ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF CLASSICAL PHILOLOGY
AND ITALIAN STUDIES

Second Cycle Degree in
Digital Humanities and Digital Knowledge

Automated Essay Scoring for Russian as a Second Language: A Deep Ordinal Learning Approach

Dissertation in
Natural Language Processing

Supervisor:
Prof. Fabio Tamburini

Co-Supervisor:
Prof. Mikhail Kopotev
(University of Helsinki)

Defended by:
Nikolai Gorbachev
Matricola: 0001073852

Graduation Session III
Academic Year 2024/2025

Abstract

Automated Essay Scoring (AES) systems are critical for scaling language assessment, yet most research focuses on English and large-scale datasets. This thesis addresses the challenge of developing reliable AES models for Russian as a Second Language (L2), a morphologically rich language, within the resource-constrained context of institutional learning. The study utilizes a real-world dataset of approximately 1,100 learner essays from the Russian language course at the Middlebury Language School (VT, USA), rated according to the ACTFL proficiency guidelines.

The research systematically evaluates three modeling paradigms: (1) feature-based statistical models relying on engineered linguistic metrics (e.g., syntactic complexity, lexical diversity); (2) deep representation learning using fine-tuned multilingual Transformers (XLM-RoBERTa); and (3) hybrid fusion strategies. A central innovation of this work is the application of Ordinal Regression objectives, specifically Consistent Rank Logits (CORAL) and Conditional Ordinal Regression (CORN), to explicitly model the ordered nature of proficiency levels, contrasting them with standard nominal classification and linear regression.

Results demonstrate that the Deep Ordinal approach (XLM-RoBERTa + CORAL) achieves state-of-the-art performance, substantially outperforming feature-based baselines and standard cross-entropy models. Error analysis reveals that while hand-crafted features provide interpretability, they are redundant when paired with high-capacity Transformer encoders, which implicitly learn these linguistic constructs. The study concludes that aligning loss functions with the ordinal structure of the ACTFL scale is the most effective strategy for mitigating data scarcity in L2 assessment.

Keywords: Automated Essay Scoring, Russian L2, Natural Language Processing, Deep Ordinal Learning, XLM-RoBERTa, Linguistic Complexity, Language Assessment, ACTFL Framework, Transformers

Contents

Chapter 1

Introduction

1.1 Background and Motivation

Automated Essay Scoring (AES) refers to the task of automatically evaluating written texts to estimate a writer's language proficiency or writing quality. In the context of Second Language (L2) learning, AES systems aim to assign proficiency levels to learner-produced essays, approximating the judgments traditionally made by trained human assessors. Such systems are increasingly relevant in educational, institutional, and professional settings, where written language proficiency is often a prerequisite for course placement, certification, immigration procedures, or employment.

The motivation for automated assessment is both practical and pedagogical. From a practical perspective, manual assessment is time-consuming, costly, and difficult to scale. Educational institutions frequently face constraints in expert assessor availability, especially when managing large cohorts or repeated assessment cycles. Automated systems alleviate this burden by providing fast, low-cost, and consistent evaluations. From a pedagogical standpoint, timely feedback on proficiency allows learners to monitor their progress, understand their current abilities, and select appropriate learning materials.

Importantly, written essays offer a particularly rich signal for assessment. Unlike multiple-choice tests or isolated grammar exercises, essay writing reflects a learner's ability to organize ideas, select appropriate vocabulary, construct syntactically complex sentences, and maintain coherence across longer stretches of discourse. Consequently, essay-based assessment remains a cornerstone of high-stakes and institutional frameworks. The challenge, however, lies in translating these complex, multi-dimensional human judgments into reliable automated systems.

These challenges are particularly acute in real-world educational contexts involving languages other than English. In such settings, annotated learner corpora are typically small, unevenly distributed across proficiency levels, and costly to collect, limiting the applicability of data-hungry benchmarks.

In this thesis, we focus on the automated assessment of learner-produced essays in **Russian as a Second Language**, using a real-world, small, and inherently skewed dataset ($N \approx 1,100$). The data originate from writing assessments administered at the Middlebury Language Schools in the United States and are aligned with the ACTFL proficiency framework. This choice of language and dataset reflects the ecological reality faced by many educational institutions: the need to build robust tools without access to large-scale annotated corpora. By grounding our study in authentic learner

texts and operating under strict constraints—morphologically rich target language, limited data, and open-source architecture—we aim to provide insights into the design of practical AES systems for low-resource educational contexts.

1.2 Proficiency Assessment and the ACTFL Framework

To automate assessment, one must first understand the human evaluation process it seeks to emulate. The Middlebury curriculum aligns with the **ACTFL Proficiency Guidelines** (American Council on the Teaching of Foreign Languages). The ACTFL scale categorizes ability into major levels ranging from *Novice* to *Superior*. With the exception of Superior, these levels are subdivided into Low, Mid, and High sublevels.

Crucially, these categories are inherently **ordinal**: higher levels presuppose mastery of lower ones, but the "distance" between adjacent levels is not uniform. For example, the pedagogical leap from *Intermediate High* to *Advanced Low* (crossing the "Intermediate Plateau") is widely considered more challenging than the progression from *Novice Low* to *Novice Mid*. While European institutions typically utilize the CEFR (Common European Framework of Reference), ACTFL is the dominant standard in US academia. The two frameworks are theoretically interoperable, but ACTFL offers a more granular subdivision at the lower levels, providing a richer learning signal for modeling.

It is also vital to distinguish proficiency assessment from standard *readability* assessment. A "Level A2" textbook passage is written by a professional to be simple and correct. A "Level A2" learner essay, conversely, is characterized by specific patterns of failure: grammatical breakdowns, phonetic spelling errors, and L1-transfer interference. Genuine learner errors cannot be reliably synthesized, meaning real-world AES systems must operate on learner corpora that are orders of magnitude smaller than standard NLP datasets. This makes efficient data utilization the primary engineering constraint.

