Murray and Catherine A. M. Clarke. 2023. Evaluating large language model creativity from a literary perspective. ArXiv, abs/2312.03746.

Shardlow, Matthew, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024. The BEA 2024 Shared Task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), Mexico City, Mexico*, pages 571–589, Association for Computational Linguistics.

Shavit, Zohar. 1999. The double attribution of texts for children and how it affects writing for children. In Sandra L. Beckett, editor, *Transcending Boundaries: Writing for a Dual Audience of Children and Adults*. Garland Publishing, New York and London, pages 83–98.

Shen, Tianxiao, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*.

Shetty, Krithi and Jagadish S. Kallimani. 2017. Automatic extractive text summarization using k-means clustering. In *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, pages 1–9.

Snover, Matthew, Nitin Madnani, Bonnie J. Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? exploring different human judgments with a tunable MT metric. In *Proceedings of the*

Fourth Workshop on Statistical Machine Translation (WMT 2009), Athens, Greece, pages 259–268, Association for Computational Linguistics.

Snowling, Margaret J., Charles Hulme, and Kate Nation. 2000. Defining and understanding dyslexia: Past, present and future. *Oxford Review of Education*, 46(4):501–513.

Spaulding, Seth. 1956. A spanish readability formula. *The Modern Language Journal*, 40(7):433–441.

Stajner, Sanja and Horacio Saggion. 2013. Adapting text simplification decisions to different text genres and target users. *Procesamiento del Lenguaje Natural*, 51.

Štajner, Sanja and Horacio Saggion. 2013. Readability indices for automatic evaluation of text simplification systems: A feasibility study for Spanish. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing, Nagoya, Japan*, pages 374–382, Asian Federation of Natural Language Processing.

Stodden, Regina. 2024. EASSE-DE & EASSE-multi: Easier automatic sentence simplification evaluation for german & multiple languages. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), Miami, Florida, USA*, pages 107–116, Association for Computational Linguistics.

Stymne, Sara, Jörg Tiedemann, Christian Hardmeier, and Joakim Nivre. 2013. Statistical machine translation with readability constraints. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), Oslo, Norway*, pages 375–386, Linköping University Electronic Press.

Sulem, Elior, Omri Abend, and Ari Rappoport. 2018. Semantic structural evaluation for text simplification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, Volume 1 (Long Papers)*, pages 685–696, Association for Computational Linguistics.

Surya, Sai, Abhijit Mishra, Anirban Laha, Parag Jain, and Karthik Sankaranarayanan. 2019. Unsupervised neural text simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy*, pages 2058–2068, Association for Computational Linguistics.

Swanson, Ben, Kory Mathewson, Ben Pietrzak, Sherol Chen, and Monica Dinalescu. 2021. Story Centaur: Large language model few shot learning as a creative writing tool. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 244–256, Association for Computational Linguistics.

Thiel, Elizabeth. 2013. Downsizing Dickens: Adaptations of Oliver Twist for the child reader. In Anja Müller, editor, *Adapting Canonical Texts in Children's Literature*. Bloomsbury Academic, London, pages 143–162.

Tikhonov, Alexey and Ivan P. Yamshchikov. 2018. What is wrong with style transfer for texts? ArXiv, abs/1808.04365.

Toral, Antonio and Andy Way. 2018. What level of quality can neural machine translation attain on literary text? In *Translation Quality Assessment: From Principles to Practice*. Springer, pages 263–287.

Tratz, Stephen and Eduard Hovy. 2008. Summarization evaluation using transformed basic elements. In *Proceedings of the First Text Analysis Conference (TAC 2008), National Institute of Standards and Technology Gaithersburg, Maryland, USA*.

Trăuşan-Matu, Ştefan. 2019. Computer-based story generation: An analysis from a phenomenological standpoint. *International Journal of User-System Interaction*, 12(1):39–53.

Turing, Alan. 1950. Computing machinery and intelligence. *Mind*, 59(236):433–460.

Vajjala, Sowmya and Ivana Lučić. 2018. OneStopEnglish Corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, New Orleans, Louisiana*, pages 297–304, Association for Computational Linguistics.

Vajjala, Sowmya and Ivana Lucic. 2019. On understanding the relation between expert annotations of text readability and target reader comprehension. In *Proceedings of the Fourteenth Workshop on Innovative Use of Natural Language Processing for Building Educational Applications, Florence, Italy*, pages 349–359, Association for Computational Linguistics.

Van Lierop-Debrauwer, Helma. 2000. Over de 'Grote Gelijkenis': adolescentenromans voor jongeren en voor volwassenen. *Literatuur Zonder Leeftijd*, 14:336–351.

Vazov, Ivan Minchov. 2024. *Pod igoto.* Chanev, Rusi (ed.). Iztok-Zapad, Sofia. Adapted edition.

Wachowiak, Lennart, Dagmar Gromann, and Chao Xu. 2022. Drum up support: Systematic analysis of image-schematic conceptual metaphors. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 44–53.

Wang, Jiaan, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023a. Zero-shot cross-lingual summarization via large language models. In *Proceedings of the 4th Workshop on New Frontiers in Summarization (NewSum 2023), Toronto, Canada*, pages 12–23, Association for Computational Linguistics.

Wang, Longyue, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023b. Findings of the WMT 2023 Shared Task on discourse-level literary translation: A fresh orb in the cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine Translation (WMT 2023)*.

Wang, Yingjie and Jun Ma. 2013. A comprehensive method for text summarization based on latent semantic analysis. In *Proceedings of the Natural Language Processing and Chinese Computing: Second CCF Conference (NLPCC 2013), Chongqing, China*, pages 394–401, Springer, Berlin, Heidelberg.

Washburne, Carleton and Mabel Vogel. 1926. *Winnetka Graded Book List.* American Library Association, Chicago, Illinois.

Wilkens, Rodrigo, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin P. Yancey, and Thomas François. 2022. FABRA: French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France*, pages 1217–1233, European Language Resources Association.

Wu, Yunshu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2024. Less is more for long document summary evaluation by LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, St. Julian's, Malta (Volume 2: Short Papers)*, pages 330–343, Association for Computational Linguistics.

Xiao, Le and Xiaolin Chen. 2023. Enhancing LLM with evolutionary fine tuning for news summary generation. ArXiv, abs/2307.02839.

Xu, Wei. 2017. From Shakespeare to Twitter: What are language styles all about? In *Proceedings of the Workshop on Stylistic Variation, Copenhagen, Denmark*, pages 1–9, Association for Computational Linguistics.

Xu, Wei, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Xu, Wei, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Xu, Wei, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), Mumbai, India*, pages 2899–2914.

Yang, Ziyu. 2024. *Enhancing the Comprehension: Text Simplification Approaches and the Role of Large Language Models.* Phd dissertation, Temple University.

Yuan, Weizhe, Graham Neubig, and Pengfei Liu. 2021. BARTSCORE: evaluating generated text as text generation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Curran Associates Inc., Red Hook, NY, USA.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*, OpenReview.net.

Zhang, Tianyi, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Zhang, Xiang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, MIT Press, Cambridge, MA, USA.

Zhang, Xingxing and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark*, pages 584–594, Association for Computational Linguistics.

Zhang, Yang, Hanlei Jin, Dan Meng, Jun Wang, and Jinghua Tan. 2026. A comprehensive survey on automatic text summarization with exploration of LLM-based methods. *Neurocomputing*, 663:131928.

Zhao, Junbo (Jake), Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. 2018. Adversarially regularized autoencoders. ArXiv, abs/1706.04223.

Appendix

A. Evaluation Metrics

Metric	Type	Application	Benefits/Characteristics
BLEU	n-gram overlap	summarisation, simplification, style transfer, creative generation	widely used; good for precision; simple to compute; multiple variants
SacreBLEU	n-gram overlap	summarisation, simplification, creative generation	reproducible; standardised;
ROUGE	n-gram overlap	summarisation, simplification, style transfer	good for recall; widely used; multiple variants
NIST	weighted n-gram overlap	simplification	accounts for informativeness; improves over BLEU
TERp	edit distance	simplification	interpretable; reference-based
METEOR	n-gram + semantic matching	summarisation, simplification, style transfer	accounts for synonyms; considers word order; better correlation than BLEU
SARI	operation-based (add/delete/keep)	simplification	task-specific; captures simplicity; interpretable
SAMSA	semantic structure	simplification	reference-free; captures semantics; focuses on predicate-argument relations
BERTScore	semantic similarity (embedding-based)	summarisation, simplification, style transfer	captures semantics; robust to paraphrasing; widely used

Metric	Type	Application	Benefits/Characteristics
COMET	learned metric	simplification	models human judgement; multilingual; uses source and reference
BLEURT	learned metric (BERT-based)	summarisation	models human judgement; robust with limited references
BARTScore	LLM-based (seq2seq scoring)	summarisation	correlates with fluency; generation-based evaluation
QuestEval	QA-based	summarisation, simplification	evaluates factual consistency; reference-free; correlates with human judgement
LENS	learned metric	simplification	correlates with fluency, meaning, simplicity; high correlation with humans
FKBLEU	hybrid (BLEU + readability)	simplification	combines similarity and simplicity; uses source and reference
Flesch Reading Ease	readability	simplification, style transfer	simple to compute; reference-free
Flesch–Kincaid Grade Level (FKGL)	readability	simplification, style transfer	interpretable; grade-level estimate; widely used
Perplexity	language model-based	style transfer, creative generation	correlates with fluency; reference-free

Table 1
Automatic evaluation metrics mentioned within the survey as relevant to the NLP tasks of simplification, summarisation, style transfer and creative generation. Individual shortcomings are not focused on, as the purpose is to provide an overview of different methods' strengths that allow for their effective use when applied in combinations rather than independently.

Readability Criteria for Bulgarian: Lexical and Grammatical Dimensions

Ivelina Stoyanova
Institute for Bulgarian Language
Bulgarian Academy of Sciences
Sofia, Bulgaria
iva@dcl.bas.bg

This paper presents an empirical investigation into text readability for Bulgarian, with a focus on the lexical and grammatical features that predict word-level and sentence-level reading difficulty for primary-school children. We review established readability indices such as Flesch, LIX, Coleman-Liau, and discuss their limitations when applied to a morphologically rich Slavic language such as Bulgarian. We then propose a feature-based approach to readability assessment grounded in the actual reading behaviour of Bulgarian learners.

The empirical work is based on a corpus of finger-tracking reading data from 83 children across grades 2–5, reading age-appropriate narrative texts both aloud and silently. Each word is classified as easy, average, or difficult based on its letters-per-second speed distribution, and we examine the relationship between this difficulty classification and a comprehensive set of phonological, morphological, syntactic, and textual features.

Results identify a small set of strong and weak predictors of readability level. Word length, part-of-speech class (especially the content-vs-function-word distinction) are strong predictors, as well as the sentence-initial position and the position of clause and phrase boundaries, which are consistent across all grades. Repetition within the text reduces the processing cost of difficult content words with subsequent occurrences. Morphological complexity and stem alternation, as well as dependency relations have weak contribution to readability. We report specific findings on pronoun subtypes, on verb-initial and subject-initial sentences, and on POS-bigram syntactic structures.

The analysis also focuses on grade-by-grade comparison, as well as the differences between reading aloud and silent reading.

Keywords: readability, Bulgarian, lexical complexity, morphosyntax

1. Introduction

Readability is a multidimensional property of the written text that determines how easily a given reader can decode and comprehend the text. It depends jointly on properties of the text itself – its phonological, lexical, morphological, and syntactic complexity, and on the characteristics and abilities of the reader, including age, prior exposure to reading, vocabulary knowledge, and the degree to which decoding has become automatised. Good

https://doi.org/10.47810/JCLIB.2.2026.05

reading and text comprehension skills are key competences and an essential prerequisite for high-quality education. Reading fluency is a strong predictor of performance across all literacy-based subjects, with reading speed being one of the most robust indicators of overall reading proficiency. In the long term, students with early reading difficulties face serious challenges in learning, academic performance, and social integration.

Existing readability indices, in turn, were developed primarily for a particular language, and rely on surface measures such as average sentence length and average word length that are highly language-dependent and do not straightforwardly transfer to morphologically rich languages such as Bulgarian, thus the direct application of established readability formulae is not suitable.

Bulgarian has a relatively transparent orthography, with largely predictable grapheme–phoneme correspondences. Transparency facilitates early mastery of reading. However, Bulgarian exhibits a rich inflectional and derivational morphology, postpositional definite article realised as an enclitic, and relatively free word order – properties that influence readability in ways that surface formulae cannot capture.

The present work is part of the efforts to address this gap by developing a Bulgarian-specific, empirically grounded set of readability criteria, based on reading data collected from primary school children from grade 2 to grade 5.

Reading data were collected using the finger-tracking technique implemented in the *ReadLet* infrastructure (Taxitari et al. 2021; Crepaldi et al. 2022; Nadalini et al. 2023), a tablet-based platform that records the continuous movements of a reader's finger across a connected text, together with the voice signal in reading aloud sessions. Finger-tracking provides a fine-grained, temporally sensitive measure of reading dynamics at the word level, validated against eye-tracking in both adult and developing readers (Crepaldi et al. 2022; Nadalini et al. 2023), and is portable and classroom-friendly enough to be deployed at scale in primary schools (Lento et al. 2024; Marzi et al. 2026).

Crucially, by aligning per-word tracking times with linguistic annotations of the reading texts, we are in a position to relate the time a child takes to read each word to the phonological, lexical, morphological, and syntactic properties of that word, and to track how this relationship evolves across grades. The data analysed in the present study were collected within the bilateral project *Assessing reading literacy and comprehension of early graders in Bulgaria and Italy* (2023–2025),[1] jointly funded by the Bulgarian Academy of Sciences and the Italian National Research Council. Reading materials were originally authored in Italian and translated into Bulgarian, with sentence number and sentence type, age-appropriate vocabulary, and culturally neutral narrative content in similar proportions across the two languages (Pirrelli and Koeva 2024). Children read one text aloud and another text silently, on a tablet running the ReadLet application, in their regular classroom setting. The two texts are in the same age category. The present paper focuses exclusively on the Bulgarian data, with a view to identifying the text-internal features that determine word-, phrase-, and sentence-level reading difficulty and to laying the groundwork for a language-specific readability evaluation.

1 https://dcl.bas.bg/en/literacy/

The present study focuses on decoding speed as measured by per-word reading times, on the assumption that fluent decoding is a necessary component of, and a reliable proxy for, overall reading proficiency in early primary grades. Comprehension questions administered after each reading episode were recorded but fall outside the scope of the present paper. In Section 8 we briefly discuss the ways in which fast decoding can dissociate from full understanding of the text.

The objective of the study is to use reading speed data to derive a set of evidence-based readability criteria for Bulgarian texts. While established readability formulae such as Flesch, LIX, and Coleman–Liau, operate on a small number of surface variables and yield a single text-level score, the present approach grounds readability in observed word-level and sentence-level reading patterns and exploits a substantially richer feature inventory that includes phonological, morphological, syntactic, and textual properties. Identifying which of these features contribute robustly across all grades, and which are diagnostic only for beginner readers, is a necessary step towards a complex, developmentally sensitive readability index for Bulgarian. The design and validation of such an index is left to future work.

Building on the empirical data, we address the following research questions:

RQ1. Which phonological, lexical, morphological, and syntactic features of Bulgarian most strongly predict word-level and sentence-level reading speed in primary school children?

RQ2. Which features remain stable in importance as reading skills develop, and which lose (or gain) weight as decoding becomes increasingly automatised?

RQ3. Do the same features predict reading speed in reading aloud and in silent reading, or do the two modalities differ, and is that consistent across grades? Do phonological and articulatory features weigh more heavily in reading aloud, and are features equally pronounced?

The remainder of the paper is organised as follows. Section 2 reviews related work on readability assessment and on empirical finger-tracking studies of reading development, with Section 2.3 presenting the established readability assessment frameworks and their limitations for Bulgarian. Section 3 describes the dataset and the automatic linguistic analysis pipeline applied to the reading texts. Section 4 discusses the word-level and sentence-level feature inventory, including length of the word in number of letters and number of syllables, part of speech, with focus on function vs. content words, morphological complexity, syntactic structure, position of the word in the sentence. Section 5 presents the key findings that are directly relevant to the development of a readability index for Bulgarian. Section 6 draws practical recommendations for reading instruction and text design, Section 7 concludes and outlines directions for future work, and Section 8 briefly outlines the limitations of the work presented in the paper.

2. Related Works

Research relevant to a language-specific readability evaluation for Bulgarian spans across several directions. The first concerns the empirical methods used to observe reading be-

haviour, from eye-tracking studies to finger-tracking patterns among primary-school children. The literature also focuses on the linguistic features that most reliably modulate reading speed and comprehension – word length and frequency, morphological structure, part-of-speech class, and syntactic complexity. The developmental trajectory along which each of these features gains or loses weight with age and skills is relevant to establishing reliable criteria for readability. This section aims to outline these directions in the literature and to identify the limitations of established readability indices when applied to Bulgarian.

2.1 Works on Assessment of Reading Ability

Most work on the temporal dynamics of reading has relied on eye-tracking, which has established robust effects of word length, word frequency, and predictability on fixation durations and patterns (Rayner 1998, 2009). Eye-tracking, however, requires laboratory equipment and is difficult to deploy at the scale needed for developmental and educational research. Finger-tracking has recently emerged as a portable, classroom-friendly alternative that records the continuous movements of a reader's index finger across a connected text on a tablet screen. The technique has been validated against eye-tracking in both adult and developing readers, with finger movements showing strong correlations with fixation durations, saccade patterns, and articulation timing (Crepaldi et al. 2022; Nadalini et al. 2023). The data analysed in the present paper were collected using the *ReadLet* infrastructure (Taxitari et al. 2021; Ferro et al. 2018), a tablet-based platform purpose-built for large-scale, multimodal reading data collection in primary school settings.

Reading acquisition is a multidimensional developmental process that depends not only on the ability to decode written symbols into sounds, but also on efficient lexical access, working memory, attentional control, and the gradual automatisation of orthographic patterns (Ehri 2005; Share 2008; Perfetti 2007). Automatic word recognition is considered a prerequisite for fluent reading: as decoding becomes effortless, cognitive resources are freed for comprehension (Breznitz 2006).

One of the clearest behavioural markers of this developmental shift is considered to be the divergence between reading aloud and silent reading. In early grades, the two modalities are closely aligned, as silent reading still engages phonological processing and covert, subvocal articulation (LaBerge and Samuels 1974; Perfetti 1985; Wright, Sherman, and Jones 2004). With increasing proficiency, silent reading becomes faster than reading aloud and progressively less dependent on phonological decoding, reflecting the acquisition of larger orthographic units and more efficient lexical access (Brysbaert 2019; Rayner 2009). This divergence has been documented in English and Italian (Zoccolotti et al. 2009; Kim, Wagner, and Foster 2011) and in cross-linguistic finger-tracking work on Bulgarian and Italian primary school children (Lento et al. 2024; Marzi et al. 2026).

2.2 Studies on Aspects of Readability

The most extensively documented predictors of reading speed are word length and word frequency, both of which interact with orthographic transparency and with reading proficiency. Relatively transparent orthographies such as Bulgarian, with highly consistent

grapheme–phoneme correspondences and rules for exceptions, support early mastery of decoding (Schüppert et al. 2017; Seymour, Aro, and Erskine 2003; Share 2008), and have been argued to allow learners to rely on small, consistent grain sizes from the onset of literacy (Ziegler and Goswami 2005; Ziegler et al. 2010). Within this setting, the effect of word length is strongest in early grades and progressively diminishes with age, as serial decoding gives way to more holistic recognition (De Luca et al. 2008; Joseph et al. 2009; Zoccolotti et al. 2009; Marzi et al. 2020).

Word frequency shows the opposite trajectory: its facilitatory effect sharpens with experience, with high-frequency items accessed increasingly efficiently while low-frequency items remain costly (Lento et al. 2024; Marzi et al. 2026). These two effects are robust across measures of both reading time and reading speed and provide the empirical baseline against which any feature-based readability model must be evaluated.

Beyond length and frequency, morphological structure also plays a role in word recognition. Developing readers gradually learn to exploit morphological regularities to anticipate upcoming segments and reduce processing load (Beyersmann and Grainger 2018; Beyersmann et al. 2015; Carlisle 2000). Morphological decomposition effects have been observed across Spanish (Casalis et al. 2009), French (Beyersmann et al. 2015), German (Hasenäcker and Schroeder 2017), Italian (Burani 2010; Burani et al. 2018; Traficante et al. 2011), among others, suggesting a robust role for morphology in reading acquisition across languages. These effects become particularly visible in silent reading, where articulatory constraints no longer impose a strict left-to-right pace (Marzi et al. 2026). For Bulgarian, with its rich inflectional and derivational morphology and its postpositional definiteness marker, morphological cues are expected to be especially informative.

A further distinction that has emerged is the contrast between content and function words. Function words, being shorter, more frequent, and less semantically demanding, tend to reach automaticity earlier in development, whereas content words remain more sensitive to lexical variables such as length and frequency (Marzi et al. 2026). However, as function words play a distinct role in syntactic segmentation and express syntactic relations, their function also plays a role in reading speed.

These observations only emphasise the need to build a complex readability evaluation approach that incorporates all major features that determine the level of difficulty of a text relative to an age group or reading skills level.

2.3 Readability Assessment Frameworks

A number of indices for measuring text readability have been proposed in the literature, typically based on surface characteristics such as sentence and word length. Although they were developed independently and calibrated against different reference populations, these indices share a common assumption: that the difficulty of a text can be captured by a small number of easily computable variables – chiefly the average length of its sentences and the average length of its words. They differ mainly in how these variables are combined and in the scale on which the resulting score is reported.

The **Flesch Reading Ease Index** (FLI) (Flesch 1948), originally developed for English, assigns a score from 100 (very easy) to 0 (very difficult). It combines average sentence length, expressed as the number of words per sentence, with average word length, expressed as the number of syllables per word, and weights the latter heavily on the assumption that polysyllabic vocabulary is the dominant source of reading difficulty:

$$\text{FLI} = 206.835 - 1.015 \times (\text{avg. sentence length}) - 84.6 \times (\text{avg. word length}) \quad (1)$$

The two large multiplicative constants and the intercept of 206.835 were derived empirically by fitting the formula to reading comprehension data from school-age readers of English. A score above 90 corresponds to text accessible to a fifth-grade reader, whereas scores below 30 indicate text requiring college-level education.

The **LIX index** (Läsbarhetsindex) (Björnsson 1968) was developed for Swedish and is widely used across European languages. It ranges, in practice, from roughly 20 (easy) to 60 (difficult) and combines average sentence length with the percentage of long words, where long words are conventionally defined as those with more than six characters:

$$\text{LIX} = \frac{\#\text{ words}}{\#\text{ sentences}} + \frac{\#\text{ long words} \times 100}{\#\text{ words}} \quad (2)$$

By thresholding word length at six characters rather than averaging it, LIX is somewhat less sensitive to the presence of a few very long words and is in principle more portable across languages with different average word lengths than indices that rely on syllable counts.

The **Coleman–Liau Index** (CLI) (Coleman and Liau 1975) avoids syllable counting altogether and operates entirely on letter and sentence counts, which makes it particularly well-suited to automatic computation:

$$\text{CLI} = 0.0588 \times \frac{\#\text{ letters}}{\#\text{ words}} \times 100 - 0.296 \times \frac{\#\text{ sentences}}{\#\text{ words}} \times 100 - 15.8 \quad (3)$$

The resulting value is interpreted as the approximate number of years of formal education required to understand the text: a CLI of 8 corresponds to eighth-grade level, a CLI of 12 to the end of secondary school, and so on.

The dominant computational approach to text readability assessment treats reading difficulty as a classification problem driven by a feature inventory organised into four groups: text, lexical, morphological, and syntactic features (Dell'Orletta, Montemagni, and Venturi 2011). This framework was originally developed for Italian and has since served as a template for readability work in other languages, including Bulgarian (Pirrelli and Koeva 2024).

Text characteristics include sentence length, calculated as the average number of words per sentence, and word length, calculated as the average number of letters per word. Lexical characteristics comprise the percentage of unique words relative to an age-appropriate reference list, the distribution between basic vocabulary, high-frequency words, and relatively low-frequency words, and the number of occurrences of different words in the text. Morphosyntactic features refer to lexical density, defined as the ratio of content words

(verbs, nouns, adjectives, and adverbs) to the total number of words. Syntactic features include the depth of the syntactic trees, the overtness and transparency of the components of the clause, the average depth of the syntactic tree of subordinate clauses, the word order of subordinate clauses, and the length of dependent clauses.

2.4 Limitations for Bulgarian

Although there is general agreement between the assessments provided by different readability indices, there is no complete overlap. Critically, the indices were created and tested primarily for English and Swedish, and require adaptation for Bulgarian and other Slavic languages (for Russian, formulae with adjusted constants have been proposed).

More fundamentally, the assessment of texts using these measures is incomplete, as it does not account for the lexical, morphological, and syntactic features of Bulgarian words, nor for the semantic, selective, and grammatical restrictions arising from Bulgarian word order, lexical combinations, and syntactic constructions.

As an illustration, the excerpt from the short story *Lesson in Japanese* by Bulgarian author Viktoriya Beshliyska (Table 1) was assigned $FLI = 19.1$ (very difficult; college level), $LIX = 39.5$ (medium to difficult; approx. 9th grade), and $CLI = 10.1$ (difficult; approx. 10th/11th grade). While these results reflect the general difficulty of the text, they cannot precisely classify it into an age-related or difficulty-based category, and do not reflect the specific lexical, morphological, and syntactic sources of its difficulty. In the example, we notice a concentration of longer words such as *японското* 'Japanese', *семейство* 'family', *приближава* 'approaches' (verb), *вцепенявам* 'freeze' (verb), *качулката* 'hood', but they are part of the general vocabulary and do not pose a difficulty to fluent readers at the end of primary school.

Урок по японски Виктория Бешлийска	**Lesson in Japanese** Viktoriya Beshliyska
Сутринта отново виждаме японското семейство. Към нас приближава мъжът и аз се вцепенявам. Когато очите му срещат моите, той повдига любопитно вежди. Вероятно очаква да отвърна на поздрава му. Бързо нахлупвам качулката си. Тогава мъжът се усмихва и сякаш казва: „Не се притеснявай, разбирам те!“. В този миг се случва нещо странно: от устата ми започват да се нижат думи на японски.	In the morning, we see the Japanese family again. The man approaches us, and I freeze. When his eyes meet mine, he raises his eyebrows curiously. He's probably expecting me to return his greeting. I quickly pull my hood over my head. Then the man smiles, as if to say, "Don't worry, I understand!" At that moment, something strange happens: words in Japanese begin to flow out of my mouth.

Table 1
Short story excerpt demonstrating various readability indices.

3. Dataset and Automatic Preprocessing

Empirical reading data were collected with the *ReadLet* tablet-based platform (Taxitari et al. 2021; Ferro et al. 2018), which records the continuous movements of a reader's index finger across a connected text together with the voice signal in reading aloud sessions. Eighty-three Bulgarian children from grades 2 to 5 at a primary school in Sofia took part (Marzi et al. 2026), with no selection bias for reading skills, all native speakers of Bulgarian and with no documented learning-related special needs. Each child read two age-appropriate narrative texts on a tablet – one aloud and another one silently of similar length and complexity appropriate for the age, and answered two comprehension questions following each episode of the text. Texts ranged from approximately 280 words at grade 2 to approximately 700 words at grade 5, matched for narrative complexity to the age of the target readers (Pirrelli and Koeva 2024). Responses to the comprehension questions were recorded but are not analysed in the present paper, which focuses exclusively on decoding-speed measures; the analysis of comprehension data is left to future work.

All reading data are anonymised at the point of collection: each session is identified by a numeric code, with no personally identifying information retained beyond grade level, age in months, and gender. Analyses in the present study are conducted exclusively on aggregated data, grouped by grade and reading modality, with no individual reader being identifiable in any of the reported results. The dataset records the duration of the finger tracing within the token's bounding box, the reading modality (reading aloud or silent reading), and the resulting per-token reading speed, computed in both letters per second and syllables per second.

Table 2 summarises the distribution of the Bulgarian finger-tracking dataset across the eight groups categorised by grade (2 – 5) and modality (aloud / silent), comprising a total of 248 reading sessions from 83 children and a total of 128,386 tracked words.

Grade	Modality	Sessions	Children	Words read	Words/session
2	aloud	23	21	6,675	290
2	silent	21	21	6,088	290
3	aloud	38	38	16,970	447
3	silent	38	38	16,698	439
4	aloud	32	32	18,928	592
4	silent	32	32	18,980	593
5	aloud	32	31	21,750	680
5	silent	32	31	22,297	697
Total	**aloud**	**125**	–	**64,323**	–
Total	**silent**	**123**	–	**64,063**	–
Total		**248**	**83**	**128,386**	–

Table 2
Distribution of the Bulgarian finger-tracking dataset across the eight groups (grade × modality).

Preliminary automatic processing of the texts involves annotation at the following linguistic levels: sentence segmentation; tokenisation; part-of-speech (POS) tagging; lemmatisation; and sentence structure identification using a Universal Dependencies (UD) model. The analysis was performed using the Bulgarian multi-component system for processing and linguistic annotation (Koeva, Obreshkov, and Yalamov 2020).

The annotation supports statistical investigation of readability features across different linguistic levels:

- phonological and general lexical complexity (word length and lexical diversity);
- coverage of different parts of speech and their word forms;
- lexical typicality (statistical frequency of usage and distribution of words);
- morphological complexity (composition of words);
- syntactic functions of words in the sentence;
- word order positions and word combinations.

4. Analysis of Readability Features

The objective of this study is to investigate the features at phonological, lexical, morphological and syntactic level, that contribute to the classification of words and larger language units (phrases, sentences) according to their reading difficulty.

Our current specific tasks are related to the research questions set in the Introduction in Section 1. Our future work aims at developing a comprehensive system of criteria for assessing reading difficulty which can be formalised and implemented to perform automatic text classification based on readability.

Difficulty measure. Reading difficulty was operationalised as the average reading speed for a word, calculated as the number of letters read per second. Words are classified as: *difficult*/slow to read (read at a slow speed), *average* (read at the most common speed), and *easy*/fast to read (read at a high speed). The class boundaries were determined from the distribution of reading speeds: the speed recorded for the majority of words (60%) is considered normal; the slowest 20% constitute the difficult class, and the fastest 20% the easy class.

The following features were defined at the word-level:

1. length in number of letters and number of syllables;
2. ratio of consonants to total number of letters and number of clusters of 3 or more consonants;
3. part of speech;
4. number of prefixes and suffixes, as well as the definite article;
5. phonological irregularities in wordforms;

6. number of occurrences in the text so far;
7. syntactic function;
8. position of the word in the sentence.

4.1 Average Reading Speed Thresholds by Grade and Modality

For each grade × modality, the 20th and 80th percentiles of per-token reading speed (in letters per second) were taken as the boundaries between the three difficulty classes: tokens read at a speed below the 20th percentile are labelled *difficult*, those between the 20th and 80th percentiles – *average*, and those above the 80th percentile – *easy*. The resulting thresholds are reported in Table 3 and Figure 1.

Grade	Modality	Difficult (slow)	Average	Easy (fast)
2	aloud	< 5.49	5.49–17.09	≥ 17.09
2	silent	< 5.84	5.84–17.61	≥ 17.61
3	aloud	< 6.54	6.54–20.16	≥ 20.16
3	silent	< 7.76	7.76–23.92	≥ 23.92
4	aloud	< 7.75	7.75–22.45	≥ 22.45
4	silent	< 9.22	9.22–26.61	≥ 26.61
5	aloud	< 9.98	9.98–23.95	≥ 23.95
5	silent	< 10.93	10.93–29.76	≥ 29.76

Table 3
Reading speed thresholds (letters per second) for the three difficulty classes.

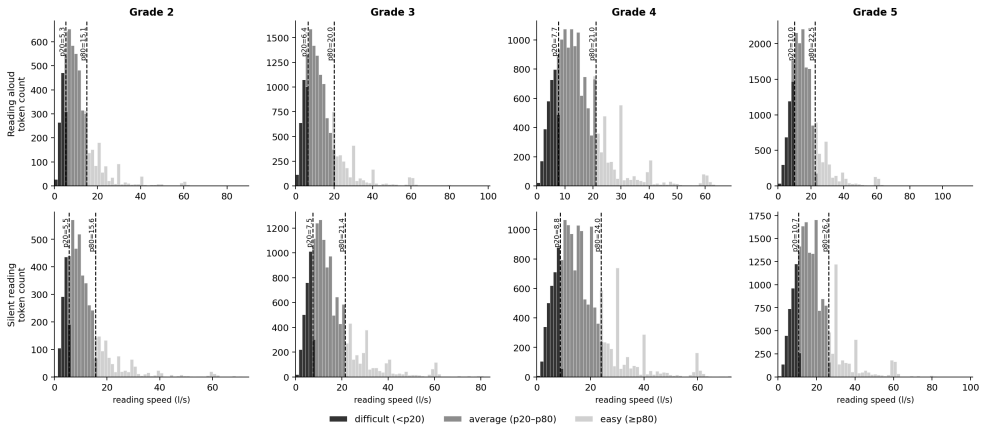


Figure 1
Per-token reading speed distributions (letters per second) for each grade × modality. Dashed lines mark the 20th and 80th percentiles, which define the boundaries between the three difficulty classes (dark – *difficult*; gray – *average*; light gray – *easy*).

Two patterns are immediately visible: distributions shift progressively rightward across grades, reflecting developmental gains in reading speed, and silent thresholds sit systematically above aloud thresholds within each grade, i.e. silent reading is faster, with the gap widening with grade progress.

4.2 Phonological Features

Word Length in Letters and syllables. Table 4 shows that word length is significantly associated with reading difficulty in every grade and modality as the average length (both in terms of number of letters and in terms of number of syllables) of easier/faster to read words is shorter than for more difficult words. However, the difficult and average classes are not separable by length alone.

		Length in letters				Length in syllables			
Gr	**Mod**	H	p	D / A	E	H	p	D / A	E
2	aloud	292	$<10^{-60}$	4.63 / 4.60	3.24	287	$<10^{-62}$	1.98 / 1.92	1.33
2	silent	340	$<10^{-70}$	4.79 / 4.45	3.11	346	$<10^{-75}$	2.04 / 1.87	1.25
3	aloud	411	$<10^{-89}$	4.61 / 4.62	3.60	399	$<10^{-86}$	1.95 / 1.94	1.49
3	silent	685	$<10^{-148}$	4.67 / 4.55	3.31	682	$<10^{-148}$	1.97 / 1.90	1.35
4	aloud	835	$<10^{-181}$	4.51 / 4.76	3.30	774	$<10^{-168}$	1.90 / 2.01	1.39
4	silent	1200	$<10^{-260}$	4.69 / 4.65	3.06	1114	$<10^{-241}$	1.99 / 1.96	1.25
5	aloud	1027	$<10^{-223}$	4.51 / 4.80	3.30	986	$<10^{-214}$	1.92 / 2.03	1.38
5	silent	1520	$<10^{-300}$	4.75 / 4.68	3.00	1454	$<10^{-300}$	2.02 / 1.98	1.23

Table 4
Word length in letters (left) and in syllables (right) across the three difficulty classes, by grade and modality. H is the Kruskal-Wallis statistic on 2 degrees of freedom. Mean length is reported as D / A (the *difficult* and *average* classes do not differ significantly) and E (*easy*).

Consonant Contents. Table 5 shows that both the consonant-to-letter ratio and the presence of a 3+ consonant clusters are significantly associated with reading difficulty for all grades and both modalities, but the differences are small between the difficult/average and the easy category. Only 2.5% of tokens overall contain a 3+ consonant cluster, so the results reported on this feature are not reliable.

4.3 Part of speech

Results show that part of speech is significantly associated with reading difficulty (Figure 2): function words (coordinating and subordinating conjunctions, particles, prepositions, pronouns) are over-represented in the easy class, while content words (nouns, verbs, adjectives, numerals, proper nouns) are over-represented in the average and difficult classes. This function/content asymmetry replicates the one reported by Marzi et al. (2026) for Bulgarian and Italian children and is consistent with the broader claim that function words automatise earlier in development than content words.

		Consonant-to-letter ratio				3+ consonant cluster			
Gr	**Mod**	H	p	D / A	E	χ^2	p	D / A	E
2	aloud	11.1	.004	0.538 / 0.537	0.527	10.6	.005	3.65% / 3.57%	1.74%
2	silent	3.1	.209	0.545 / 0.532	0.530	6.2	.045	4.09% / 3.12%	2.30%
3	aloud	10.4	.006	0.538 / 0.538	0.529	13.3	.001	3.36% / 2.71%	1.89%
3	silent	8.6	.013	0.541 / 0.537	0.532	10.6	.005	2.90% / 2.74%	1.78%
4	aloud	17.1	$<10^{-3}$	0.531 / 0.535	0.523	24.7	$<10^{-5}$	2.71% / 2.70%	1.27%
4	silent	30.3	$<10^{-6}$	0.531 / 0.536	0.512	29.7	$<10^{-6}$	2.78% / 2.68%	1.17%
5	aloud	27.1	$<10^{-5}$	0.532 / 0.534	0.524	28.0	$<10^{-6}$	2.14% / 2.67%	1.26%
5	silent	52.8	$<10^{-11}$	0.537 / 0.535	0.514	35.2	$<10^{-7}$	2.42% / 2.53%	1.02%

Table 5
Consonant-to-letter ratio (left, Kruskal-Wallis) and presence of at least one cluster of three or more consonants (right, chi-square on the binary feature), across the three difficulty classes by grade and modality. Mean ratio and % of tokens containing a 3+ cluster are reported as D / A (the *difficult* and *average* classes do not differ meaningfully) and E (*easy*).

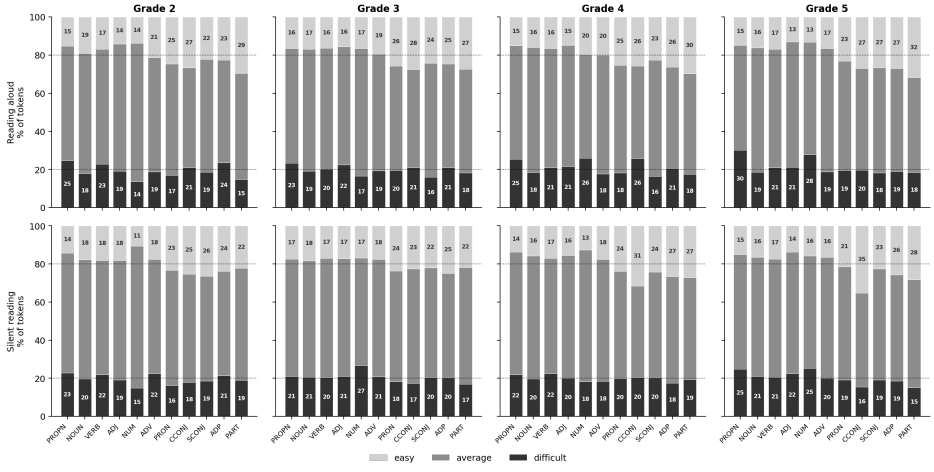


Figure 2
Classification of each POS (grade × modality) into the three difficulty classes. Colours: *difficult* (dark gray, bottom), *average* (gray, middle), and *easy* (light gray, top). The dashed lines mark the 20% and 80% boundaries that would obtain under independence.