In practice, human assessment relies on detailed rubrics evaluating dimensions such as grammatical accuracy, lexical range, syntactic complexity, and coherence. Some dimensions are countable (e.g., error frequency, sentence length), while others are latent and holistic (e.g., argumentative flow). This duality presents a fundamental difficulty for AES: while feature engineering can capture the countable metrics, modern neural language models (Transformers) are required to capture the latent semantic properties. This thesis systematically investigates the comparative and combined utility of these two paradigms.

1.3 Language-Specific and Institutional Challenges

Developing an AES system for a Russian language program involves both language-internal and institutional constraints that distinguish this setting from the large-scale, English-centric conditions typically assumed in AES research.

1. **Morphological Richness:** Russian is a fusional language with a rich case system and flexible word order. Grammatical errors often manifest as broken dependency relations (e.g., incorrect

case endings) rather than simple word omissions. This challenges shallow n-gram approaches that rely on fixed surface patterns.

2. **Data Scarcity and Imbalance:** Real-world educational data reflects the natural "funnel" of enrollment: intermediate levels are overrepresented, while the lowest levels and the highest levels are comparatively rare. A naïvely trained classifier will therefore bias predictions toward majority classes, obscuring pedagogically important distinctions at higher proficiency levels.

3. **Resource Constraints:** Educational institutions require solutions that are reproducible, affordable, and privacy-conscious. Reliance on large proprietary APIs is often unsustainable due to cost and data protection concerns. Consequently, this thesis restricts its design space to open-source architectures (e.g., XLM-RoBERTa) that can be fine-tuned on modest hardware.

1.4 Research Problem and Research Questions

Against this background, this thesis addresses the following central research problem:

> *How can an effective and pedagogically meaningful AES system be designed for a morphologically rich language, using a small and imbalanced real-world dataset, under strict computational constraints?*

To address this problem, we must investigate three distinct dimensions of modeling strategy. First, given the morphological complexity of Russian, we must determine whether traditional linguistic feature engineering remains necessary or if modern deep learning has rendered it obsolete. Second, given the scarcity and imbalance of the data, we must explore loss functions that are more data-efficient than standard classification. Finally, we must assess whether combining explicit linguistic features with deep representations yields complementary benefits or merely redundant signals.

Therefore, this thesis investigates three specific research questions:

- **RQ1: Model Paradigm Comparison.** Do fine-tuned multilingual Transformer models outperform traditional feature-based approaches for Russian L2 Automated Essay Scoring under low-resource conditions?

- **RQ2: Ordinal versus Nominal Modeling.** Does explicitly modeling proficiency as an ordinal construct (e.g., via ordinal regression objectives) yield measurable and consistent performance gains over nominal classification and linear regression approaches?

- **RQ3: Hybridization and Complementarity.** Does combining engineered linguistic features with deep contextual embeddings (via early or late fusion architectures) provide additional predictive benefit beyond strong Transformer-based ordinal models?

1.5 Research Goals and Contributions

In addressing these questions, the goal of this thesis is not merely to maximize predictive performance on a benchmark dataset, but to determine which modeling strategies are most reliable, interpretable, and stable under realistic institutional constraints.

The main contributions are threefold:

1. Ordinal Modeling for Russian AES. We provide a systematic evaluation of ordinal regression techniques for Russian L2 AES, comparing the CORAL (Consistent Rank Logits) and CORN (Conditional Ordinal Regression) frameworks to standard cross-entropy classification and linear regression baselines.

2. Evaluation of Hybrid Fusion under Data Scarcity. We empirically test whether explicit linguistic knowledge, encoded through engineered features, complements contextualized Transformer embeddings in small-data settings.

3. Comparative Low-Resource Benchmarking. We document model behavior on a real-world, imbalanced Russian L2 dataset, providing a practical blueprint for AES development in morphologically rich, non-English contexts.

1.6 Evaluation Criteria and Notion of Success

Success in this domain cannot be defined by accuracy alone. Misclassifying an *Advanced* essay as *Novice* constitutes a more serious error than misclassifying it as *Intermediate High*. Consequently, model evaluation prioritizes ordinal-aware metrics, particularly **Quadratic Weighted Kappa (QWK)** and **Mean Absolute Error (MAE)**.

Beyond numerical performance, four criteria define success:

1. **Statistical Superiority:** Proposed models must outperform both dummy baselines and strong feature-based baselines.
2. **Architectural Insight:** The study must determine whether ordinal objectives provide a consistent advantage over nominal and linear formulations.
3. **Metric Stability:** Preferred models must demonstrate low variance across cross-validation folds.
4. **Qualitative Validity:** Model errors should be linguistically interpretable rather than arbitrary.

1.7 Structure of the Thesis

The remainder of this thesis is organized as follows:

- **Chapter 2** reviews prior work on AES, contrasting traditional feature engineering with neural approaches.
- **Chapter 3** describes the Middlebury dataset and label distribution.

- **Chapter 4** details the engineered linguistic features used in baseline models.
- **Chapter 5** defines the modeling architectures, including XLM-R and ordinal heads (CORAL/CORN).
- **Chapter 6** specifies training protocols and validation procedures.
- **Chapter 7** presents experimental results and error analysis.
- **Chapter 8** interprets the findings, discussing the "Ordinal Advantage" and the limited utility of fusion.
- **Chapter 9** concludes the thesis and outlines future directions.

Chapter 2

Related Work and Theoretical Foundations of Automated Essay Scoring

This chapter situates the present study within the broader literature on Automated Essay Scoring (AES), second language assessment, and neural natural language processing. It first traces the historical evolution of AES methodologies, highlighting the persistent tension between surface fluency, semantic modeling, and construct validity. It then reviews linguistically grounded approaches rooted in Second Language Acquisition (SLA) research, followed by contemporary Transformer-based models and ordinal learning frameworks. Finally, it examines hybrid and ensemble strategies that seek to reconcile predictive performance with interpretability. Together, these strands of research define the conceptual and methodological foundations for the modeling decisions explored in this thesis.

2.1 Automated Essay Scoring: Historical Context

Research on Automated Essay Scoring (AES) has evolved through several distinct methodological phases, reflecting broader developments in natural language processing and educational measurement. From early surface-based statistical models to contemporary deep learning architectures, each stage has sought to approximate human judgment of writing quality while grappling with fundamental questions of construct validity, interpretability, and data efficiency.

2.1.1 Early Statistical Systems and the Construct Problem

The inception of automated essay scoring is widely attributed to Ellis Page's *Project Essay Grade* (PEG) system in the 1960s [19]. PEG relied on a small set of easily computable surface features—most notably essay length, average word length, and punctuation counts—which were combined using linear regression to predict human-assigned scores. Despite its simplicity, PEG demonstrated surprisingly high correlations with human raters, a result that fueled early optimism about the feasibility of automated scoring.

However, subsequent critiques revealed a fundamental limitation of these early systems: they primarily measured *fluency* or productivity rather than true linguistic proficiency. Length-based features, in particular, were shown to dominate model predictions, raising concerns that systems could

be "gamed." For instance, a student could theoretically achieve a high score by writing a nonsensical but lengthy essay replete with complex punctuation and polysyllabic words, exposing the disconnect between what the model optimized and the pedagogical construct of writing ability. This limitation became known as the *construct validity problem* in AES.

Later systems, such as ETS's *e-rater* [2], sought to address these shortcomings by incorporating more linguistically informed features. In addition to length measures, e-rater included grammar error counts, usage patterns, and stylistic features derived from shallow parsing. While these enhancements improved interpretability, they still relied heavily on hand-crafted heuristics and struggled to capture deeper aspects of discourse organization.

2.1.2 The Semantic Turn: Latent Semantic Analysis

A major conceptual shift occurred with the introduction of Latent Semantic Analysis (LSA) for essay evaluation, exemplified by the *Intelligent Essay Assessor* (IEA) [15]. Unlike earlier approaches that focused on surface form, LSA aimed to model semantic content by representing texts in a high-dimensional vector space derived from word co-occurrence patterns.

Technically, LSA constructs a term–document matrix from a large corpus and applies Singular Value Decomposition (SVD) to reduce this matrix to a lower-dimensional latent space. In this space, similarity is computed using cosine distance. This allowed systems to evaluate essays based on their semantic proximity to high-scoring reference texts rather than grammatical correctness alone.

This represented a significant breakthrough: automated systems could finally assess whether an essay addressed the intended topic. However, LSA remained limited by its "bag-of-words" assumption. Because it ignores word order, LSA cannot distinguish between "The dog bit the cat" and "The cat bit the dog," meaning semantic adequacy could be awarded even to essays that were grammatically incoherent.

2.1.3 The Neural Turn: Sequential Representation Learning

The next major transition was driven by the deep learning revolution, particularly the introduction of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models [22]. These architectures offered a principled way to process text as a sequence, enabling the modeling of local syntactic dependencies and context.

Pioneering work by Taghipour and Ng [22] demonstrated that LSTM-based models could achieve state-of-the-art performance on benchmarks like the ASAP dataset without manual feature engineering. By learning representations directly from raw token sequences, these models implicitly captured lexical choice and syntactic variation. Compared to statistical baselines, LSTMs showed improved robustness, particularly when trained on sufficiently large datasets.

Despite these advances, early neural AES models exhibited notable limitations. Recurrent architectures process text sequentially, which makes them susceptible to vanishing gradients and memory bottlenecks when dealing with long essays. While LSTMs mitigate these issues through gating mechanisms, they still struggle to retain information over very long contexts, often prioritizing recent tokens

over earlier content. Moreover, their sequential nature limits parallelization, leading to high computational costs during training.

From an interpretability standpoint, LSTM-based models also introduced new challenges. Although they outperformed feature-based systems in predictive accuracy, they offered little transparency regarding the linguistic properties driving their decisions. This opacity raised concerns in educational settings, where automated scores are expected not only to be accurate but also to be explainable and pedagogically meaningful.

Toward Contextualized Representations. These limitations set the stage for the adoption of Transformer-based architectures in AES, which will be discussed in detail in Section 2.3. By replacing recurrence with self-attention mechanisms, Transformers enable direct modeling of long-range dependencies and global context, addressing many of the structural weaknesses of earlier neural models. At the same time, the transition from hand-crafted features to fully end-to-end learning has revived long-standing debates about construct validity and interpretability—debates that motivate the hybrid and ordinal approaches explored in this thesis.

In summary, the historical evolution of AES reflects a tension between surface fluency, semantic adequacy, and structural linguistic competence. Each methodological phase has addressed some limitations of its predecessors while introducing new challenges. Understanding this trajectory is essential for situating contemporary AES models and for motivating the integration of linguistically informed features with modern deep learning techniques.

2.2 Hand-crafted Linguistic Features in SLA and AES

While deep learning approaches have achieved dominance in benchmark performance, the theoretical foundation of automated assessment remains rooted in Second Language Acquisition (SLA) research. Unlike end-to-end neural models, which learn opaque representations, SLA-informed approaches seek to quantify proficiency through explicit, linguistically motivated constructs. Research in this domain has shifted from a deficit-oriented focus on error analysis to a complexity-oriented view, emphasizing the sophistication of the learner's interlanguage. This section reviews the three primary dimensions of linguistic complexity—syntactic, lexical, and phraseological—that serve as the basis for the feature engineering methodology adopted in this thesis.

2.2.1 Syntactic Complexity and Developmental Profiling

Syntactic complexity—the range and sophistication of grammatical structures produced by a learner—is perhaps the most robust indicator of L2 writing development. In the context of Russian, a morphologically rich language with flexible word order, syntactic metrics have proven particularly discriminative.