However, the pattern is not uniform across function words. Despite being short and high-frequency, pronouns and conjunctions do not share the same level of automatisation as the other function-word categories (Figure 3).

Pronouns require additional processing time to resolve their referential meaning to an antecedent – a cognitive cost that reflects in the reading speed.

Similarly, conjunctions require processing time to uncover the structure of the sentence and the relation between the phrases or clauses they connect.

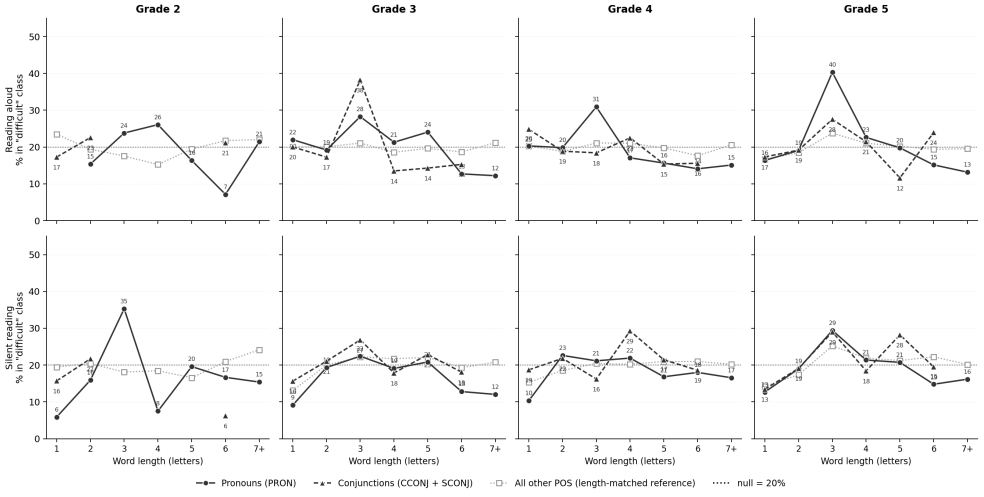


Figure 3
Comparison on pronouns and conjunctions classified into the difficult category, for aloud and silent reading, by grade.

Another diversion is that of proper nouns – they show the largest share of difficult tokens of any category (24.6%), reflecting the cost of reading unfamiliar names (*Зиги*, *Андрей*, etc.) for which children have no pre-existing lexical knowledge.

A closer look at pronouns and conjunction reveals that in some cases they are classified as difficult (read slowly by children) much more often than compared to function words in general. Two important observations can be made about the results on Figure 3. First, we are interested in the peaks for short pronouns and conjunctions which are clear outliers, signifying particular lexical elements that cause problems and slow reading. Secondly, we notice that these differences are dropped for silent reading, which can have a significance for the quality of reading and comprehension in silent mode if the reading speed of these items is in fact linguistically motivated.

Figure 4 compares the different pronoun subtypes, separately for each grade and modality. Two patterns dominate. First, articulation while reading aloud exposes systematic differences between pronoun subtypes that silent reading largely neutralises. Second, the indefinite and negative pronouns are read substantially more slowly than the others. Developmental change across grades 2–5 is small for all subtypes, suggesting that the general cost reflects the inherent referential demands of these pronouns rather than a transient difficulty that primary-school readers outgrow.

Figure 5 of the most difficult pronoun forms show two different trends across grades. For some items difficulty declines monotonically from grade 2 to grade 5, indicating genuine automatisation. However, for most of the pronouns difficulty plateaus or rises, reflecting the increasing referential complexity of the constructions in which these pronouns appear in higher-grade texts.

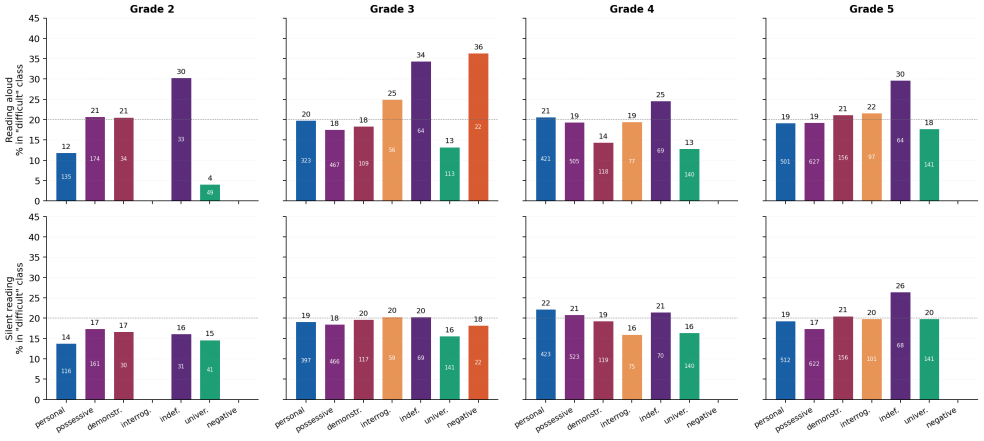


Figure 4
Reading difficulty by pronoun subtype, by grade and modality.

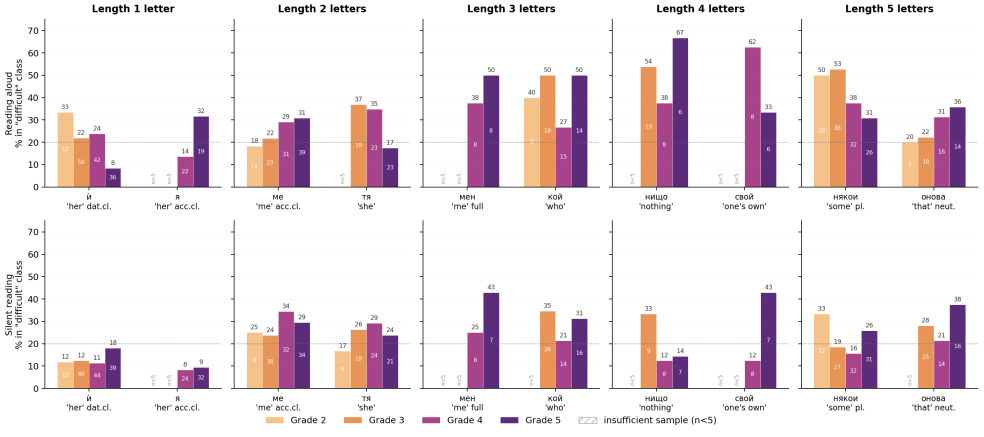


Figure 5
Most difficult pronoun forms of lengths 1–5 letters, by grade and modality.

Figure 5 also reveals a consistent modality asymmetry at individual pronoun level as seen in generalised terms before (Figure 3): reading aloud sits above silent reading in almost all cases. As discusses, this may be a sign of better automatisation while reading silently, but does not eliminate the risk of mechanical reading with less comprehension depth. These observations show the need to perform a more focused analysis of reading patterns with view to function words, and in particular pronouns.

The difficult nouns in grade 2 are mostly definite plural or definite singular forms of moderate length (5–7 letters), e.g. *вълните* ‘wave’ (fem., pl., def.), *балкона* ‘balcony’ (masc., sg., def.), *стаята* ‘room’ (fem., sg., def.). From grade 3 onwards, the difficult nouns are

longer and not part of everyday vocabulary, e.g. *небосвода* 'sky dome' (figurative, masc., sg., def.), *аквариум* 'aquarium' (masc., sg.), *въртележки* 'carousel' (fem., pl.), *лунапаркът* 'amusement park' (masc., sg., def.).

Grade 2 difficult verbs are short (3–7 letters) and high-frequency but either impersonal (*има* '(there) is/are') or in past tense (*имаше* 'was having'), and in some cases there is grammatical ambiguity of the verb form (*видя* can be 1st person singular present tense 'see', or 2nd/3rd person singular aorist 'saw'). From grade 3 onwards, the difficult category is dominated by past participles used adjectivally (*набраздена* 'furrowed', *обрасла* 'overgrown', *обзаведена* 'furnished', *вкаменен* 'petrified'), which are formally verbs but functionally adjective-like and morphologically complex.

Grade 2 difficult adjectives are usually phonologically or morphoogically challenging, e.g. *кръгъл* 'round' (contains two letters 'ъ'), *внушителни* 'impressive' (pl.), *сияещо* 'glowing' (neut.), etc. From grade 3 onwards difficult adjectives are long, low-frequency, semantically rich and inflectionally demanding, e.g. , *благородната* 'noble' (fem., sg., def.), *теракотена* 'terracotta' (fem.), *причудливи* 'whimsical' (pl.).

The auxiliary verb *съм* 'to be' appears to be predominantly easy to read (Figure 6), which is not surprising since its forms are also short. More thorough analysis is required to determine what causes the difficulty and the differences across modality. Across grades, while general verb difficulty follows consistent distribution, the difficulty of the auxiliary verb shows irregular patterns. The fluctuations in difficulty are likely to be caused by the more complex verb forms, e.g. perfect, passive, etc., and constructions that involve the auxiliary (*трябва да беше дошъл* 'should have probably arrived'; *са били категорично против* 'have been firmly against').

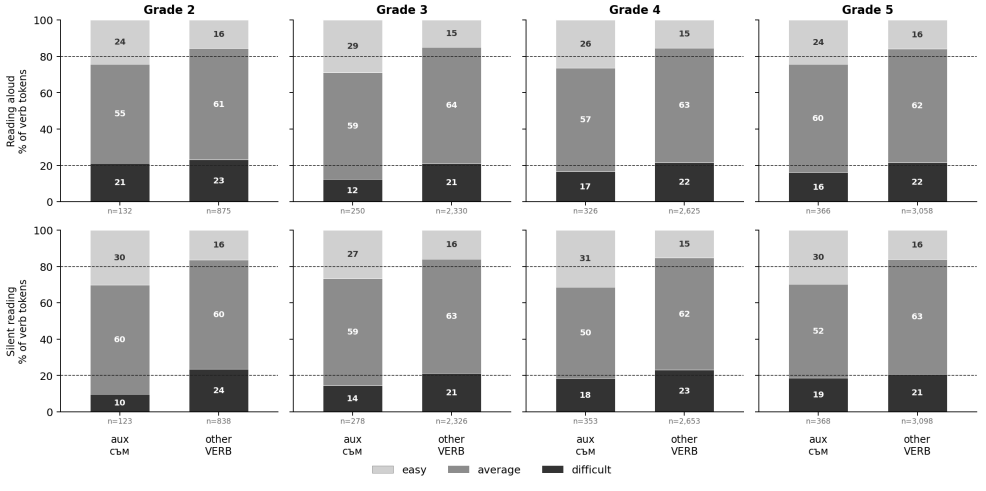


Figure 6
Difficulty level of the auxiliary verb, by grade and modality, compared to general verb reading difficulty.

4.4 Morphological Complexity

As already seen, morphologically complex words usually take longer time to read, but since they are also longer in length, it is difficult to judge these two features separately. Here we examine which particular morphemes may pose more challenges and whether the difficulty fades with grade progress.

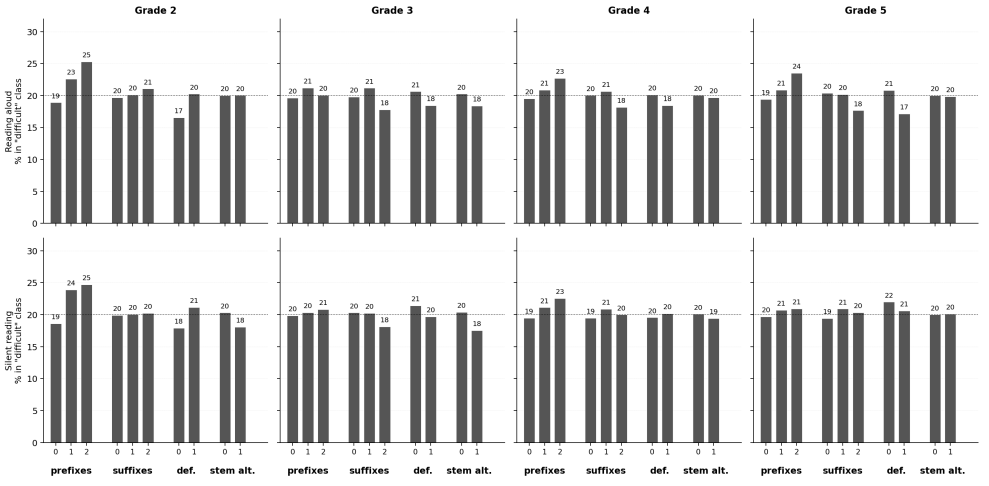


Figure 7
Difficulty of morphologically complex words, by grade and modality.

Figure 7 shows the influence of four morphological and phonological features – number of prefixes, number of suffixes, presence of a definite article, and stem alternation in wordforms. The results show that these are statistically significant but weak associations with reading difficulty (much smaller than the effect of word length).

Among the four, prefixation impedes reading the most. Multisuffix words are generally no slower to read and at higher grades are even slightly faster. This confirms the common belief that when a multi-morphemic ending is a familiar pattern, the reader processes it as a unit and the complexity of the form does not require longer processing time.

This is further confirmed by the fact that the definite article shows almost no effect across grades, indicating that this overt Bulgarian-specific morphological marker is largely automatised very early, even in grade 2.

Stem alternations also show no consistent effect, but since the lexemes that undergo such changes in the texts are high-frequency items (*ръка – ръце*, *око – очи*, *голям – големи*) whose forms are learned early and retrieved holistically, no reliable conclusions can be drawn.

The fact that none of the four features shows a clear developmental trajectory across grades 2–5 reinforces the overall picture that morphological properties are not slowly automatised across primary school, but are either trivial from the start (definite article, regular suffixation) or remain a relatively trivial challenge (prefixation). For a readability index, this

means that morphological complexity should have a lower weight alongside other more pronounced features.

4.5 Frequency Properties

In general, across content categories, the slowest words are low-frequency words that likely the readers encounter for the first time. Figure 8 shows a robust trend stable across grade and modality, that subsequent occurrences are read faster than the first one, and most often the difficulty level moves out of the *difficult* into the *average* or *easy* category.

Silent reading produces slightly larger differences with repetition than reading aloud at grades 2–3, and the two modalities converge at grades 4–5. Some lemmas, particularly *вещица* 'witch' and *кралица* 'queen' – remain in the *difficult* class even after multiple encounters, indicating that within-session repetition does not always influence lexically demanding items; for these cases, the contextual complexity of the constructions in which they appear continues to impose a processing cost that repetition cannot fully overcome.

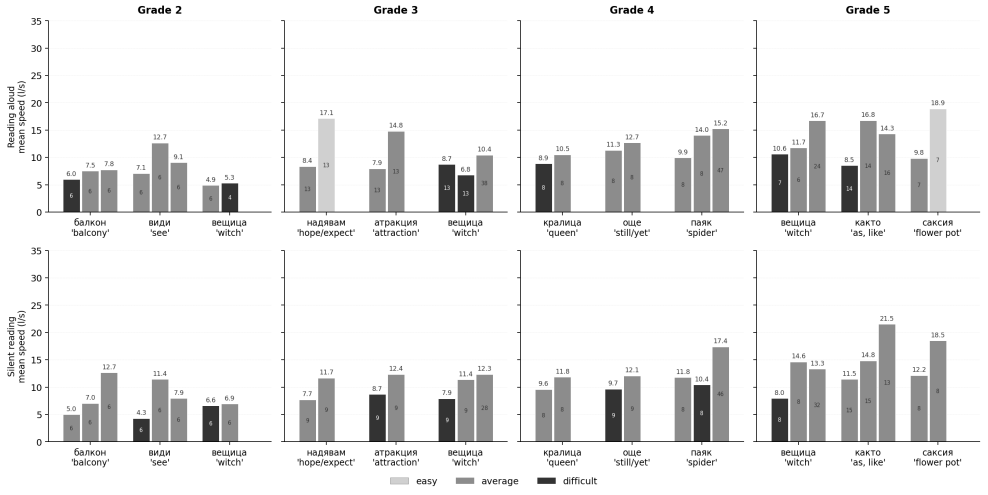


Figure 8
Progression of reading speed for content words at first, second and any further occurrences of the words within a reading session, by grade and modality.

The detailed analysis of word frequency in the language and their reading difficulty is beyond the scope of the current study. Here we only acknowledge that with each subsequent occurrence of a word, in general its reading difficulty decreases, at least for the span of the text. This is valid for both silent reading and reading aloud.

Overall, feature-based readability index should incorporate token-recurrence count alongside the static lexical features of word length, frequency, and part of speech. Moreover, this can be implemented as a strategy for learning new and more complex words by repetition in the learner texts.

4.6 Syntactic Features

Syntactic function correlates with reading difficulty (Figure 9), but the effect is derived in combination with the lexical content typically appearing in each role. It is not entirely clear what the relation contributes once those lexical correlates are accounted for, and further analysis is needed in order to determine the weight of generalised dependency features in a feature-based readability index.

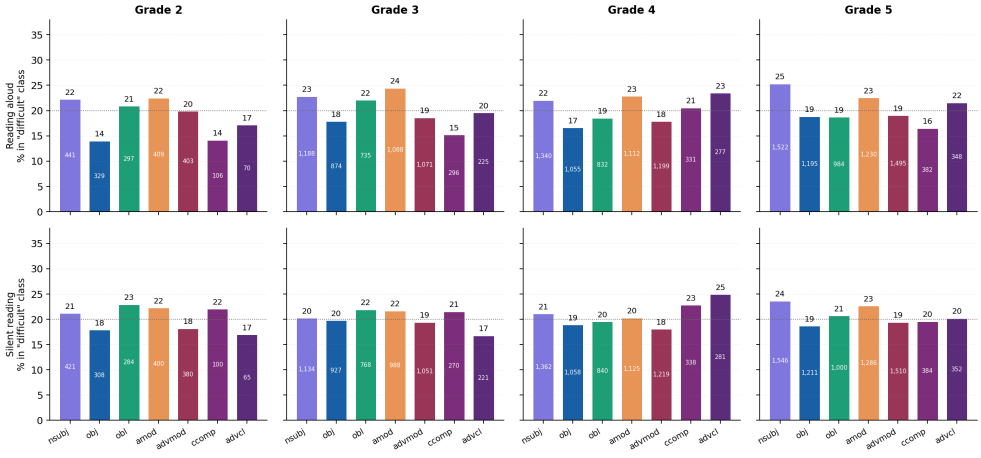


Figure 9
Reading difficulty for seven core dependency relations (UD schema), by grade and modality.

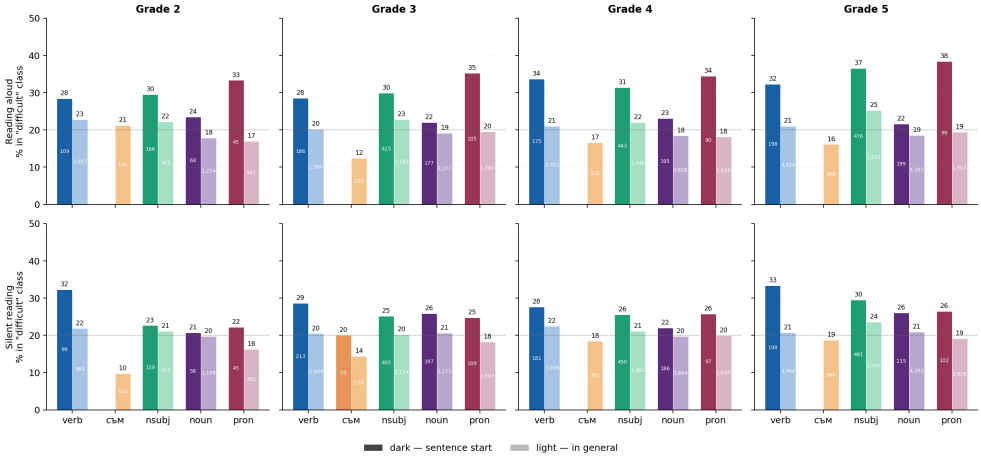


Figure 10
Comparing reading difficulty of verbs and nouns at the start of the sentence and in general, by grade and modality.