Recent work by Kisselev, Klimov, and Kopotev [11] has been instrumental in establishing dependency-based metrics as valid indices of proficiency for Russian. Their research highlights that traditional surface measures, such as sentence length, often conflate distinct linguistic phenomena. Instead, they propose metrics derived from dependency parsing, such as Mean Dependency Distance (MDD) and

Average Tree Depth, which capture the hierarchical complexity of sentences independent of linear length. Crucially, their work distinguishes between *Heritage Learners* (HL) and *Second Language* (L2) learners, noting that while HLs may exhibit native-like intuition for certain structures, L2 learners—the focus of this study—tend to show a more linear progression in syntactic elaboration, moving from simple coordinate structures to complex subordination as formal instruction advances [11].

Further investigating developmental trajectories, Kisselev et al. [12] provided a "developmental profile" of learner Russian, demonstrating that different syntactic strategies emerge at different proficiency stages. For instance, novice and intermediate learners often expand sentence length by adding coordinate phrases (e.g., "and" conjunctions), whereas advanced learners achieve length through subclausal elaboration and embedded dependent clauses. This finding directly motivates the inclusion of specific syntactic features in our model, such as the *clauses-per-sentence* ratio and *average tree depth*, which serve as proxies for this shift from horizontal (coordinate) to vertical (subordinate) complexity.

Despite their strong theoretical grounding, dependency-based metrics are not without limitations. Automatic parsers are typically trained on native-language corpora and may introduce noise when applied to learner data, particularly in morphologically ambiguous contexts such as Russian case marking and agreement. Nevertheless, prior work suggests that aggregate syntactic measures remain robust at the population level, even in the presence of sentence-level parsing errors. As a result, dependency-derived indices provide a practical compromise between linguistic validity and computational feasibility, making them suitable for large-scale proficiency modeling despite imperfect annotations.

2.2.2 Lexical Diversity and the Length Problem

Lexical proficiency is multidimensional, encompassing both the breadth of vocabulary known (diversity) and the rarity of the words used (sophistication). Historically, AES systems relied on the Type-Token Ratio (TTR) to measure diversity. However, TTR is notoriously sensitive to text length: as a text grows longer, the probability of repeating function words increases mechanically, artificially lowering the TTR score of proficient (and thus prolific) writers.

To address this methodological artifact, modern SLA research advocates for length-invariant metrics. The Moving-Average Type-Token Ratio (MATTR), utilized in this study, mitigates the length bias by computing TTR within a fixed sliding window (e.g., 50 tokens) and averaging the results [6]. This approach allows for a fair comparison of lexical density between short essays produced by novices and longer essays by advanced learners. Additionally, indices such as Guiraud's Index ($Types/\sqrt{Tokens}$) have been employed to normalize for volume, ensuring that models measure the density of information rather than the sheer quantity of text production.

From a psycholinguistic perspective, lexical diversity reflects not only vocabulary size but also the efficiency of lexical access during production. Advanced learners tend to exhibit greater lexical availability, allowing them to avoid excessive repetition and rely less on high-frequency scaffolding words. Consequently, diversity-based measures capture the interaction between lexical knowledge and fluency, making them indirect but informative indicators of proficiency rather than purely stylistic properties of texts.

2.2.3 Phraseology and Collocational Competence

A critical yet often overlooked dimension of L2 proficiency is phraseological competence—the ability to use multi-word expressions and collocations naturally. As noted by Sinclair's "idiom principle," native speakers do not construct sentences word-by-word but rather retrieve holistic chunks of language. For L2 learners, acquiring these prefabricated patterns is a late-stage developmental milestone.

Kopotev et al. [14] conducted a corpus-driven analysis of collocational complexity in L2 Russian, revealing that advanced proficiency is characterized not just by using *correct* collocations, but by using *statistically strong* ones. They introduced metrics based on association measures (such as *t*-score and MI-score) to quantify the strength of bigrams used by learners. Their findings suggest that while intermediate learners rely heavily on high-frequency, generic combinations (e.g., *big house*), near-native learners successfully employ strong, specific collocations (e.g., *vnesenny vklad* — "made contribution") that mirror native usage patterns.

Phraseological features are particularly valuable for distinguishing higher proficiency levels, where grammatical accuracy often reaches a ceiling and traditional error-based metrics lose discriminative power. In this sense, collocational strength functions as a late-stage proficiency marker, complementing syntactic and lexical indices that are more sensitive at lower levels. This research underscores that standard vocabulary lists are insufficient for capturing the upper bounds of proficiency: a learner may use relatively simple lexical items yet combine them in highly idiomatic ways that signal advanced control. Consequently, integrating features that approximate collocational sensitivity is essential for an automated system to distinguish between grammatically correct intermediate text and truly advanced, idiomatic writing.

2.3 Transformer-based Approaches to Proficiency Prediction

The landscape of Automated Essay Scoring (AES) has been substantially reshaped by the advent of large pre-trained language models (PLMs) based on the Transformer architecture. Unlike preceding recurrent neural networks (RNNs) or statistical models, Transformers leverage self-attention mechanisms to process input sequences in parallel, capturing long-range dependencies and bidirectional context with unprecedented efficacy. This section reviews the transition to Transformer-based scoring, focusing on the pre-training paradigms that enable their success, the architectural adaptations required for document-level assessment, and the methodological trade-offs they introduce in educational contexts.

2.3.1 The Pre-training and Fine-tuning Paradigm

The core advantage of models such as BERT (Bidirectional Encoder Representations from Transformers) lies in their two-stage training process: pre-training on massive unlabelled corpora followed by fine-tuning on downstream tasks. During pre-training, BERT employs two objectives: Next Sentence Prediction (NSP), and Masked Language Modeling (MLM). The latter is an objective in which random tokens are masked and the model must predict them based on their surrounding context. This

objective enables the model to acquire rich, general-purpose linguistic representations encoding lexical semantics, syntactic structure, and aspects of discourse coherence before being exposed to any task-specific supervision.

For AES, this paradigm is particularly transformative due to the chronic scarcity of labeled data. Whereas earlier neural approaches required large annotated corpora to learn basic linguistic regularities from scratch, fine-tuning a pre-trained Transformer allows models to achieve strong performance with comparatively few scored essays. The model effectively transfers its prior knowledge of language structure to the task of proficiency prediction. In multilingual settings, this effect is further amplified by models such as XLM-RoBERTa [5], which are pre-trained on text from over 100 languages. Such models enable cross-lingual transfer, allowing systems trained on high-resource languages to be adapted to lower-resource contexts, including learner Russian, with minimal architectural modification.

Despite these advantages, fine-tuning-based AES models remain sensitive to domain and prompt variation. Because pre-training corpora differ substantially from learner essays in genre, register, and communicative intent, Transformer-based scorers may inadvertently learn prompt-specific or topical cues rather than generalizable indicators of proficiency. This sensitivity has motivated research into prompt-aware modeling and cross-prompt evaluation protocols, highlighting that strong aggregate performance does not necessarily guarantee robustness across assessment contexts.

2.3.2 Multi-Scale Essay Representation

A central challenge in applying standard Transformer architectures to AES lies in the mismatch between their typical training units and the document-level nature of essays. Vanilla BERT models impose a maximum sequence length of 512 tokens, which may truncate longer submissions. Moreover, representing an entire essay as a single flat sequence risks obscuring the hierarchical structure of writing, in which coherence and organization operate at multiple levels.

To address these limitations, recent work has explored hierarchical and multi-scale modeling strategies. Wang et al. [25] proposed a framework for the joint learning of multi-scale essay representations that moves beyond the conventional reliance on the special `[CLS]` token as a global summary. Their architecture simultaneously models linguistic information at multiple granularities:

- **Token-scale:** capturing local lexical choice and grammatical correctness;
- **Segment-scale:** aggregating information across coherent spans of text to model discourse flow;
- **Document-scale:** synthesizing a global semantic representation of the essay.

By integrating these complementary views, the model can detect localized weaknesses—such as inconsistencies or breakdowns in argumentation—that may be masked by a single pooled embedding. Empirical results on the ASAP dataset demonstrate that multi-scale representations outperform standard BERT fine-tuning, particularly for prompts that require sustained reasoning or structured argumentation. This line of research highlights that while pre-trained embeddings provide powerful linguistic priors, their effective application to essay scoring necessitates architectural adaptations that respect the hierarchical organization of extended discourse.

2.3.3 Ordinality, Loss Functions, and Score Modeling

An additional methodological challenge in Transformer-based AES concerns the nature of the target variable itself. Essay scores are inherently *ordinal*: adjacent proficiency levels are ordered but not necessarily equidistant. Nevertheless, many Transformer-based systems formulate AES as a standard regression problem, optimizing mean squared error over scalar predictions. While this approach is computationally convenient and often yields strong evaluation scores, it implicitly assumes linear spacing between proficiency levels and ignores the ordered structure of the scale.

This mismatch between model objective and assessment construct has important implications. Treating scores as continuous may obscure nonlinear developmental progressions between adjacent proficiency bands and fails to encode the intuition that misclassifying an essay by one level is less severe than misclassifying it by several. Recent work has therefore explored ordinal regression objectives and ranking-based losses for AES, though such approaches remain less common than standard regression. The tension between representational power and construct-aware modeling provides further motivation for integrating neural representations with explicitly structured, linguistically grounded features.

2.3.4 The Interpretability Trade-off

Despite their predictive superiority, Transformer-based AES systems introduce a pronounced interpretability trade-off. Although they consistently achieve higher Quadratic Weighted Kappa (QWK) scores than feature-based or recurrent models, they operate as high-dimensional black boxes. A fine-tuned Transformer typically maps an essay to a dense embedding vector—often hundreds of dimensions—which is then projected to a scalar score with little transparency regarding the linguistic evidence underlying the prediction.

This opacity poses a significant obstacle in educational contexts, where assessment systems are expected not only to rank learners but also to provide diagnostically meaningful feedback. Unlike feature-based approaches, which can directly attribute a low score to specific factors such as limited lexical diversity or high error rates, Transformer-based models offer no straightforward mechanism for explaining their decisions. As a result, their pedagogical validity remains contested despite strong empirical performance.

These limitations motivate the hybrid modeling strategy adopted in this thesis. By combining the representational strength of Transformer-based encoders with explicit, interpretable linguistic features grounded in SLA research (see Section 2.2), we aim to reconcile predictive accuracy with transparency. Such an approach seeks to leverage the semantic and contextual sensitivity of modern neural models while retaining the explanatory power required for educational assessment.

2.4 Ordinal Classification Methods in NLP

A fundamental methodological challenge in Automated Essay Scoring (AES) concerns the representation of the target variable. Proficiency labels—such as those defined by the ACTFL or CEFR frame-

works—are inherently *ordinal*, reflecting a ranked progression of language ability (e.g., Novice $<$ Intermediate $<$ Advanced). Despite this, most machine learning approaches model these labels either as nominal categories, using standard multi-class classification, or as continuous values, using regression. Both formulations impose mathematical assumptions that are often misaligned with the psychometric and developmental nature of language proficiency. This section reviews the limitations of these conventional approaches and motivates the adoption of ordinal regression methods in neural NLP models.

2.4.1 Limitations of Nominal and Metric Formulations

In standard multi-class classification, proficiency levels are treated as mutually independent categories and optimized using the Cross-Entropy loss. This formulation ignores the ordered structure of the label space: all misclassifications are penalized equally, regardless of their severity. For instance, predicting "Novice High" for a "Superior" essay incurs the same loss as predicting "Advanced High." From a pedagogical and assessment perspective, however, such errors are not equally serious. Formally, this setup induces a Hamming-distance-like loss landscape, which fails to encourage "near-miss" predictions that preserve ordinal proximity.

An alternative formulation casts AES as a regression problem, typically optimized using Mean Squared Error (MSE). While this approach accounts for ordering, it assumes that proficiency labels lie on a metric scale with equal intervals. In other words, it presupposes that the developmental distance between adjacent levels is constant across the scale. This assumption is rarely justified in second language acquisition (SLA): empirical evidence and assessment practice suggest that progression slows at higher proficiency levels, making transitions such as *Advanced* $\rightarrow$ *Superior* substantially more demanding than those at lower levels. As a consequence, regression-based models often exhibit "regression toward the mean," systematically underestimating high proficiency essays and overestimating low ones.