However, a more focused investigation of syntactic properties involving other aspects of the syntactic structure and in particular, features specific for Bulgarian, reveal more relevance.

Figure 10 shows that the position in the sentence is correlated with difficulty, and this is consistent across categories, grades, and modalities. Sentence-initial position shows more difficulty compared to the corresponding overall rate. The effect is more pronounced for verbs as an initial verb requires the reader to recover the non-overt subject, or process an impersonal sentence.

Subjects in initial position also require more processing time, most notably if it is expressed by a pronoun, which also requires anaphora resolution. Nouns are the least affected category.

Across modalities, reading aloud and silent reading produce comparable positional effects, but in general, for silent reading the difficulty gap is less pronounced.

For a feature-based readability index, sentence-initial position deserves an independent entry – particularly when scoring sentences whose first word is a verb or a pronoun subject, because the cost it imposes is not predicted by length, frequency, or POS alone.

Further, our observations show that reading speed of word bigrams can also serve as a predictor of grade-level fluency (Figure 11). Speed of main syntactic structures rises monotonically with grade, the slope is steeper than any single-word feature. For a feature-based readability index, the type of POS sequence a sentence is composed of carries diagnostic value beyond its individual word features. Figure 11 also shows that the most frequent syntactic structures are read with speed close to the average speed for the age group. This means that typical structures are generally not problematic for the children, and more focus is needed on the use of structures of lower frequency.

One final observation concerns phrase and clause boundaries within the sentence marked by punctuation (commas) or conjunctions. Figure 12 show that the average reading speed of bigrams that include the boundaries falls at the lower end of the *average* category. Phrase and clause boundaries cause slight slowing and this effect is consistent through grades and modalities, which shows that this is a regular reading feature rather than a learner's deficiency.

5. Discussion of Results

The results reveal several key findings that are directly relevant with a view to designing a readability index in Bulgarian.

1. **Word length is by far the strongest single predictor of reading difficulty.** Every other lexical, morphological, or syntactic feature we tested shows a lesser effect. Length emerges as the dominant predictor consistently across grade and modality. This in fact reaffirms the previously discussed indices such as FLI, LIX, CLI, etc. (see Section 2.3).

2. **Part-of-speech class is also a strong predictor and the results reveal a clear function-vs.-content words division.** The split is symmetric across modal-

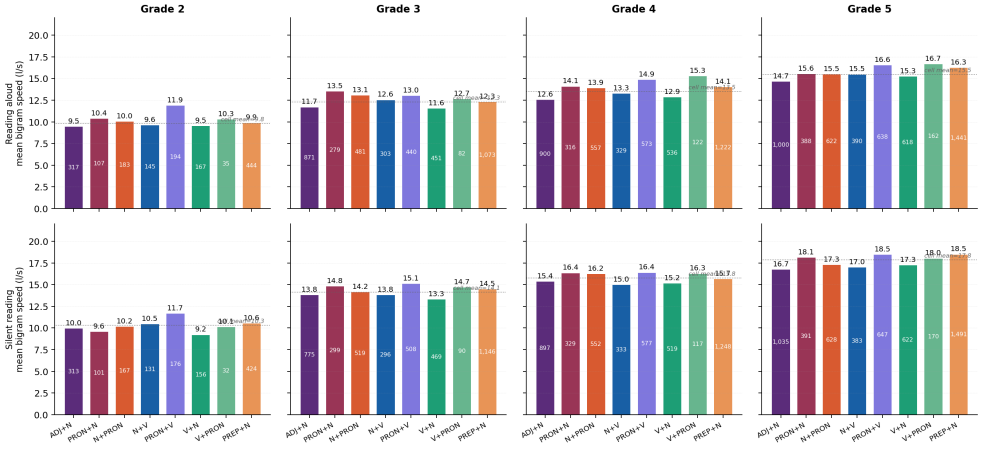


Figure 11
Reading speed of selected syntactic structures (bigrams), by grade and modality.

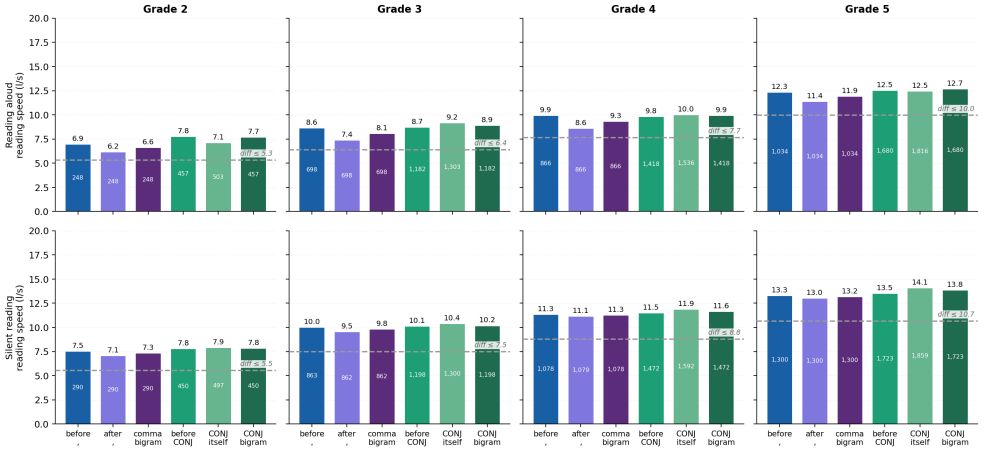


Figure 12
Reading speed at mid-sentence punctuation and conjunction boundaries, by grade and modality.

ities and stable across grades. In general, the auxiliary verb *съм* does not pose any significant difficulties, except in certain more complex structures in higher grade texts.

3. **Pronouns are associated with increased processing cost consistent across grades.** Among pronoun subtypes, indefinite (*някой, нещо*) and negative pronouns (*никой, нищо*) bear the heaviest cost because they introduce or exclude referents rather than retrieve them from prior discourse.

4. **Sentence-initial position imposes an additional processing cost that does not fade across grades.** Verbs and pronoun subjects at sentence-initial position show more challenges relative to the general baseline. The cost is consistent across grades, indicating a structural rather than developmental effect.

5. **Clause and phrase boundaries within the sentence require an additional processing cost consistent across grades.** Words before and after punctuation, as well as conjunctions, which mark phrase and clause boundaries, show lower reading speed. The cost is consistent across all four grades, indicating a structural property.

6. **Within-session word repetition is one of the most consistent accelerators of reading fluency.** The share of tokens classified as *easy* grows with every subsequent occurrence of a word in the text. The effect is strongest at grade 2 and gradually attenuates, but does not disappear at higher grades.

7. **Morphological complexity features show small influence.** The number of prefixes, suffixes, and the presence of a definite article, as well any stem irregularity that appear in a wordform, correlate with difficulty in the predicted direction but are not able to predict difficulty on their own. Notably, the effect disappears at higher grades, indicating that compound morphology is chunked and automated efficiently with progress of reading skills. This is a feature that can be used to evaluate readability only for beginner learners and to identify problems early on.

8. **Syntactic function carries a small but reliable signal that overlaps with POS.** Across the core dependency relations (`nsubj`, `obj`, `obl`, etc.), the slowest is the subject, which also is associated often with a sentence-initial position, and is a content word. The effect is largely a consequence of the lexical content typical of each role rather than the role itself.

9. **Reading speed of bigram POS structures is a developmental signal.** Mean reading speed for bigrams (ADJ+N, N+V, ADP+N, *etc.*) rises monotonically from grade 2 to grade 5, both in silent reading and reading aloud.

10. **Silent reading is consistently faster than reading aloud across all grades and all features, and the gap widens with age and development of reading skills.** Silent reading compresses the variability across some features, e.g. the difference between the slowest and fastest pronoun subtypes. On one hand, articulation exposes processing differences that silent reading partly absorbs. However, on the other hand, irregular patterns in speed in silent reading may suggest distraction and lower quality of reading comprehension.

Concerning the features most strongly predicting reading speed (**RQ1**, see Section 1), the empirical data provides clear insights. Word length is the dominant lexical predictor and outweighs every other tested feature in isolation. Other important features are POS, primarily the content-vs-function words distinction that drives much of the variance in

Feature	Strong	Weak	Stable across grades	Stable across modality
Word length	✓	×	✓	partially
POS class (content vs. function)	✓	×	✓	✓
Auxiliary *съм* vs. lexical verbs	✓	×	✓	✓
Sentence-initial position	✓	×	✓	partially
Clause / phrase boundary	✓	×	✓	×
Within-text repetition	✓	×	×	partially
POS bigram structures	✓	×	×	✓
Number of prefixes	×	✓	✓	✓
Suffixal complexity	×	✓	partially	✓
Definite article	×	✓	✓	✓
Stem alternations	×	✓	✓	✓
Syntactic function (UD dependency)	×	✓	✓	✓

Table 6
Summary of the features and their classification according to their contribution to readability. **Strong** = strong predictor of difficulty in the empirical data; **Weak** = statistically significant but weak predictor; **Stable across grades** = the effect magnitude does not change substantially between grades 2 and 5; **Stable across modality** = the effect magnitude is approximately the same in reading aloud and silent reading. The label 'partially' indicates that the effect varies systematically or shows only a partial dependency.

reading speed, sentence-initial position with a focus on sentences starting with a verb (impersonal or pro-drop), clause and phrase boundary positions. Morphological features – prefixes, suffixes, definiteness, stem alternation, carry only small independent weight. Syntactic function contributes only marginally to what POS already captures. Within-text repetition modulate the cost of individual challenging words by 30–70 % in each subsequent mention which suggests this as a possible strategy to improve readability.

For a feature-based readability index, these observations give a minimal effective feature set consisting of strong basic features and weaker secondary features to evaluate the readability of a text.

Concerning the features that remain stable across grades and the ones which gain or lose weight as decoding becomes more automatised (**RQ2**, see Section 1), the analysis points to two qualitatively different developmental patterns. The first applies to the lexical predictors (length, POS, function-vs-content split), which attain stable importance although with decreasing magnitude – the effects remain detectable at grade 5 although the absolute reading speed at every feature rises. The second pattern is characteristic of within-session repetition and the processing cost of sentence-level components, which show stable magnitude – the repetition gain and the cost of processing pronouns, sentence-initial words and clause and phrase boundaries are present at both grade 2 and grade 5, even as children

become more fluent in reading. The morphological features (prefixes, suffixes, definiteness, stem alternations) show only small magnitude and do not exhibit a clear developmental trajectory, suggesting that these features represent stable, low-weight processing costs that are neither automatised away nor accentuated as proficiency grows.

The most significant developmental feature is the bigram-speed trajectory of typical and high frequency syntactic sequences (ADJ+N, N+V, etc.), which doubles between grade 2 and grade 5 in roughly parallel fashion, indicating that what improves with development is the reader's general fluency on multi-word chunks of particular syntactic relations. More thorough analysis of syntactic features is necessary in order to integrate syntactic information into a readability index.

Concerning the features that are consistent or differ in reading aloud and in silent reading (**RQ3**, see Section 1), we make two main observations. First, silent reading consistently raises absolute reading speed across all features and grades. Second, and more interesting, in some cases silent reading drops some of the relevant features. One possible explanation is that silent reading neutralises features tied to articulatory or phonological complexity. However, smaller difficulty gaps under silent reading in structural positions requiring extra processing cost (e.g., anaphora resolution, sentence structure, etc.) may suggest loss of attention and deteriorated comprehension. While in general, the hypothesis articulated in RQ3 — that phonological and articulatory features weigh more heavily in reading aloud, is supported by the data, any further conclusions need to be based on a detailed examination of reading comprehension in reading aloud as compared to reading silently, especially in lower grades.

Several findings of the present analysis converge with those reported by Lento et al. (2024) and Marzi et al. (2026) on the same finger-tracking dataset. In particular, the studies confirm a robust developmental trajectory of reading speed from grade 2 to grade 5 and a stable content-vs-function-word contrast in which function words are read faster and automatised earlier than content words in both modalities. The consistency of these observations supports their status as developmental phenomena in Bulgarian primary-school reading, and contribute to the efforts to distinguish reading developmental features (which evolve across grades) from stable decoding and comprehension features (which are related to text processing and information comprehension).

6. Notes on Reading Development Strategies

Based on the findings presented in the paper, we can compile a list of practical recommendations to support reading instruction and development of reading skills and text design for Bulgarian learners.

Vocabulary and Morphology. Before or after reading, attention should be paid to unfamiliar or more difficult words and those that require more processing time. In learners' texts, new, unfamiliar or long words should appear more than once, in different morphological forms and in different syntactic positions.

Pronoun Use. Reading should be combined with exercises on the use and comprehension of pronouns, including the correct identification of antecedents in more complex cases.

Particular attention should be paid to possessive and personal reflexive pronouns, indefinite and negative pronouns, and to the correct resolution of antecedents from the context.

Verb-Initial Constructions and Sentence Segmentation. Sentences that begin with a verb deserve special attention, especially those with complex verb forms where an auxiliary verb appears in sentence-initial position. Text and sentence segmentation during reading should also be explicitly taught, with particular emphasis on pausing correctly at the boundaries of phrases and sentences. Syntactic constructions that appear frequently in the language are more suitable for beginner readers, and gradual introduction of more complex syntactic structures could facilitate their acquisition without impeding reading speed significantly.

7. Conclusion

This paper has presented an analysis of reading data from children from grades 2 to 5. The main objective was to identify features measuring and predicting text readability for Bulgarian, with a focus on lexical and grammatical features. Our analysis confirms that established readability indices developed for English and Swedish are applicable but insufficient for Bulgarian, as they do not provide in-depth information to support any teaching or text design approaches to improve reading skills of learners.

The empirical findings yield a small but consistent set of strong predictors for text readability that need to be investigated further in order to be implemented into a readability index for Bulgarian.

The observations also show grade-by-grade developmental tendencies which reflect the gradual improvement of reading skills.

The aloud-vs-silent reading comparison reveals a systematic but bounded asymmetry. While silent reading is faster than reading aloud across all grades as it generally removes the cost of articulatory processing, it also raises the important concern of the quality of comprehension in younger readers.

Beyond the immediate application to readability scoring, the findings reported here suggest several directions for future work. First, the within-session repetition effect deserves dedicated study with materials designed specifically to test priming gradients across longer time intervals and across different text positions. Second, the modality discrepancies could be tested experimentally with stimuli that controllably vary articulatory load, sentence structure, and referent accessibility. Third, the lexicon-specific findings (the difficult words at each grade, the bigram patterns) could inform the construction of graded reading materials in which the introduction of demanding vocabulary is paced to take advantage of the priming gains demonstrated in the study.

The finger-tracking paradigm proved well-suited for collecting naturalistic reading-time data from young children at scale, and the resulting dataset provides an empirical foundation that text-corpus statistics alone cannot offer. The continued development of such corpora, in Bulgarian and in other under-resourced languages, is a productive route toward creating readability frameworks that are grounded in the actual reading behaviour of the children and the language specifics for which they are designed.

8. Limitations

Several caveats should be taken into account when interpreting the findings reported in this paper.

Sampling and population coverage. The data were collected at a single school across grades 2–5. The sample is therefore not statistically representative of Bulgarian primary-school readers as a whole, and the absolute reading-speed values, the thresholds, and the magnitudes of the effects should be interpreted as characterising this cohort rather than the wider population. The analyses report on correlations between linguistic features and reading difficulty and point to candidate predictors worth further investigation on larger and more diverse samples.

Difficulty thresholds. Reading difficulty is operationalised throughout a 20/60/20 split of the letters-per-second speed distribution, and words are classified as *difficult*, *easy*, and *average*. This threshold choice was made for analytic convenience, but this is a coarse discretisation of a continuous and skewed distribution of word reading speeds which may be approximated more precisely with other thresholds (e.g., a quintile-based segmentation). The analyses should be read as illustrative of tendencies rather than as precise quantifications, and any operational readability scoring system would need to justify any threshold choice in light of its intended end use.

Feature representation sets are mostly small. Some of the observations are made on a limited number of occurrences, which is insufficient to draw reliable conclusions. For some features a more targeted experiment would be needed in order to determine its significance for the readability of the text (e.g., pronouns in sentence-initial position).

Confounds between features. The features we examined are not statistically independent. Word length correlates with morphological complexity, frequency with the division between content and function words, part of speech with syntactic function, etc. A more complex analysis should be performed, e.g. multivariate regression, to disentangle the contributions of correlated features.

Reading comprehension was not measured. Reading speed is a measure of fluency, but does not reflect reading comprehension and information extraction. The difficulty classification applied here captures the *decoding* difficulty of words and sentences, and not necessarily the content difficulty of the text and the ability to process the information structure of the text. A future study pairing the finger-tracking data and decoding with post-reading comprehension and information extraction questions would be necessary to validate that the features identified as difficulty-correlated are also features that impede understanding, not just features that slow down decoding.

Acknowledgments

The work presented here is carried out with the support of the National Scientific Program "Development of Scientific Research and Innovation in Bulgarian Preschool and School Education", Institute of Education, Ministry of Education and Science, Agreement SP-13/04.11.2025.

References

Beyersmann, Elisabeth and Jonathan Grainger. 2018. Support from the morphological family when unembedding the stem. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(1):135–142.

Beyersmann, Elisabeth, Jonathan Grainger, Séverine Casalis, and Johannes C. Ziegler. 2015. Effects of reading proficiency on embedded stem priming in primary school children. *Journal of Experimental Child Psychology*, 139:115–126.

Björnsson, Carl Hugo. 1968. *Läsbarhet*. Liber, Stockholm.

Breznitz, Zvia. 2006. *Fluency in Reading: Synchronization of Processes*. Routledge.

Brysbaert, Marc. 2019. How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, 109.

Burani, Cristina. 2010. Word morphology enhances reading fluency in children with developmental dyslexia. *Lingue e Linguaggio*, 9(2):177–198.

Burani, Cristina, Stefania Marcolini, Daniela Traficante, and Pierluigi Zoccolotti. 2018. Reading derived words by Italian children with and without dyslexia: The effect of root length. *Frontiers in Psychology*, 9:647.

Carlisle, Joanne F. 2000. Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing*, 12(3):169–190.

Casalis, Séverine, Mathilde Dusautoir, Pascale Colé, and Stéphanie Ducrot. 2009. Morphological effects in children word reading: A priming study in fourth graders. *British Journal of Developmental Psychology*, 27(3):761–766.

Coleman, Meri and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Crepaldi, Davide, Marcello Ferro, Claudia Marzi, Andrea Nadalini, Vito Pirrelli, and Loukia Taxitari. 2022. Finger movements and eye movements during adults' silent and oral reading. In Ronit Levie, Amalia Bar-On, Orit Ashkenazi, Elitzur Dattner, and Gilad Brandes, editors, *Developing Language and Literacy: Studies in Honor of Dorit Diskin Ravid*. Springer International Publishing, pages 443–471.

De Luca, Maria, Laura Barca, Cristina Burani, and Pierluigi Zoccolotti. 2008. The effect of word length and other sublexical, lexical, and semantic variables on developmental reading deficits. *Cognitive and Behavioral Neurology*, 21(4):227–235.

Dell'Orletta, Felice, Simonetta Montemagni, and Giulia Venturi. 2011. READ-IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies, Edinburgh, Scotland, UK*, pages 73–83, Association for Computational Linguistics.

Ehri, Linnea C. 2005. Learning to read words: Theory, findings, and issues. In *Scientific Studies of Reading*, volume 9. Taylor & Francis, pages 167–188.