2.4.2 Consistent Rank Logits (CORAL)

To address these limitations, Cao et al. [3] proposed the Consistent Rank Logits (CORAL) framework, which explicitly models the ordinal structure of the label space. CORAL reformulates an ordinal classification problem with K ordered levels as a set of $K-1$ binary classification tasks, each corresponding to a cumulative comparison.

Rather than predicting a single class label, the model answers a sequence of questions of the form: "Is the essay's proficiency level greater than k?" for $k = 0, \ldots, K-2$. The model learns a shared feature representation and a single weight vector, while maintaining distinct bias terms for each binary classifier. This weight-sharing mechanism enforces a strict *monotonicity constraint*: the predicted probability of exceeding a higher proficiency threshold cannot exceed the probability of exceeding a lower one.

This constraint guarantees *rank consistency*, ensuring that the model's outputs respect the ordinal ordering by construction. As a result, CORAL aligns the optimization objective with the developmen-

tal logic underlying proficiency scales, making it particularly suitable for educational and assessment-oriented NLP tasks.

2.4.3 Conditional Ordinal Regression (CORN)

An alternative ordinal approach is Conditional Ordinal Regression for Neural Networks (CORN), introduced by Shi et al. [21]. Instead of modeling cumulative probabilities, CORN decomposes the ordinal prediction problem using the chain rule of probability. Specifically, the model estimates conditional probabilities of the form

$$P(y > k \mid y \geq k),$$

corresponding to the probability that an instance surpasses level k, given that it has already reached at least level k.

This formulation converts the ordinal task into a sequence of conditional binary decisions, each focused on distinguishing between two adjacent levels. During inference, absolute class probabilities are recovered by multiplying the relevant conditional probabilities along the chain. One practical advantage of CORN is its robustness to class imbalance, a pervasive issue in learner corpora where intermediate proficiency levels are typically overrepresented. By isolating decision boundaries between adjacent classes, CORN prevents dominant classes from overwhelming the learning signal and allows finer-grained modeling of underrepresented proficiency transitions.

2.4.4 Relevance to the ACTFL Proficiency Scale

The use of ordinal regression methods is particularly well motivated by the structure of the ACTFL proficiency scale adopted in this study. The scale consists of both major levels (Novice, Intermediate, Advanced) and finer-grained sublevels (Low, Mid, High). While distinctions between adjacent sublevels often rely on subtle quantitative cues—such as frequency of grammatical control or lexical diversity—transitions between major levels reflect qualitative shifts in communicative competence, such as the move from sentence-level to paragraph-level discourse.

Ordinal frameworks such as CORAL and CORN are well suited to capture this non-linearity. By avoiding the equal-distance assumptions of regression and the independence assumptions of nominal classification, these models allow the learning process to reflect the variable "difficulty" associated with different proficiency transitions. This alignment between the loss function and the assessment construct constitutes a central methodological motivation for their adoption in the present work.

2.5 Feature Fusion and Model Ensembling

While deep learning and feature engineering have historically been viewed as competing paradigms in Automated Essay Scoring (AES), contemporary research increasingly treats them as complementary rather than mutually exclusive. Empirical and theoretical work suggests that neural models and hand-crafted linguistic features capture fundamentally different aspects of language proficiency.

Transformer-based architectures have been shown to be particularly effective at modeling semantic coherence, topical relevance, and global discourse patterns, whereas explicit linguistic features tend to provide high-resolution signals related to morphosyntactic accuracy, lexical diversity, and mechanical correctness. This section reviews the primary methodological strategies for integrating these heterogeneous sources of information: feature fusion (early fusion) and model ensembling (late fusion).

2.5.1 Feature Fusion Strategies

Feature fusion, often referred to as *early fusion*, involves integrating engineered linguistic features directly into the neural architecture prior to the final prediction stage. The most common implementation relies on vector concatenation. In this setup, an essay is passed through a Transformer encoder (e.g., BERT or XLM-RoBERTa) to obtain a dense representation $h_{\text{CLS}} \in \mathbb{R}^d$, typically derived from the special `[CLS]` token. In parallel, a vector of engineered linguistic features $f \in \mathbb{R}^k$ is computed using external NLP tools such as parsers, taggers, and lexical resources.

The two representations are concatenated to form a composite vector $v = [h_{\text{CLS}}; f]$, which is then provided as input to the final prediction layer. Prior to concatenation, engineered features are typically standardized to ensure numerical compatibility with neural embeddings. Uto et al. [23] demonstrated the effectiveness of this approach, showing that augmenting neural representations with rubric-based linguistic metrics yields significant improvements in trait-specific scoring tasks compared to models relying on either representation alone.

Early fusion allows the decision layer to condition its predictions jointly on latent semantic representations and explicit linguistic signals. This enables the model to learn interaction effects that are difficult to capture with either paradigm in isolation—for example, distinguishing an essay that is topically relevant but grammatically unstable from one that is linguistically well-formed but only loosely aligned with the prompt.

2.5.2 Ensembling and Late Fusion

An alternative integration strategy is *late fusion*, commonly implemented through model ensembling. In this paradigm, independent models are trained on different representations of the same input, and their predictions are combined only at inference time. A typical configuration pairs a fine-tuned Transformer model with a feature-based model (e.g., linear regression or gradient-boosted trees trained on linguistic metrics). Each model produces a probability distribution over proficiency levels, which are then aggregated via weighted averaging or a meta-learner (stacking).

The motivation for ensembling lies in the principle of error diversity. Neural models often exhibit high variance and may overfit to spurious lexical or topical cues in limited-data regimes, a phenomenon commonly described as *Clever Hans* behavior [16]. In contrast, feature-based models are typically high-bias but low-variance: their predictions are constrained by linguistically interpretable signals and are therefore less susceptible to superficial correlations.