Ferro, Marcello, Claudia Cappa, Sara Giulivi, Claudia Marzi, Ouafae Nahli, Franco Alberto Cardillo, and Vito Pirrelli. 2018. ReadLet: Reading for understanding. In *Proceedings of the 5th IEEE Congress on Information Science and Technology (IEEE CiSt'18), Marrakech, Morocco*, pages 1–6.

Flesch, Rudolf. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Hasenäcker, Jana and Sascha Schroeder. 2017. Syllables and morphemes in German reading development: Evidence from second graders, fourth graders, and adults. *Applied Psycholinguistics*, 38(3):733–753.

Joseph, Holly S. S. L., Simon P. Liversedge, Hazel I. Blythe, Sarah J. White, and Keith Rayner. 2009. Word length and landing position effects during reading in children and adults. *Vision Research*, 49(16):2078–2086.

Kim, Young-Suk, Richard K. Wagner, and Elizabeth Foster. 2011. Relations among oral reading fluency, silent reading fluency, and reading comprehension: A latent variable study of first-grade readers. *Scientific Studies of Reading*, 15(4):338–362.

Koeva, Svetla, Nikola Obreshkov, and Martin Yalamov. 2020. Natural language processing pipeline to annotate Bulgarian legislative documents. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020), Marseille, France*, pages 6988–6994, European Language

Resources Association.
LaBerge, David and S. Jay Samuels. 1974. Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6(2):293–323.
Lento, Alessandro, Andrea Nadalini, Marcello Ferro, Claudia Marzi, Vito Pirrelli, Tsvetana Dimitrova, Hristina Kukova, Valentina Stefanova, Maria Todorova, and Svetla Koeva. 2024. Assessing reading literacy of Bulgarian pupils with finger-tracking. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024), Sofia, Bulgaria*, pages 140–149.
Marzi, Claudia, Marcello Ferro, Andrea Nadalini, Vito Pirrelli, Maria Todorova, Tsvetana Dimitrova, Valentina Stefanova, Hristina Kukova, and Svetla Koeva. 2026. Comparable reading development in Bulgarian and Italian: Cross-linguistic insights from a finger-tracking study. *Languages*, 11(4):70.
Marzi, Claudia, Anna Rodella, Andrea Nadalini, Loukia Taxitari, and Vito Pirrelli. 2020. Does finger-tracking point to child reading strategies? In *Proceedings of the 7th Italian Conference on Computational Linguistics (CLiC-it 2020), Bologna, Italy*, volume 2769.
Nadalini, Andrea, Claudia Marzi, Marcello Ferro, Loukia Taxitari, Alessandro Lento, Davide Crepaldi, and Vito Pirrelli. 2023. Eye-voice and finger-voice spans in adults' oral reading of connected texts: Implications for reading research and assessment. *The Mental Lexicon*, 18(3):366–400.
Perfetti, Charles. 1985. *Reading Ability*. Oxford University Press, New York.
Perfetti, Charles. 2007. Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, 11(4):357–383.
Pirrelli, Vito and Svetla Koeva. 2024. Developing materials for assessing reading literacy and comprehension of early graders in Bulgaria and Italy. *Foreign Language Teaching (Chuzhdoezikovo Obuchenie)*, 51(1):29–37.
Rayner, Keith. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.
Rayner, Keith. 2009. Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, 62(8):1457–1506.
Schüppert, Anja, Wilbert Heeringa, Jelena Golubovic, and Charlotte Gooskens. 2017. Write as you speak? A cross-linguistic investigation of orthographic transparency in Germanic, Romance and Slavic languages. In Martijn Wieling, Martin Kroon, Gertjan van Noord, and Gosse Bouma, editors, *From Semantics to Dialectometry: Festschrift in Honour of John Nerbonne*. College Publications, pages 312–427.
Seymour, Philip H. K., Mikko Aro, and Jane M. Erskine. 2003. Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, 94(2):143–174.
Share, David L. 2008. On the Anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin*, 134(4):584–615.
Taxitari, Loukia, Claudia Cappa, Marcello Ferro, Claudia Marzi, Andrea Nadalini, and Vito Pirrelli. 2021. Using mobile technology for reading assessment. In *Proceedings of the 6th IEEE Congress on Information Science and Technology (CiSt), Agadir, Morocco*, pages 302–307.
Traficante, Daniela, Stefania Marcolini, Alessandra Luci, Pierluigi Zoccolotti, and Cristina Burani. 2011. How do roots and suffixes influence reading of pseudowords: A study of young Italian readers with and without dyslexia. *Language and Cognitive Processes*, 26(4–6):777–793.
Wright, Gary, Ross Sherman, and Timothy B. Jones. 2004. Are silent reading behaviors of first graders really silent? *The Reading Teacher*, 57(6):546–553.
Ziegler, Johannes C., Daisy Bertrand, Dénes Tóth, Valéria Csépe, Alexandra Reis, Luís Faísca, Nina Saine, Heikki Lyytinen, Anniek Vaessen, and Leo Blomert. 2010. Orthographic depth and its impact on universal predictors of reading: A cross-language investigation. *Psychological Science*, 21(4):551–559.
Ziegler, Johannes C. and Usha Goswami. 2005. Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, 131(1):3–29.
Zoccolotti, Pierluigi, Maria De Luca, Gloria Di Filippo, Anna Judica, and Marialuisa Martelli. 2009. Reading development in an orthographically regular language: Effects of length, frequency, lexicality and global processing ability. *Reading and Writing*, 22(9):965–992.

Extraction of Handwritten and Printed Cyrillic Text from Documents: A Resource-Efficient Pipeline

Daniel Halachev
Sofia University "St. Kliment Ohridski"
Sofia, Bulgaria
danihalachev@gmail.com

Ivan Koychev
Sofia University "St. Kliment Ohridski"
Sofia, Bulgaria
koychev@fmi.uni-sofia.bg

The digitisation of historical, administrative, and personal documents in Bulgarian faces considerable challenges due to the lack of robust Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) systems tailored for the Cyrillic alphabet. While modern Vision-Language Models (VLMs) and large transformer-based architectures achieve state-of-the-art results, their performance and resource efficiency on low-resource languages remain prohibitive for decentralised, privacy-preserving applications. In this paper, we present a comprehensive, resource-efficient pipeline for extracting printed and handwritten Cyrillic text. Our system integrates advanced image preprocessing, YOLO-based document structure and table recognition, and a Permuted Autoregressive Sequence (PARSeq) model trained specifically for Bulgarian. We generated custom synthetic Bulgarian cursive datasets to mitigate the severe lack of real-world training data. Our evaluation indicates that the specialised PARSeq model outperforms traditional OCR tools such as Tesseract and EasyOCR on our custom degraded printed test set, and provides a practical, resource-efficient baseline for handwriting recognition compared to a modern local VLM (Qwen3-VL-4B). Finally, we discuss the discrepancy between synthetic and real handwritten data, highlighting the urgent need for a standardised, annotated Bulgarian HTR dataset.

Keywords: Optical Character Recognition, Handwritten Text Recognition, Bulgarian Language, Cyrillic Alphabet, PARSeq, Document Layout Analysis

1. Introduction

The modern world relies heavily on digital information, yet a vast amount of historical, administrative, and cultural data in Bulgarian remains locked in physical paper documents. The automatic extraction of text from these documents—Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR)—is essential for preserving, searching, and analysing this data. Documents such as exams, homework, essays, forms, books, magazines, and archival records contain valuable information that is yet to be digitised and analysed in its entirety. Such a system would improve accessibility and inspire future Bulgarian OCR research.

Despite substantial advancements in OCR technologies, most existing libraries and commercial systems are heavily optimised for the Latin alphabet. Tools that support Cyrillic

https://doi.org/10.47810/JCLIB.2.2026.06

often exhibit unsatisfactory accuracy, particularly when processing highly degraded printed text or cursive handwriting. Furthermore, while recent Large Language Models (LLMs) and Vision-Language Models (VLMs) have revolutionised natural language processing, their application to document recognition introduces severe drawbacks regarding high computational costs and privacy concerns when sending sensitive documents to third parties.

This research investigates the feasibility of using lightweight, highly efficient architectures for Bulgarian text recognition. Rather than relying on massive, resource-intensive models, we focus on optimising the quality-speed-cost trilemma through targeted experiments and architectural choices. A driving factor for our methodology was the strict requirement that the system must run locally on consumer-grade hardware (e.g. a standard laptop with 16 GB of RAM and an 8-core CPU). This constraint is essential to ensure data privacy, decentralisation, and accessibility. The primary goal was to design and evaluate a complete pipeline for recognising printed and handwritten Cyrillic text that is accurate, usable, and resource-efficient. To support practical digitisation workflows, the system was designed to expose a REST API and output results in standardised formats such as plain text, ALTO XML, and HOCR. Table 1 summarises the practical constraints that shape the full system.

Aspect	Constraint or Objective
Hardware	Local execution on 16 GB RAM and 8 CPU cores
Deployment	Privacy-preserving, open-source, locally runnable
Interface	Browser GUI and REST API
Outputs	Plain text, ALTO XML, HOCR

Table 1
Practical constraints shaping the system architecture.

To achieve these goals, we developed a complete document processing pipeline. We investigated the feasibility of using lightweight, highly efficient architectures for Bulgarian text recognition. We trained and evaluated a Permuted Autoregressive Sequence (PARSeq) (Bautista and Atienza 2022) model and benchmarked it against widely used open-source tools, Tesseract (Smith 2007) and EasyOCR (Kittinaradorn et al. 2020), as well as a local VLM, Qwen3-VL-4B (Qwen Team 2025).

In summary, our main contributions are:

- A practical, resource-efficient OCR/HTR pipeline prototype tailored for Bulgarian documents.
- A Bulgarian-adapted PARSeq training recipe utilising custom synthetic datasets to overcome the low-resource barrier.
- An empirical analysis of the discrepancy between synthetic and real-world handwriting, establishing a useful pilot baseline for future research.

To support reproducibility, the source code of the system is provided as anonymised supplementary material for review and will be made publicly available upon publication.

2. Related Work

The field of text recognition has evolved substantially from early mechanical devices to modern deep learning architectures. Early OCR systems relied on explicit character segmentation followed by classification using matrix matching or feature extraction combined with classifiers like Support Vector Machines (SVMs) or k-Nearest Neighbours (k-NN). Systems like Tesseract and EasyOCR remain standard open-source baselines due to their accessibility. However, they struggle considerably with cursive handwriting and heavily degraded documents due to their reliance on brittle segmentation algorithms.

The paradigm shifted with the introduction of Convolutional Neural Networks (CNNs) for feature extraction, which replaced manual engineering of shape descriptors with automatic representation learning. The subsequent introduction of Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, and the Connectionist Temporal Classification (CTC) loss function (Graves et al. 2006) allowed for end-to-end sequence recognition without explicit segmentation. This led to the popular Convolutional Recurrent Neural Network (CRNN) architecture (Shi, Bai, and Yao 2017).

More recently, transformer-based architectures have dominated academic research in text recognition. Models such as ABINet (Fang et al. 2021), ViTSTR (Atienza 2021), and TrOCR (Li et al. 2022) leverage self-attention mechanisms to capture complex spatial and linguistic dependencies in parallel. While models like TrOCR achieve state-of-the-art results by combining powerful visual encoders with language-aware decoders, they often demand substantial computational resources and large labelled corpora, making them difficult to deploy in decentralised, low-resource environments.

To balance accuracy and efficiency, the PARSeq (Permuted Autoregressive Sequence) architecture was introduced. It utilises a Vision Transformer (ViT) (Dosovitskiy et al. 2021) encoder and a permutation-based autoregressive decoder. By training on permuted reading orders, it learns ensembles of internal language models, making it highly robust to noise and occlusion while remaining lightweight and fast during inference.

Modern multimodal large language models and document-understanding systems, such as Donut or GPT-family models, can perform OCR-like tasks as part of broader image understanding. However, these systems are not optimised specifically for text recognition, consume disproportionate resources, and create privacy concerns when documents must be sent to third-party infrastructure.

Cyrillic and Bulgarian text recognition. High-accuracy Cyrillic OCR systems are rare, and most public resources target Russian and Kazakh. For Bulgarian, existing work focuses on printed and historical text: using Bulgarian language resources to post-correct OCR output (Kratchanov, Laskova, and Simov 2020), and a benchmark with methods for post-OCR correction of historical documents in pre-1945 orthography (Beshirov et al. 2025). Tesseract and EasyOCR handle printed Bulgarian reasonably well, but no system reliably recognises Bulgarian cursive, which is the gap we address.

3. Proposed Methodology

To handle the high variance of real-world documents, the system is designed as a modular pipeline: Image Input → Preprocessing → Text Detection → Structure Recognition → Text Recognition → Output. Full-document recognition was rejected due to the lack of suitable annotated datasets and the insufficiency of synthetic data alone for reliable real-world performance.

3.1 System Architecture and Preprocessing

Document images often suffer from noise, poor lighting, stains, and geometric distortions. The system implements a suite of selectable preprocessing techniques to normalise the input. Crucially, preprocessing is treated as a configurable search space rather than a fixed recipe. Because the effect of preprocessing cannot be predicted reliably in advance, the system allows arbitrary user-selected preprocessing sequences. These dependencies form a Directed Acyclic Graph (DAG) of valid combinations.

The image representation is another critical design choice. Images are stored in the LAB colour space, where the L channel acts as a luminance-based black-and-white representation, while the A and B channels preserve colour information. This allows grayscale-style algorithms to operate only on the L channel, while spatial transformations (such as deskewing or dewarping) are propagated consistently to all channels. This enables in-place processing with good cache locality and minimal copying.

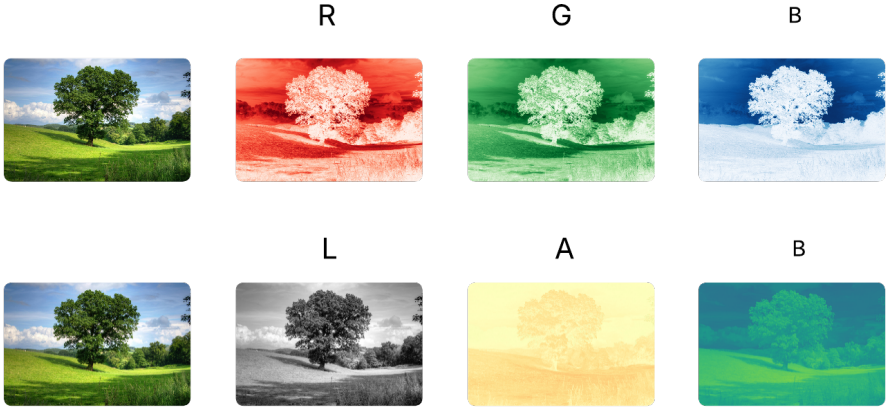


Figure 1
Comparison of processing in RGB vs. LAB colour space. Operating on the L channel preserves colour information while allowing grayscale algorithms to function correctly.

The implemented preprocessing techniques are categorised into four main areas to handle the wide variety of degradations found in historical and administrative documents:

- **Noise Reduction:** Documents often suffer from sensor noise, paper degradation, and ink bleed-through. The system implements linear filters (Average, Gaussian) for high-frequency noise, non-linear filters (Median) for salt-and-pepper noise, and advanced edge-preserving filters (Wiener, Bilateral, Non-Local Means) that smooth homogeneous regions while keeping text boundaries sharp.
- **Contrast Enhancement:** Faded ink and uneven illumination are addressed using Histogram Equalisation for global contrast and CLAHE (Contrast Limited Adaptive Histogram Equalisation) to prevent noise amplification in homogeneous areas while locally enhancing contrast.
- **Binarisation:** Converting images to black and white is essential for legacy OCR engines. The system includes global methods (Otsu's, Kittler-Illingworth) and local adaptive methods (Niblack's, Sauvola's) that calculate thresholds for each pixel based on its local neighbourhood, effectively separating text from varying backgrounds.
- **Geometric Correction:** Skewed and warped pages severely degrade recognition accuracy. The system uses the Hough Transform and Projection Profiles to detect and correct global page skew, and attempts Piecewise Vertical dewarping and Thin Plate Spline transformations to flatten curved text lines.

3.2 Text Detection and Structure Recognition

Accurate text recognition requires precise localisation. To locate text bounding boxes, the system supports either CRAFT (Baek et al. 2019) or YOLO-family text detectors (Jocher, Chaurasia, and Qiu 2024), exposed as interchangeable backends so users can choose whichever works best on their material without tying the write-up to a specific model version or training recipe.

For document structure recognition, we use an off-the-shelf layout detector from the YOLO family (Ultralytics 2023) (release/version not fixed in our pipeline) whose published weights were trained on *DocLayNet* (Pfitzmann et al. 2022); we did not train this component ourselves. It exposes high-level region classes (paragraphs, titles, lists, figures). However, most layout detection algorithms only recognise these broad semantic elements and do not support individual text line extraction. Furthermore, bounding-box regression models can be highly sensitive to document defects, often fragmenting paragraphs incorrectly when faced with skewed or degraded scans.

To address this, we cluster word-level detections into lines (and optionally group lines into blocks) with DBSCAN (Ester et al. 1996): the algorithm does not require a fixed number of clusters in advance and drops isolated noise. Radius ϵ is set proportional to row height to track font size, with boxes sorted by reading order. This is a lightweight geometric glue layer between detection and recognition, not a full document-understanding model.

Additionally, we integrate a YOLO-based table detector using weights trained on *PubTables-1M* (Smock, Pesala, and Abraham 2022) (released checkpoints; we did not train this submodule from scratch), extending the prototype toward broader document understanding.

3.3 Text Recognition Model Selection

Our approach relies on word-level recognition, offering the optimal compromise between visual difficulty, available data, latency, and the ability to leverage implicit language modelling.

Before selecting the final architecture, we conducted controlled experiments comparing three distinct paradigms: a custom CNN + Transformer Decoder, a Deformable CNN (DCNN) (Zhu et al. 2018) + Transformer Decoder, and PARSeq.

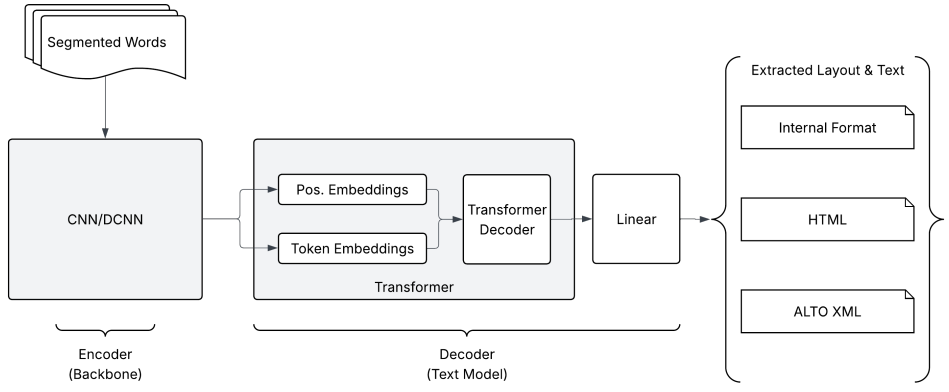


Figure 2
The baseline architecture evaluated during preliminary experiments, consisting of a ResNet-50 feature extractor coupled with a Transformer Decoder.

The CNN + Transformer Decoder served as our baseline. Our experiments revealed that a full encoder-decoder transformer is unnecessary; a lightweight transformer decoder paired with a strong CNN feature extractor is sufficient for the language-modelling component. We then hypothesised that Deformable Convolutions (DCNN), which adapt their receptive fields to the geometric variations of the input, would excel at capturing the high variance of cursive handwriting.

Surprisingly, replacing standard convolutions with DCNNs resulted in slower convergence and inferior accuracy (Figure 3). The deformable filters struggled to find consistent patterns in the highly irregular handwritten strokes.

Ultimately, PARSeq demonstrated better efficiency, faster convergence, and lower Character Error Rate (CER). The PARSeq model represents a substantial departure from traditional CRNN architectures by addressing the limitations of both purely autoregressive (AR) models (which suffer from error accumulation) and non-autoregressive (NAR) models (which lack language modelling capabilities).

PARSeq employs a Vision Transformer (ViT) as its visual encoder. The ViT processes the input image as a sequence of non-overlapping patches, utilising self-attention to capture global spatial dependencies across the entire image. This is particularly advantageous for

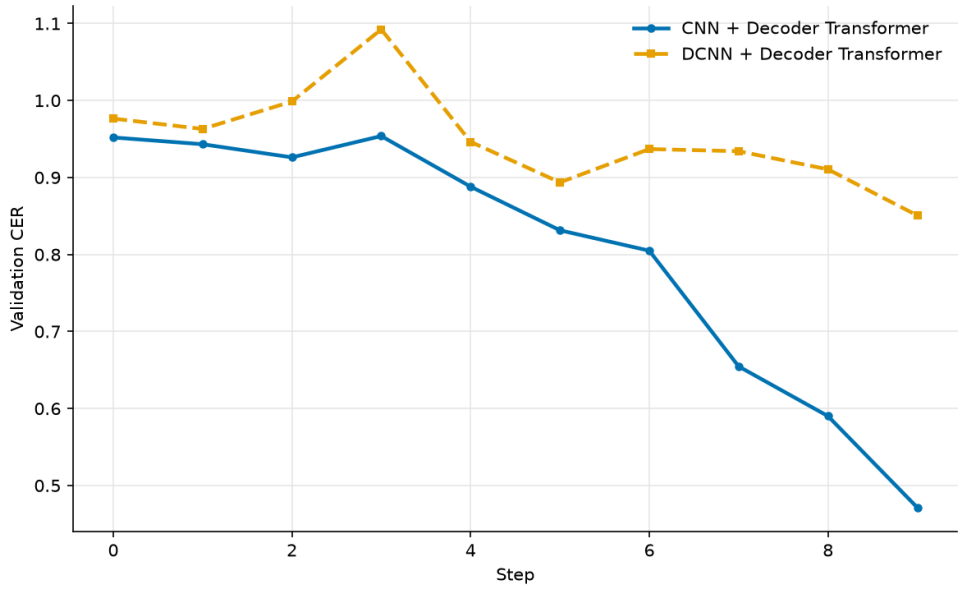


Figure 3
Validation *CER* for *CNN + DecoderTransformer* vs. *DCNN + DecoderTransformer* (same decoder). The DCNN encoder minimises error more slowly than the standard CNN backbone.

highly degraded text, where local features might be obscured, but global context can aid recognition.

The decoder is a lightweight, permutation-based autoregressive transformer. During training, PARSeq generates the target text sequence using multiple permuted reading orders. Predicting characters from arbitrary context subsets enables the model to learn an internal language model ensemble. During inference, it can operate in a standard left-to-right AR mode for maximum accuracy, or in an iterative refinement mode where it uses bidirectional context to correct initial predictions. This flexibility makes PARSeq robust to the high variance and occlusions typical of handwritten Cyrillic text.

Handling both printed and handwritten styles with a single model presents challenges like catastrophic forgetting. To mitigate the issue, we considered explicit style tokens (e.g. `[HAND]` and `[PRINT]`) as a lightweight way to condition a *single* recogniser so print and handwriting share weights instead of separate heads. This approach is applicable to all architectural variants, including PARSeq, but we did not pursue it, for three reasons. First, the architecture comparison had already consumed much of our experimental budget, and evaluating style conditioning properly would have required a separate set of experiments. Second, we expected little benefit: PARSeq is a strong recogniser and should handle both styles without an explicit switch, whereas a style token adds a single point of failure, since a misclassified style propagates errors through the rest of the pipeline. Third, the ViT encoder

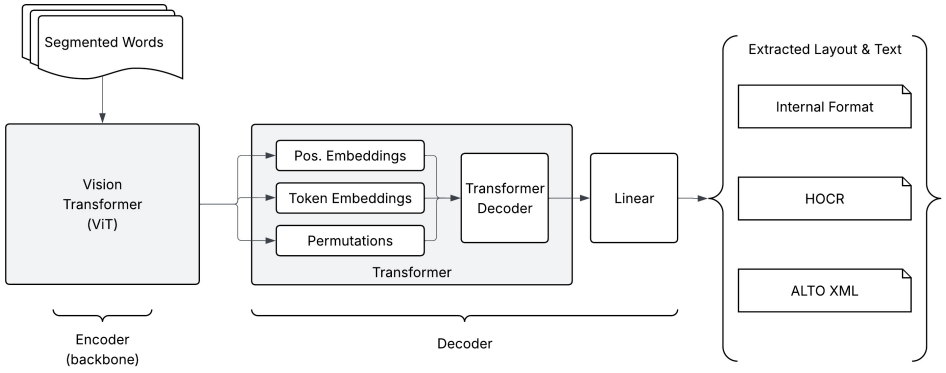


Figure 4
The PARSeq architecture, utilising a ViT encoder and a permutation-based autoregressive decoder.

already handles most of the recognition, so conditioning the smaller decoder would have limited effect.

3.4 Datasets and Training Curriculum

A major challenge in developing HTR systems for Bulgarian is the severe lack of annotated real-world datasets. To overcome this, we utilised a combination of synthetic and real-world Cyrillic datasets. For printed text, we utilised the MJSynth dataset (Jaderberg et al. 2014) (a massive collection of synthetic Latin scene text) to provide a strong baseline for visual feature extraction and regularisation. For printed Cyrillic, we aggregated several synthetic datasets (OCR Cyrillic Printed 1, 9, and 10), which contain millions of rendered words in various standard Cyrillic fonts.

For handwriting, due to the severe lack of real-world Bulgarian data, we generated four custom synthetic corpora based on free fonts that mimic different handwriting and typewritten styles: Pacifico (connected cursive), MarckScript (loose cursive), Comforter (informal handwriting), and GNU Typewriter (typewritten). The generation process involved rendering text with random variations in font size, spacing, and stroke width. During training, rendered line images were further augmented (Gaussian and motion blur, elastic warping, random paper textures and stains) to mimic degraded scans and historical documents. These were combined with a corpus of the 714,876 most frequent words from the Bulgarian Wikipedia (Kostov 2011). All single-letter words except the high-frequency prepositions/conjunctions were excluded. This exposes the model to Bulgarian vocabulary and orthographic patterns rather than relying only on foreign-language Cyrillic corpora.

For real-world handwriting, we incorporated the Cyrillic Handwriting Dataset (Werner 2021) (primarily Russian cursive) and the Kazakh Offline Handwritten Text Dataset (KOHTD) (Toiganbayeva et al. 2022). The KOHTD corpus was further processed to remove examples

containing characters specific to the Kazakh alphabet that were outside our defined support set.

Training a single model to recognise both styles requires careful curriculum learning. We utilised a four-phase strategy:

1. **Phase 0 (Alphabet Adaptation):** The primary goal of this initial phase was to warm up the newly initialised weights of the expanded classification head. The base PARSeq model, pre-trained on Latin text, was adapted to the Cyrillic alphabet. Weights for homoglyphs were transferred from the Latin model, while new Cyrillic characters were initialised randomly. Training utilised a 20% Latin and 80% Cyrillic printed mix for just 2 epochs to prevent the random weights from corrupting the pre-trained feature extractor.
2. **Phase 1 (Stabilisation):** With the classification head adapted, this phase performed full-network fine-tuning to solidify the model's understanding of Cyrillic typography. The model was trained for 10 epochs on a 25% Latin and 75% Cyrillic printed mix, establishing a stable, high-accuracy baseline on printed text before introducing the complex variance of handwriting.
3. **Phase 2 (Handwritten Introduction):** Introducing synthetic handwriting. The dataset mixture included the custom Bulgarian synthetic fonts alongside printed data to prevent forgetting. The model successfully learned to read synthetic handwriting despite heavy augmentations.
4. **Phase 3 (Real Handwritten Adaptation):** Fine-tuning exclusively on real-world handwritten datasets (KOHTD and Cyrillic Handwriting Dataset) to minimise the domain gap and handle the high variance of real human handwriting.

3.5 Implementation Details

The entire system was developed to run under resource constraints, ensuring it remains accessible to the general public, not just well-funded research institutions. The system is formally constrained to run locally on consumer-grade hardware on the order of 16 GB RAM and an 8-core CPU. This constraint guarantees privacy, as sensitive documents never leave the user's device.

To maximise extensibility, the backend exposes its functionality via a REST API. This allows the recognition engine to be integrated into broader document management systems or archival workflows. The system supports multiple standardised output formats: plain text for simple search and retrieval tasks, ALTO XML for describing the layout and content of physical text documents (preserving bounding box coordinates and reading order), and HOCR for embedding layout information directly into HTML.

Training deep learning models on millions of high-resolution images required a deliberate data pipeline. An initial attempt to *stream* examples directly from remote storage proved too slow for our setup and made proportional mixing across corpora difficult. We therefore converted all training corpora into compact *WebDataset* tar shards, stored them locally on the training machine, and fed them through PyTorch `Dataset`/`DataLoader` with multiple worker processes to overlap decoding, on-the-fly augmentation, and GPU compute.

4. Experiments and Results

We run two experiments. The first measures the final model on the held-out test splits of the training corpora, under conditions close to the training distribution, to establish an upper bound. The second measures it on real, unseen Bulgarian documents and compares it against widely used OCR tools and a local vision-language model, to test how far that accuracy carries to practical use.

We report CER and Word Error Rate (WER) throughout. Both are standard for OCR and HTR: CER captures character-level mistakes, and WER reflects whole-word usability, which is what matters for downstream search and indexing.

4.1 Test Set Performance

This first experiment measures the final model on the held-out test split of every training corpus, so accuracy is assessed under conditions that match training. To quantify uncertainty, we computed Wilson 95% confidence intervals (CIs) and two-proportion z-tests between the Phase 2 and final Phase 3 checkpoints on the real handwritten corpora. The final PARSeq model achieved strong results on the test splits of the training data. Table 2 summarises point estimates. Held-out sample sizes ranged from 14,831 characters (2,156 words) on the Cyrillic Handwriting Dataset to 5,398,419 characters (822,035 words) on Synthetic Cyrillic Large (PUMB AI 2024); normal approximations for the z-tests were appropriate throughout. Table 3 reports CIs at the final checkpoint. Table 4 summarises phase-to-phase significance. The model achieves high accuracy on stylistically controlled synthetic data, and scores on real-world sets provide an adequate baseline for generalisation to unseen Bulgarian cursive.

Dataset	CER (%)	WER (%)
MJSynth (Latin)	5.63	10.61
OCR Cyrillic Printed (Combined)	3.87	17.70
BulgarianPacifico	0.03	0.29
BulgarianMarckScript	0.03	0.23
BulgarianComforter	0.05	0.41
BulgarianGNUTypewriter	0.03	0.27
Synthetic Cyrillic Large	6.15	27.20
Cyrillic Handwriting Dataset	9.03	39.28
KOHTD	7.38	25.46

Table 2
Detailed Character Error Rate (CER) and Word Error Rate (WER) of the final PARSeq model on the test splits of the training corpora.

4.2 Real-World Comparison

To evaluate practical applicability, we constructed a custom test set. For printed text, we randomly selected highly degraded documents from a small subset of the Bulgarian National

Dataset	CER	95% CI	WER	95% CI
Cyrillic Handwriting	9.03%	[8.58, 9.50]	39.28%	[37.24, 41.35]
KOHTD	7.38%	[6.98, 7.80]	25.46%	[23.91, 27.07]

Table 3
Wilson 95% CIs for the final Phase 3 checkpoint on real handwritten test splits.

Dataset	Metric	Phase 2	Final	p (z-test)
Cyrillic Handwriting	CER	28.31%	9.03%	7.37×10^{-7}
Cyrillic Handwriting	WER	82.42%	39.28%	6.88×10^{-4}
KOHTD	CER	37.65%	7.38%	≈ 0
KOHTD	WER	86.58%	25.46%	≈ 0

Table 4
Phase 2 vs. final Phase 3 checkpoints on real handwritten corpora (two-proportion z-tests).

Radio's archive.[1] For handwritten text, we collected paragraphs randomly chosen from Wikipedia, featuring three distinct cursive styles. The ground truth for these samples was manually transcribed and verified by a native Bulgarian speaker. For these multi-line documents, predicted and ground-truth lines are aligned via the Hungarian algorithm on edit-distance costs before CER/WER are computed; unpaired lines count as full errors. We compared our PARSeq implementation against Tesseract and EasyOCR (for printed) and Qwen3-VL-4B-8bit (for handwritten).[2] To ensure a fair zero-shot evaluation on real-world document conditions, Tesseract and EasyOCR were evaluated out-of-the-box using their standard Bulgarian extensions and dictionaries without any further fine-tuning. Similarly, our PARSeq model was trained strictly on the training corpora splits and was not fine-tuned on any samples from this custom test set.

Results on Printed Text: As shown in Table 5, PARSeq outperformed both Tesseract and EasyOCR on degraded typewritten documents. The bidirectional context of PARSeq allowed it to correctly infer characters obscured by noise, whereas traditional OCR tools failed at the segmentation stage.

Results on Handwritten Text: Recognising real-world Bulgarian handwriting proved considerably more challenging. Table 6 compares PARSeq against Qwen3-VL-4B-8bit. While dedicated HTR architectures like TrOCR exist, they typically lack out-of-the-box support for Bulgarian cursive and would require resource-intensive pre-training. Instead, we selected Qwen3-VL-4B-8bit as our primary baseline because it was the strongest open-weights Vision-Language Model we identified that supports Bulgarian and still fits within our strict

1 These typewritten transcripts were provided by the Bulgarian National Radio through private correspondence for thesis research; their public disclosure would require separate permission.
2 Specifically, the MLX 8-bit checkpoint `https://huggingface.co/mlx-community/Qwen3-VL-4B-Instruct-8bit`, an 8-bit quantisation of `Qwen/Qwen3-VL-4B-Instruct`.

Model	CER (%)	WER (%)	Time (s)
Tesseract	12.9	33.5	8.2
EasyOCR	14.1	43.8	3.1
Ours	**9.4**	**28.1**	5.1

Table 5
Average results on degraded printed documents.

hardware constraints. Furthermore, benchmarking against a VLM reflects the current real-world trend of relying on general-purpose assistants for document understanding tasks. In this comparison, PARSeq outperformed the local VLM on the handwritten test set. Qwen3-VL suffered from context explosions and hallucinations on high-resolution images and sometimes stopped reading abruptly. Resource consumption was also much lower: roughly 4–6 GB vRAM for DocRec versus 12–16 GB vRAM for Qwen3-VL-4B-8bit.

Model	CER (%)	WER (%)
Qwen3-VL-4B-8bit	56.4	84.1
Ours	**17.2**	**45.3**

Table 6
Average results on handwritten documents across three cursive styles.

5. Discussion and Limitations

The Synthetic vs. Real Data Discrepancy: Loss curve analysis (Figure 5) reveals a negative correlation between synthetic and real handwriting data: while they share useful structure, synthetic cursive often fails to capture the variability, slant, and stroke continuity of real handwriting. This is reflected in the phase-wise accuracy trends (Table 7), where performance on real data improves substantially only during the final adaptation stage. Overall, data scarcity, rather than model complexity, remains the primary constraint, limited by the lack of large-scale, annotated Bulgarian HTR datasets.

Stage	Printed CER (%)	CHD CER (%)	KOHTD CER (%)
Phase 0	7.40	–	–
Phase 1	3.10	–	–
Phase 2	3.43	28.31	37.65
Phase 3	3.98	6.11	6.68
Final Test	3.87	9.03	7.38

Table 7
Phase-wise development of recognition accuracy. Real-handwriting performance improves sharply only in the final adaptation stage, justifying the staged curriculum.

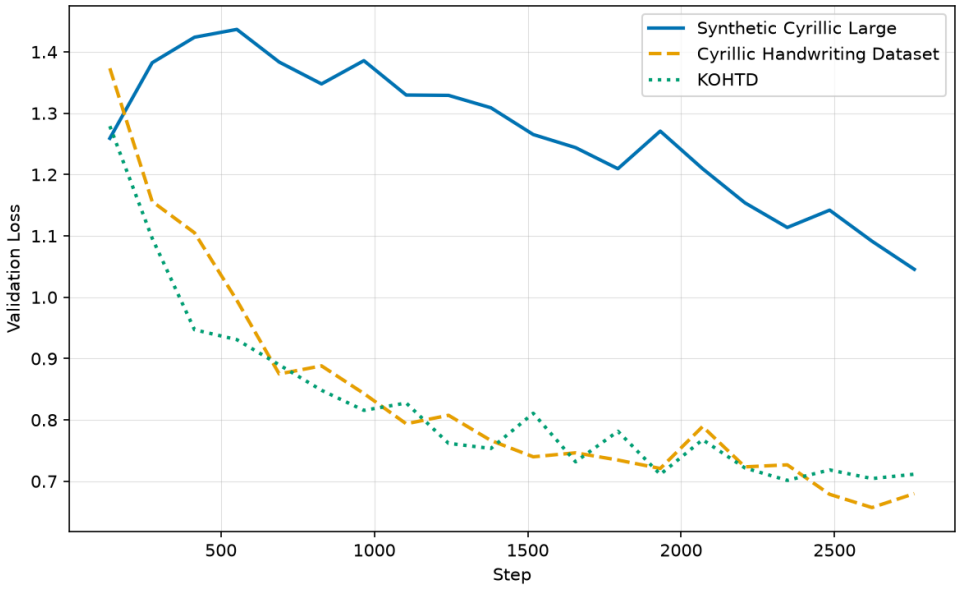


Figure 5
The observed discrepancy between validation loss on synthetic datasets and real-world handwritten datasets, highlighting the limitations of purely synthetic training.

Error Analysis: The real-world results are lower than the test-split results for two main reasons:

1. The real-world set is genuinely unseen Bulgarian handwriting; no real-world corpus of handwritten Bulgarian was available for training.
2. Even the real handwriting datasets suffer from limited variance. Their training, validation and test splits are mostly words segmented from forms, student essays and notebooks. The dataset details are unclear, but despite their large size these corpora were likely built from a relatively small number of independent documents that were then segmented into words. The words are therefore many, but the writers behind them may be few.

Per-character errors were limited mostly to confusions between visually similar letters. On the real-world samples these were о/с (where ink had worn away), г/т in plain handwriting, and ъ/ь. In cursive handwriting, where ligatures are denser, the most frequently confused pairs were а/о, и/ш, е/с, н/п and к/н. The high character and word error rates were mostly driven by ligatures in dense handwriting, where the model struggled to separate individual characters. This is somewhat surprising, considering the model does not rely on explicit segmentation and should, in principle, cope with connected letters. In certain limited cases, the slant of the lines also caused confusion, as words were appended to the wrong line.

Opportunities for Improvement: To improve accuracy on highly variable real-world Bulgarian handwriting, addressing the data deficit remains the most critical step. Building large, annotated corpora is essential and can be achieved through institutional data collection. For example, universities could systematically gather student-written materials and integrate lightweight annotation into coursework, enabling continuous corpus growth.

In parallel, exploring larger model architectures, combined with efficiency techniques such as quantisation or knowledge distillation, offers a practical path to increased capacity within fixed hardware constraints. Recent methods such as Google's TurboQuant further highlight the potential for optimisation without prohibitive computational cost.

The Role of Local VLMs in Document Processing: Our experiments with Qwen3-VL-4B-8bit highlight a crucial limitation in the current trend toward using general-purpose Vision-Language Models for specialised tasks like dense document OCR. While these models possess remarkable zero-shot capabilities, their autoregressive nature over massive context windows makes them highly inefficient for extracting dense, full-page text. The observed hallucinations and abrupt stopping are likely symptoms of the model losing its spatial grounding within the high-resolution image encoding. For tasks requiring deterministic, high-throughput text extraction, specialised architectures like PARSeq remain more practical in both speed and resource efficiency.