Research in educational data mining indicates that ensembling is particularly effective in low-resource and domain-shift scenarios. When a neural model encounters out-of-distribution inputs (e.g.,

unseen prompts or genres), its confidence may remain artifactually high. Feature-based models, relying on relatively stable measures such as lexical richness or syntactic complexity, provide a stabilizing signal at the decision level. In practice, this complementary behavior often yields more robust and generalizable performance than either model alone.

Importantly, both early- and late-fusion strategies are compatible with ordinal learning objectives such as CORAL and CORN, as they operate at the representation or decision level without imposing metric assumptions on the label space. This makes them particularly suitable for proficiency scales that are inherently ordered but non-linear.

2.6 Summary

This chapter has traced the evolution of Automated Essay Scoring from surface-level heuristics to deep semantic modeling. While Transformer-based models offer state-of-the-art predictive performance, they lack the interpretability and fine-grained linguistic sensitivity required for many pedagogical and assessment-oriented applications. Conversely, research in second language acquisition has identified robust feature sets—rooted in syntactic complexity, lexical sophistication, and error patterns—that closely align with human proficiency judgments.

Finally, we reviewed ordinal regression frameworks that respect the non-linear structure of proficiency scales and integration strategies that combine neural and feature-based representations. Together, these insights define the methodological blueprint for the present study. The subsequent chapters detail the dataset used, the extraction of the linguistically motivated feature set described in Section 2.2, and the implementation of hybrid, ordinal-aware models designed to address the limitations of existing AES systems.

Chapter 3

Dataset

This chapter describes the learner corpus used in this study, its annotation scheme, and the preprocessing steps required to construct a machine-learning-ready dataset under low-resource conditions.

3.1 Corpus description (Russian L2 learner essays)

The dataset used in this study consists of written essays produced by learners of Russian as a second language (L2) enrolled in intensive language programs at the Middlebury Language Schools, Middlebury College (Vermont, USA). Middlebury Language Schools are well known for their immersion-based pedagogy and long-standing expertise in foreign language assessment, making their learner data particularly valuable for research in second language acquisition and proficiency assessment.

The corpus comprises entrance and final written examination essays collected over four academic years: 2017, 2018, 2019, and 2020. These essays were produced in response to predefined prompts administered as part of the school's internal placement and exit assessment procedures. For each academic year, the number of essays per prompt and per assessment type (entrance vs. final) is relatively limited, typically on the order of up to approximately ten essays per year per prompt, reflecting the selective and small-cohort nature of intensive language programs.

The original data were provided in the form of transcribed plain-text (.txt) files, organized hierarchically by year, assessment type (entrance or final), and prompt. An example directory structure is shown below for illustration:

```
./data/preprocessed/Transcribed ORIGINAL data txt/
  Summer 2017/
  Summer 2018/
  Summer 2019/
  Summer 2020/
    Original_Writing 2020_online/
      Pre_Test/
        Pre_Test_Prompt2/
```

Each text file corresponds to a single learner essay. Some learners contributed both an entrance and a final essay, while others contributed only one. Not all enrolled students are represented in the

corpus: according to the accompanying documentation, some learners either did not consent to the use of their written production for research purposes or their essays were not transcribed by the language school staff. As a result, the set of learners for whom written texts are available constitutes a subset of the full cohort.

In addition to the essay texts, the dataset includes tabular assessment records provided in spreadsheet format. These records contain extensive information about each learner, including personal identifiers, placement decisions, oral proficiency interview (OPI) results, grammar test scores, and writing assessment outcomes for both entrance and exit stages. For the purposes of this study, only a minimal subset of these fields was used: learner names (to align essays with assessment records) and the writing proficiency scores assigned at entrance and/or exit.

Some inconsistencies were observed in the raw assessment files. In particular, learner names were not always formatted consistently across spreadsheets and text filenames, and writing proficiency scores were expressed using multiple notational systems, including both numeric internal scales (e.g., 1.5, 1.9, 2.0, 1.0+) and categorical ACTFL-style labels (e.g., IL, IL+, IM). These discrepancies required normalization and careful matching during dataset construction but did not affect the underlying validity of the annotations.

All learners were L2 Russian learners with diverse first-language backgrounds. From an NLP perspective, this diversity introduces systematic properties that distinguish the data from native-speaker corpora, such as learner-specific error patterns, non-native lexical choices, and syntactic transfer effects. These characteristics are intrinsic to the task of automated proficiency assessment and are expected to influence both feature-based and neural modeling approaches examined later in this thesis.

3.2 ACTFL annotation scheme and ordinal label distribution

The writing proficiency labels used in this study are grounded in the ACTFL Proficiency Guidelines, developed by the American Council on the Teaching of Foreign Languages (ACTFL). The ACTFL scale is a widely adopted standard in language education and assessment in the United States and is extensively used in academic, governmental, and institutional contexts. It defines proficiency as an ordered set of communicative ability levels, rather than as a continuous numerical score.

The ACTFL framework organizes proficiency into the main levels Novice, Intermediate, Advanced, and Superior, with each level (except Superior) subdivided into Low, Mid, and High sublevels. Importantly, these levels are ordinal in nature: they reflect a ranking of communicative competence but do not assume equal distances between adjacent categories.

Although the ACTFL scale is more commonly used in the U.S., while the CEFR (Common European Framework of Reference for Languages) is more prevalent in Europe, the two frameworks were explicitly designed to be interoperable. Consequently, the modeling approaches explored in this thesis are not specific to a single proficiency scale and can be transferred to alternative ordinal annotation schemes without methodological changes.

At Middlebury Language Schools, assessors internally employ an internal numerical scale (0.0 – 3.0) that corresponds directly to ACTFL proficiency levels. Following direct clarification with the

Middlebury staff, it was confirmed that these numeric labels do not encode metric distances and should not be interpreted as real-valued scores. Instead, they function as abstract ordinal identifiers aligned with ACTFL categories.