Threats to Validity: While the proposed pipeline demonstrates strong potential, several limitations must be acknowledged to contextualise the results fairly:

- **External Validity (Evaluation Size):** The real-world evaluation was conducted on a highly curated but small custom test set (10 printed pages, 12 handwritten paragraphs). While indicative of the model's potential capabilities, results should be interpreted as pilot evidence rather than definitive benchmarking.
- **Construct Validity (Baselines and Metrics):** The custom synthetic datasets lack realistic noise in their validation splits, leading to validation scores that mimic high-quality scans rather than degraded conditions. Furthermore, while Qwen3-VL-4B-8bit serves as a realistic local assistant baseline, it is not a dedicated HTR model.
- **Internal Validity (Pipeline Cascades):** The modular pipeline architecture is susceptible to cascading errors; failures in early stages (e.g. layout recognition fragility on degraded historical data) propagate to and multiply within the text recognition stage.

Ethical Considerations: All training corpora are used under their respective open licences. The system's local-only design mitigates privacy risks, but any HTR technology can be misused for bulk transcription of private correspondence and should be deployed with care.

6. Conclusion

In this paper, we presented a comprehensive, end-to-end software prototype for extracting printed and handwritten Cyrillic text from documents under real constraints. By integrating advanced preprocessing treated as a configurable search space, document structure recognition, and fine-tuning the PARSeq architecture via a staged training curriculum, the system addresses several practical limitations of existing tools.

Our results indicate that specialised, lightweight architectures can outperform standard baselines like Tesseract and EasyOCR on degraded printed text, and can be more stable and resource-efficient than local VLMs for handwriting recognition. The results suggest that specialisation and careful data design can rival larger general-purpose models under strict hardware and privacy constraints.

Crucially, our experiments show that the primary barrier to further progress in Bulgarian HTR is not algorithmic but the scarcity of real-world annotated data: synthetic data cannot fully bridge the domain gap to natural handwriting. Coordinated, community-driven data collection (e.g. university-sourced essay corpora) could thus be the most impactful next step.

Use of AI Tools

During the revision of this manuscript, the authors used a general-purpose AI assistant (Anthropic's Claude) to check the bibliography for missing fields, to normalise text formatting and correct spelling mistakes. The study's methods, experiments, results, and conclusions are entirely the authors' own. All AI-assisted changes were reviewed, edited, and verified by the authors, who take full responsibility for the content.

Acknowledgments

This work is partially supported by the project UNITe BG16RFPR002-1.014-0004 funded by PRIDST and EU NextGenerationEU project, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project SUMMIT, No BG-RRP-2.004-0008.

References

Atienza, Rowel. 2021. Vision transformer for fast and efficient scene text recognition. In *Proceedings of the 16th International Conference on Document Analysis and Recognition (ICDAR 2021), Lausanne, Switzerland, Part I*, volume 12821 of *Lecture Notes in Computer Science*, pages 319–334, Springer.

Baek, Youngmin, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character region awareness for text detection. In *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019), Long Beach, CA, USA*, pages 9365–9374, Computer Vision Foundation / IEEE.

Bautista, Darwin and Rowel Atienza. 2022. Scene text recognition with permuted autoregressive sequence models. In *Proceedings of the 17th European Conference on Computer Vision (ECCV 2022), Tel Aviv, Israel*, volume 13688 of *Lecture Notes in Computer Science*, pages 178–196, Springer.

Beshirov, Angel, Milena Dobreva, Dimitar Dimitrov, Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2025. Post-OCR text correction for Bulgarian historical documents. *International Journal on Digital Libraries*, 26.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021), Virtual Event, Austria*.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*, pages 226–231, AAAI Press.

Fang, Shancheng, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. 2021. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2021)*, pages 7098–7107, Computer Vision Foundation / IEEE.

Graves, Alex, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, Association for Computing Machinery, New York, NY, USA.

Jaderberg, Max, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Synthetic data and artificial neural networks for natural scene text recognition. In *Proceedings of the Workshop on Deep Learning (NIPS)*. ArXiv, abs/1406.2227.

Jocher, Glenn, Ayush Chaurasia, and Jing Qiu. 2024. YOLO by Ultralytics. https://github.com/ultralytics/ultralytics.

Kittinaradorn, Rakpong et al. 2020. EasyOCR. https://github.com/JaidedAI/EasyOCR.

Kostov, Nikolay. 2011. Analiz na balgarskiya ezik chrez Wikipedia (in Bulgarian). https://nikolay.it/Blog/2011/08/%D0%90%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7-%D0%BD%D0%B0-%D0%B1%D1%8A%D0%BB%D0%B3%D0%B0%D1%80%D1%81%D0%BA%D0%B8%D1%8F-%D0%B5%D0%B7%D0%B8%D0%BA-%D1%87%D1%80%D0%B5%D0%B7-Wikipedia/3.

Kratchanov, Ivan, Laska Laskova, and Kiril Simov. 2020. Towards improving OCR accuracy with Bulgarian language resources. In *Proceedings of Twin Talks 3: Understanding and Facilitating Collaboration in Digital Humanities (DH 2020)*, volume 2717 of *CEUR Workshop Proceedings*, pages 115–123.

Li, Minghao, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2022. TrOCR: Transformer-based optical character recognition with pre-trained models. ArXiv, abs/2109.10282.

Pfitzmann, Birgit, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter W. J. Staar. 2022. DocLayNet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2022), Washington, DC, USA*, pages 3743–3751, ACM.

PUMB AI. 2024. Synthetic Cyrillic Large. https://huggingface.co/datasets/pumb-ai/synthetic-cyrillic-large. Apache-2.0 license.

Qwen Team. 2025. Qwen3-VL technical report. ArXiv, abs/2511.21631.

Shi, Baoguang, Xiang Bai, and Cong Yao. 2017. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304.

Smith, Ray. 2007. An overview of the Tesseract OCR engine. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Paraná, Brazil*, volume 2, pages 629–633, IEEE.

Smock, Brandon, Rohith Pesala, and Robin Abraham. 2022. PubTables-1M: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), New Orleans, LA, USA*, pages 4634–4642, IEEE.

Toiganbayeva, Nazgul, Mahmoud Kasem, Galymzhan Abdimanap, Kairat Bostanbekov, Abdelrahman Abdallah, Anel Alimova, and Daniyar Nurseitov. 2022. KOHTD: Kazakh offline handwritten text dataset. *Signal Processing: Image Communication*, page 116827.

Ultralytics. 2023. Ultralytics YOLO. https://github.com/ultralytics/ultralytics.

Werner, Constantin. 2021. Cyrillic Handwriting Dataset. https://www.kaggle.com/datasets/constantinwerner/cyrillic-handwriting-dataset.

Zhu, Xizhou, Han Hu, Stephen Lin, and Jifeng Dai. 2018. Deformable ConvNets v2: More deformable, better results. ArXiv, abs/1811.11168.

Appendices

A. Real-world handwritten examples

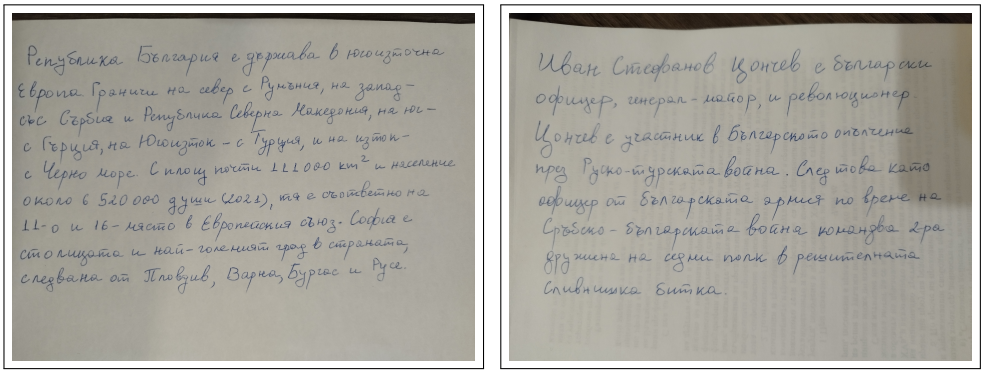


Figure 6
Examples of real-world handwritten documents from the custom test set, illustrating variation in slant, stroke connection, spacing, and line regularity.

B. Detailed Dataset Composition

Table 8 provides a comprehensive breakdown of the datasets utilised for training and testing the text recognition models. The synthetic handwritten datasets were generated using custom scripts to mimic various Bulgarian cursive styles, substantially expanding the available training data for low-resource Cyrillic handwriting.

Dataset	Style	Type	Train. subset	Test. subset
MJSynth	printed	synthetic	7,220,000	892,000
OCR Cyrillic Printed 1, 9, 10	printed	synthetic	2,700,000	300,000
BulgarianPacifico	handwritten	synthetic	643,364	71,484
BulgarianMarckScript	handwritten	synthetic	643,364	71,484
BulgarianComforter	handwritten	synthetic	643,364	71,484
BulgarianGNUTypewriter	printed	synthetic	643,364	71,484
Synthetic Cyrillic Large	handwritten	synthetic	3,120,000	779,000
Cyrillic Handwriting Dataset	handwritten	real	72,285	1,543
KOHTD	handwritten	real	49,497	5,500

Table 8
Summary of the datasets used for training and testing the recognition models.

C. Dataset Sources

Table 8 lists the datasets used in this work. The publicly available datasets can be obtained from the following sources:

- MJSynth:
  https://huggingface.co/datasets/priyank-m/MJSynth_text_recognition
- OCR Cyrillic Printed 1:
  https://huggingface.co/datasets/DonkeySmall/OCR-Cyrillic-Printed-1
- OCR Cyrillic Printed 9:
  https://huggingface.co/datasets/DonkeySmall/OCR-Cyrillic-Printed-9
- OCR Cyrillic Printed 10:
  https://huggingface.co/datasets/DonkeySmall/OCR-Cyrillic-Printed-10
- Synthetic Cyrillic Large:
  https://huggingface.co/datasets/pumb-ai/synthetic-cyrillic-large
- Cyrillic Handwriting Dataset:
  https://www.kaggle.com/datasets/constantinwerner/cyrillic-handwriting-dataset
- KOHTD:
  https://github.com/abdoelsayed2016/KOHTD

The custom synthetic Bulgarian handwriting datasets generated for this work are publicly available:

- BulgarianPacifico:
  https://huggingface.co/datasets/KotakaDanski/BulgarianPacifico
- BulgarianMarckScript:
  https://huggingface.co/datasets/KotakaDanski/BulgarianMarckScript
- BulgarianComforter:
  https://huggingface.co/datasets/KotakaDanski/BulgarianComforter
- BulgarianGNUTypewriter:
  https://huggingface.co/datasets/KotakaDanski/BulgarianGNUTypewriter

D. Supported Alphabet

The system supports a carefully curated alphabet designed to handle both standard modern Bulgarian and common historical Cyrillic characters, as well as Latin fragments frequently found in administrative documents.

A–I		J–R		S–Z	
Uppercase	**Lowercase**	**Uppercase**	**Lowercase**	**Uppercase**	**Lowercase**
A	a	J	j	S	s
B	b	K	k	T	t
C	c	L	l	U	u
D	d	M	m	V	v
E	e	N	n	W	w
F	f	O	o	X	x
G	g	P	p	Y	y
H	h	Q	q	Z	z
I	i	R	r		

Table 9
The supported Latin alphabet subset.

А–Й		К–У		Ф–Я + Archaic	
Uppercase	**Lowercase**	**Uppercase**	**Lowercase**	**Uppercase**	**Lowercase**
А	а	К	к	Ф	ф
Б	б	Л	л	Х	х
В	в	М	м	Ц	ц
Г	г	Н	н	Ч	ч
Д	д	О	о	Ш	ш
Е	е	П	п	Щ	щ
Ж	ж	Р	р	Ъ	ъ
З	з	С	с	Ь	ь
И	и	Т	т	Ю	ю
Й	й	У	у	Я	я
				Ѣ	ѣ
				Ꙗ	ꙗ

Table 10
The supported Cyrillic alphabet subset, including historical characters for archival document processing.

E. Table Detector Training Results

Table 11 reports training validation metrics on a 100,000-document subset of *PubTables-1M* (Smock, Pesala, and Abraham 2022) (8 epochs).

Ep.	Precision	Recall	mAP50	mAP50-95	Val box	Val cls	Val dfl
1	0.7182	0.6989	0.7322	0.5840	0.8186	0.6128	0.0121
2	0.8233	0.7897	0.8255	0.6891	0.6472	0.4516	0.0077
3	0.8778	0.8246	0.8724	0.7480	0.5536	0.3701	0.0060
4	0.9117	0.8448	0.8954	0.7878	0.4825	0.3132	0.0055
5	0.9260	0.8609	0.9082	0.8138	0.4457	0.2796	0.0046
6	0.9232	0.8707	0.9129	0.8254	0.4200	0.2600	0.0044
7	0.9339	0.8781	0.9207	0.8379	0.3999	0.2428	0.0041
8	0.9375	0.8821	0.9234	0.8430	0.3924	0.2331	0.0040

Table 11
Training validation metrics on a subset of PubTables-1M.

F. Training Curriculum Dataset Proportions

The four-phase curriculum learning strategy required carefully balancing the proportions of synthetic and real datasets to prevent catastrophic forgetting while adapting to new domains.

Dataset	Phase 0	Phase 1	Phase 2	Phase 3
MJSynth	0.20	0.25	0.05	0.05
OCR Cyrillic Printed	0.80	0.75	0.25	0.10
Synthetic Cyrillic Large	–	–	0.30	0.10
BulgarianPacifico	–	–	0.10	0.05
BulgarianMarckScript	–	–	0.10	0.05
BulgarianComforter	–	–	0.10	0.05
BulgarianGNUTypewriter	–	–	0.10	0.05
Cyrillic Handwriting Dataset	–	–	–	0.35
KOHTD	–	–	–	0.20

Table 12
Proportions of the training datasets across the four phases of the PARSeq training curriculum.

G. Layout Recognition Algorithms

The system offers two interchangeable families for text detection: CRAFT (Character Region Awareness for Text Detection) and YOLO-based detectors.

CRAFT is a robust text detector that localises characters and links them into lines, which suits unconstrained layouts.

YOLO-family detectors provide fast region proposals; users can swap implementations or checkpoints as needed alongside CRAFT.

H. Training Curriculum Details

The training of the PARSeq model was conducted in four distinct phases to ensure stable convergence and prevent catastrophic forgetting. The dataset proportions, hyperparameters, and results for each phase are detailed below.

H.1 Phase 0: Alphabet Adaptation

Dataset	Proportion
MJSynth	0.20
OCR Cyrillic Printed	0.80

Table 13
Proportions of the training datasets for *PARSeq* in Phase 0.

Hyperparameter	Value
Epochs	2
Batch Size	512
Learning Rate	1×10^{-4}
Weight Decay	1×10^{-2}
Warmup	5%
Patience	5
Workers	8

Table 14
Hyperparameters for training *PARSeq* in Phase 0.

Dataset	CER (%)	WER (%)
MJSynth	7.92	13.30
OCR Cyrillic Printed	7.40	23.37

Table 15
Results from the training of *PARSeq* in Phase 0.

H.2 Phase 1: Printed Cyrillic Stabilisation

Dataset	Proportion
MJSynth	0.25
OCR Cyrillic Printed	0.75

Table 16
Proportions of the training datasets for *PARSeq* in Phase 1.

Hyperparameter	Value
Epochs	10
Batch Size	512
Learning Rate	1×10^{-4}
Weight Decay	1×10^{-2}
Warmup	5%
Patience	5
Workers	8

Table 17
Hyperparameters for training *PARSeq* in Phase 1.

Dataset	CER (%)	WER (%)
MJSynth	4.29	7.74
OCR Cyrillic Printed	3.10	15.76

Table 18
Results from training *PARSeq* in Phase 1.

H.3 Phase 2: Handwritten Introduction

Dataset	Proportion
MJSynth	0.05
OCR Cyrillic Printed (1, 9, 10)	0.25
Synthetic Cyrillic Large	0.30
BulgarianPacifico	0.10
BulgarianMarckScript	0.10
BulgarianComforter	0.10
BulgarianGNUTypewriter	0.10

Table 19
Proportions of the training datasets for *PARSeq* in Phase 2.

Hyperparameter	Value
Max Epochs	5
Batch Size	512
Learning Rate	5×10^{-5}
Weight Decay	1×10^{-2}
Warmup	5%
Patience	5
Workers	8

Table 20
Hyperparameters for training *PARSeq* in Phase 2.

Dataset	CER (%)	WER (%)
MJSynth	5.17	9.69
OCR Cyrillic Printed 1	2.69	9.67
OCR Cyrillic Printed 9	7.10	48.83
OCR Cyrillic Printed 10	2.56	9.81
OCR Cyrillic Printed (concat)	3.43	17.02
Synthetic Cyrillic Large	5.72	25.74
Bulgarian Pacifico	0.03	0.29
Bulgarian MarckScript	0.02	0.19
Bulgarian Comforter	0.04	0.36
Bulgarian GNU Typewriter	0.03	0.25
Cyrillic Handwriting	28.31	82.42
KOHTD	37.65	86.58

Table 21
Results from training *PARSeq* in Phase 2.

H.4 Phase 3: Real Handwritten Adaptation

Dataset	Proportion
MJSynth	0.05
OCR Cyrillic Printed (1, 9, 10)	0.10
Synthetic Cyrillic Large	0.10
BulgarianPacifico	0.05
BulgarianMarckScript	0.05
BulgarianComforter	0.05
BulgarianGNUTypewriter	0.05
Cyrillic Handwriting Dataset	0.35
KOHTD	0.20

Table 22
Proportions of the training datasets for *PARSeq* in Phase 3.

Hyperparameter	Value
Epochs	10–20
Batch Size	512
Learning Rate	5×10^{-4}
Weight Decay	1×10^{-2}
Warmup	5%
Patience	5
Workers	8

Table 23
Hyperparameters for training *PARSeq* in Phase 3.

Dataset	CER (%)	WER (%)
MJSynth	5.50	10.49
OCR Cyrillic Printed	3.98	17.67
Synthetic Cyrillic Large	6.13	27.25
Cyrillic Handwriting	6.11	23.92
KOHTD	6.68	24.87

Table 24
Validation metrics from training *PARSeq* in Phase 3 (final).

I. Structure and Table Recognition Algorithms

For document structure recognition, we use an off-the-shelf YOLO-family layout detector (no fixed version string in our deployment) whose released weights were trained on *DocLayNet* (Pfitzmann et al. 2022); we did not train this module. It exposes the following classes: Caption, Footnote, Formula, List-item, Page-footer, Page-header, Picture, Section-header, Table, Text, and Title.

For tables lacking borders or complex spanning cells, we integrate a YOLO-based detector using weights trained on a subset of 100,000 documents from *PubTables-1M* (Smock, Pesala, and Abraham 2022) (again: third-party weights; not trained in this work).

J. Validation Metrics during Training

Figures 7–11 show validation curves by training phase. Each figure stacks two plots vertically (CER on top; WER or training loss below). Phase 3 is shown twice because we validate separately on the *Cyrillic Handwriting Dataset* and on *KOHTD*.

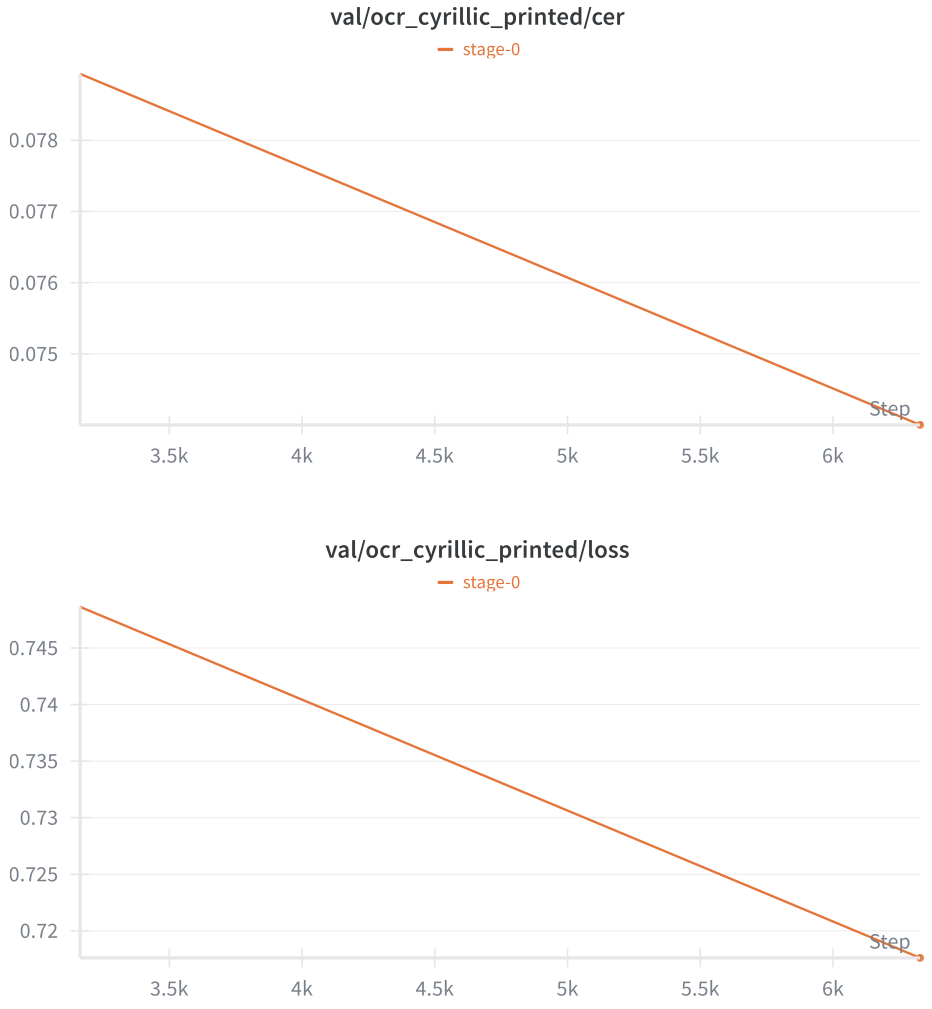


Figure 7
Phase 0 (*OCR Cyrillic Printed*): CER (top) and training loss (bottom).

Figure 8
Phase 1 (*OCR Cyrillic Printed*): CER (top) and WER (bottom).

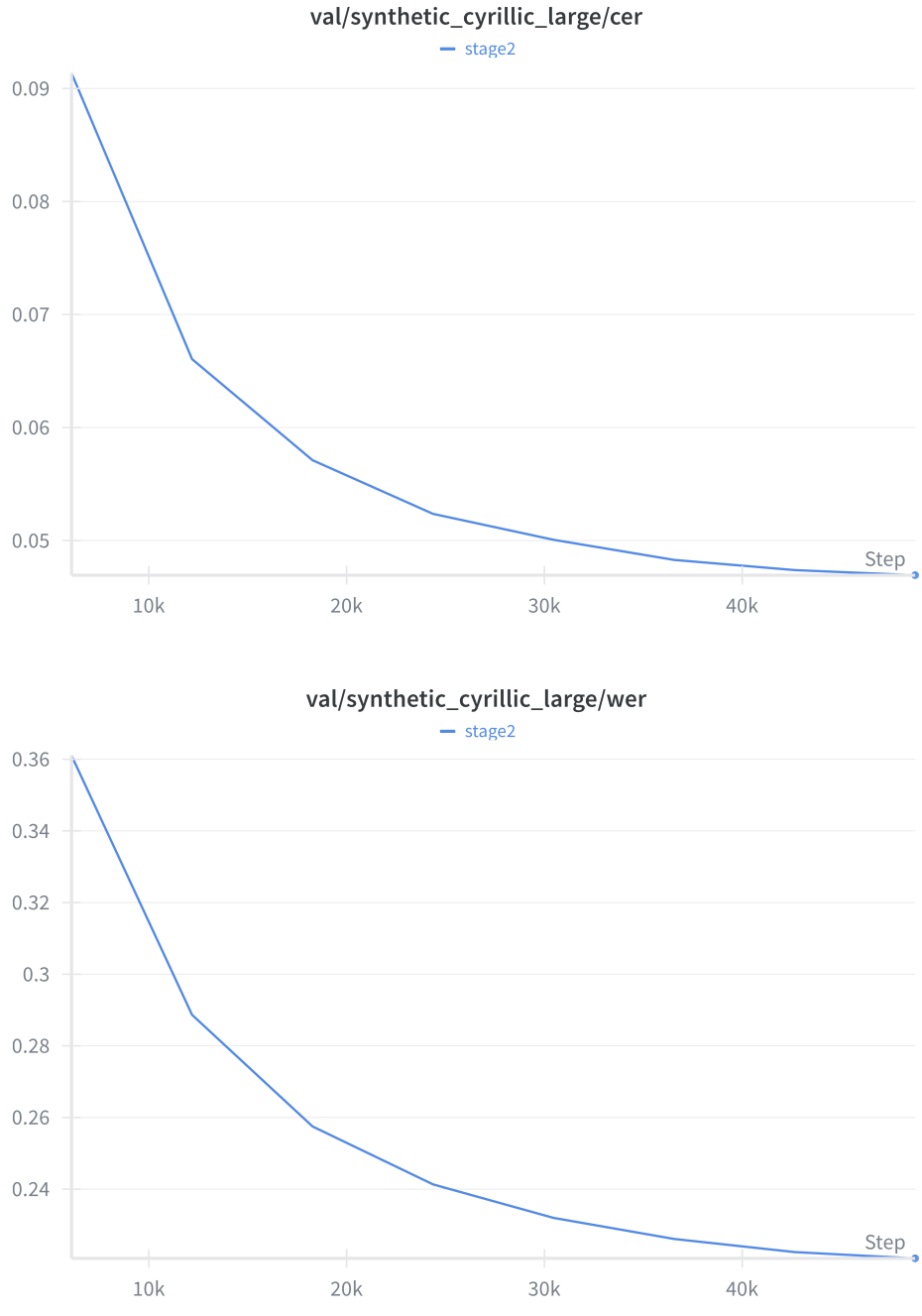


Figure 9
Phase 2 (*Synthetic Cyrillic Large*): CER (top) and WER (bottom).

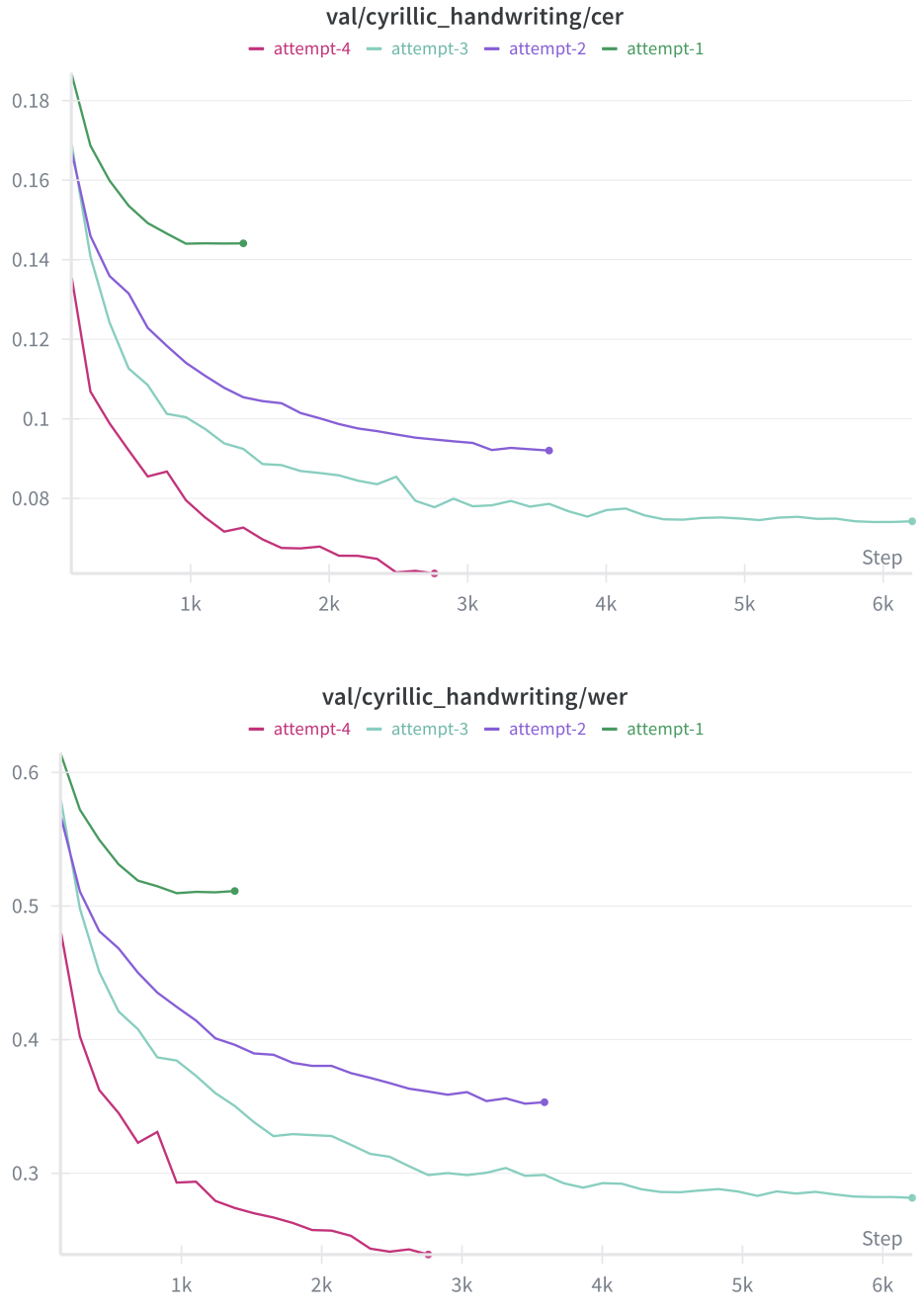


Figure 10
Phase 3, *Cyrillic Handwriting Dataset* validation split: CER (top) and WER (bottom).

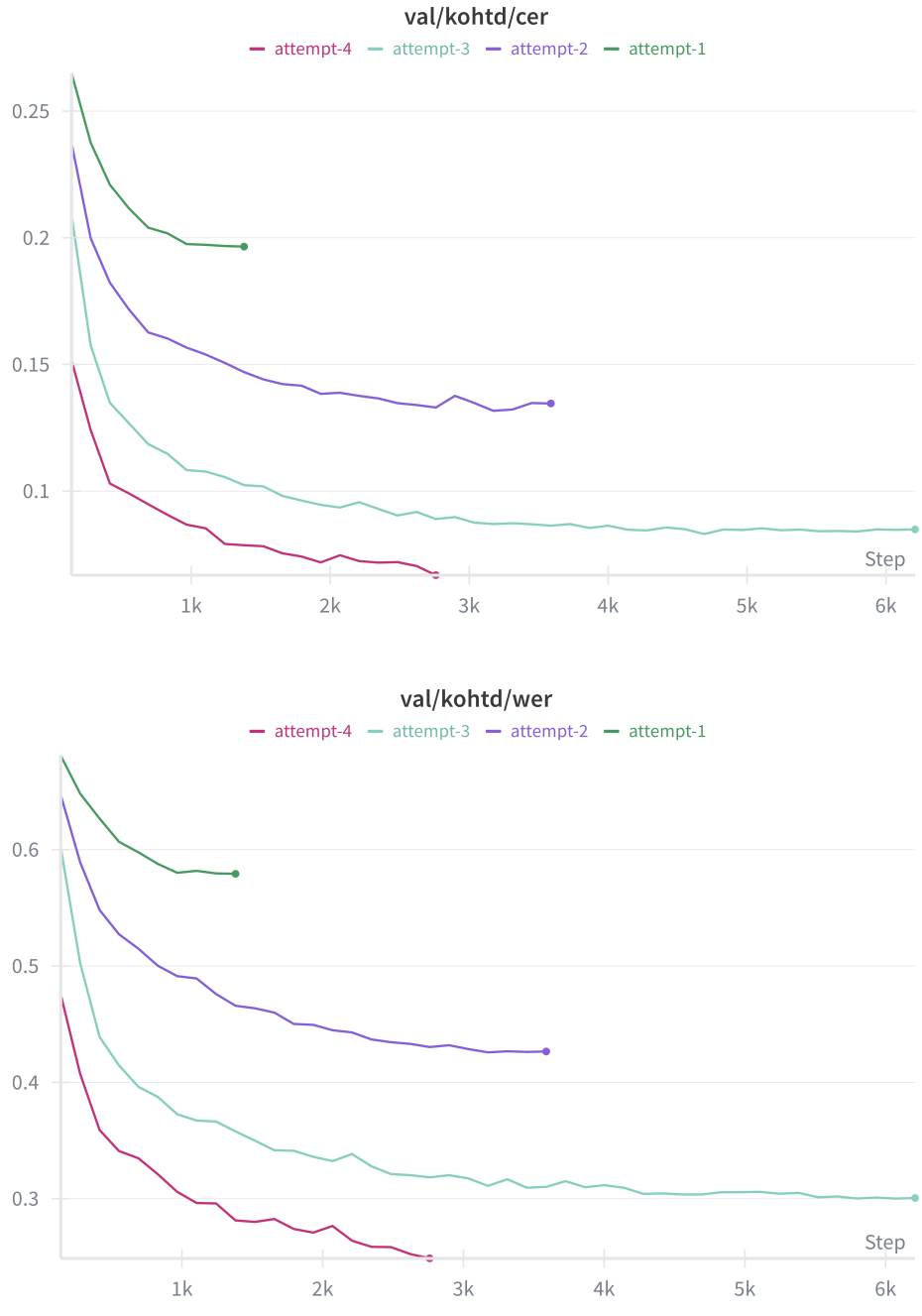


Figure 11
Phase 3, *KOHTD* validation split: CER (top) and WER (bottom).

K. Preprocessing Effects on Printed Pages

The evaluation of individual preprocessing operations on printed test pages confirms that preprocessing should remain optional and configurable. Different algorithms address specific defects, such as blur, skew, or weak contrast.

Algorithm	Doc.	Metric	Before → After (%)
Average Filter	3070-5	WER	35.66 → 25.19
Gaussian Filter	3070-3	CER	11.76 → 6.49
Hough Transform	16-3	CER	10.69 → 4.77
Projection Profile	3070-4	WER	27.27 → 22.51
Histogram Equalisation	3070-3	WER	31.69 → 27.46
CLAHE	16-2	WER	21.68 → 20.28

Table 25
Representative best-case improvements from optional preprocessing on printed documents.

K.1 Additional Architectural Comparisons

During the preliminary experiments, we compared the convergence and performance of several architectures before selecting PARSeq. The following figures illustrate these comparisons.

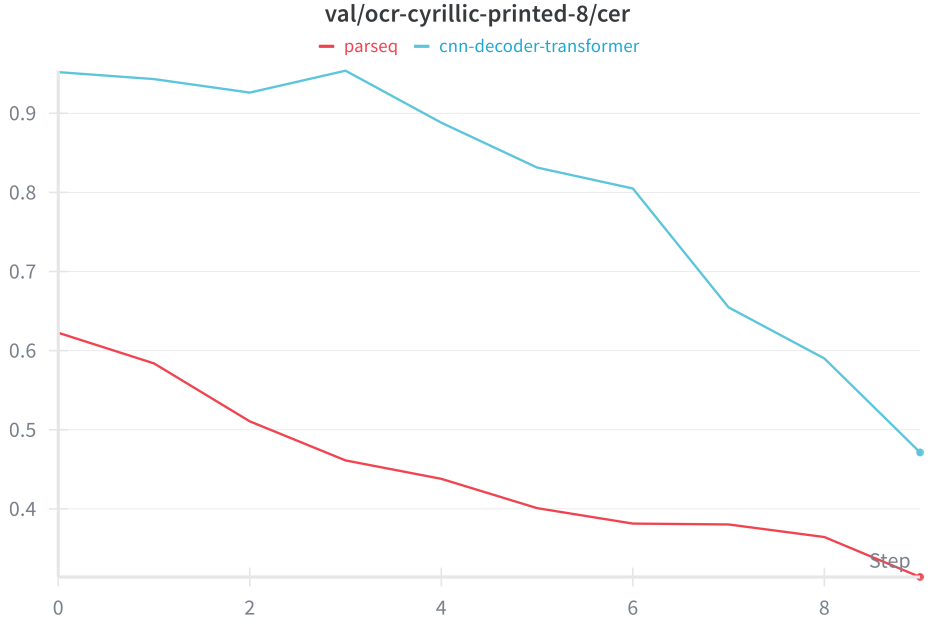


Figure 12
Preliminary experiments comparing PARSeq and CNN–Decoder Transformer on character error rate (CER).

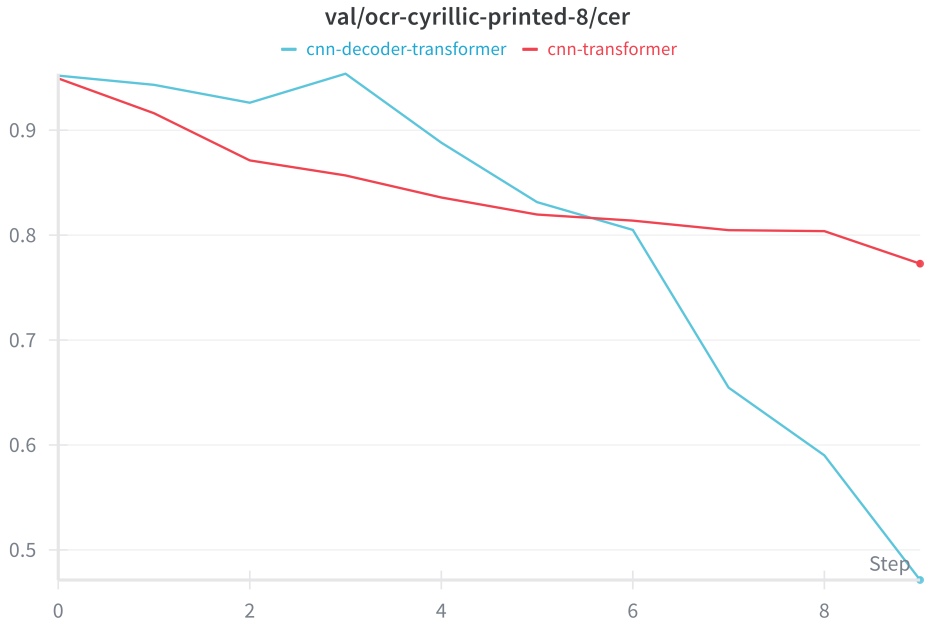


Figure 13
Preliminary experiments comparing CNN–Decoder Transformer and CNN–Transformer on character error rate (CER).

Computational Linguistics in Bulgaria
Volume 2, Issue 1, 2026

Editor-in-Chief: Svetla Koeva
Cover artwork by Bozhidar Chemshirov

ISSN 3033-1382 (print) | 3033-2397 (online)

Press sheets: 10.00
Print run: 100
Format: 70/100/16