Table 3.1 summarizes the mapping between ACTFL writing proficiency levels, Middlebury's internal labels, and the normalized ordinal labels used in this study.

ACTFL level	Middlebury label	Label used in this study
Novice Low	0.0	0
Novice Mid	0.5	1
Novice High	0.9	2
Intermediate Low	1.0	3
Intermediate Mid	1.5	4
Intermediate High	1.9	5
Advanced Low	2.0	6
Advanced Mid	2.5	7
Advanced High	2.9	8
Superior	3.0	9

Table 3.1: Mapping of writing proficiency labels

In theory, this mapping allows for ten distinct proficiency classes, ranging from Novice Low to Superior. In practice, however, the corpus only contains essays from Novice Mid to Advanced High, resulting in a total of eight observable classes. No texts were available for Novice Low or Superior learners.

This restriction reflects realistic conditions of language assessment practice. Complete beginners rarely produce extended written texts suitable for essay-based evaluation, while learners at the Superior level are unlikely to participate in institutional placement exams focused on instructional grouping. From an applied perspective, the dataset therefore captures the proficiency range that is most relevant for automated essay assessment systems deployed in educational settings.

From a machine learning standpoint, the absence of extreme proficiency levels reduces the availability of anchoring points at both ends of the scale. The implications of this limitation, as well as techniques for modeling ordinal structure under such conditions, are discussed in detail in Chapter 5.

3.3 Dataset characteristics (size, skewness, imbalance)

For the purposes of modeling, entrance and final essays were treated as equally valid observations. Although they were produced at different stages of instruction, both were assessed using the same criteria and annotation procedures by trained human raters. Consequently, no distinction was made between entrance and exit texts during dataset construction.

After preprocessing and label normalization, the final dataset consists of 1,138 essays distributed across eight ordinal proficiency levels. Table 3.2 presents the label distribution used in this study, where labels 0–7 correspond to Novice Mid through Advanced High respectively.

Model Label	ACTFL Level	Count	Percentage
0	Novice Mid	36	3.1%
1	Novice High	62	5.4%
2	Intermediate Low	65	5.7%
3	Intermediate Mid	349	30.7%
4	Intermediate High	255	22.4%
5	Advanced Low	214	18.8%
6	Advanced Mid	138	12.1%
7	Advanced High	19	1.7%
Total		**1,138**	**100%**

Table 3.2: Distribution of essays by proficiency level

The distribution is highly imbalanced and skewed, with the majority of essays concentrated in the Intermediate Mid (Class 3) and Intermediate High (Class 4) levels, which together constitute over 53% of the corpus. In contrast, the lowest and highest observed proficiency levels (Novice Mid and Advanced High) are severely underrepresented – the "Advanced High" class contains only 19 samples ($< 2\%$), and Novice Mid contains 36 samples (3.1%).

This imbalance reflects the instructional structure of intensive language programs, where most learners cluster around intermediate proficiency levels and relatively few reach advanced-high competence within the program duration. While this distribution poses challenges for standard multi-class classification approaches, it also mirrors realistic deployment scenarios for automated assessment systems.

Importantly, the labels exhibit a clear ordinal structure, with neighboring classes representing incremental changes in communicative ability rather than categorical jumps. Ignoring this ordering would discard valuable information about proficiency progression. Consequently, the dataset is well suited for investigating ordinal-aware modeling approaches, as opposed to purely nominal classification methods.

The combined effects of class imbalance, skewness, and ordinal dependency play a central role in model selection and evaluation. These aspects are explicitly addressed in later chapters through appropriate loss functions, evaluation metrics, and modeling strategies tailored to ordinal prediction tasks.

3.4 Data preprocessing

The raw data provided by Middlebury Language Schools required substantial preprocessing in order to produce a clean, consistent, and ethically compliant dataset suitable for machine learning experiments. This preprocessing pipeline was implemented in Python and executed as a reproducible notebook-based workflow. The main preprocessing stages included (1) consolidation of assessment records across years, (2) anonymization of learner identities, (3) normalization of heterogeneous proficiency score formats into a unified ordinal scale, (4) alignment of essay texts with corresponding scores, and (5) construction of the final training tables.

3.4.1 Consolidation of assessment records

Assessment metadata were originally provided as separate tabular files for each academic year (2017–2020). These files differed slightly in column naming conventions, reflecting changes in internal data collection practices over time. For example, the writing assessment columns appeared under different headers such as Writing, Writing Entrance, or Writing test Final.

To address this, all yearly files were loaded and mapped to a common schema consisting of the following fields:

- learner last name
- learner first name
- entrance writing score
- exit writing score
- academic year

The standardized yearly tables were then concatenated into a single dataframe. This step produced a unified table of raw writing scores covering all four years while preserving year-level metadata for later analysis.

3.4.2 Anonymization and learner identifiers

To ensure compliance with ethical research standards and institutional data protection requirements, all personal identifiers were removed at an early stage of preprocessing.

Each learner was assigned a unique anonymized identifier of the form `STUDENT_XXXX`, where the numeric suffix is an arbitrary sequential index. Internally, this identifier was derived by constructing a temporary key from the learner's last and first name (uppercased and normalized), which was then mapped to a stable anonymized ID. The resulting lookup table was stored separately and excluded from all modeling steps.

After anonymization, all subsequent operations relied exclusively on the anonymized `Student_ID`. No learner names or other identifying information were retained in the modeling dataset.

3.4.3 Normalization of proficiency scores

One of the most critical preprocessing steps involved standardizing writing proficiency scores, which were expressed in multiple heterogeneous formats in the original data. These included:

- ACTFL-style categorical labels (e.g., IM, IM+, IL-),
- numeric internal Middlebury scores (e.g., 1.5, 1.9, 2.0),
- locale-dependent numeric formats using commas instead of decimal points (e.g., 1,5),

- special administrative labels (e.g., NO.TEST, NOSHOW, NT).

All valid proficiency annotations were mapped onto a single unified ordinal scale, reflecting ACTFL sublevels while preserving their ordinal – but not metric – nature. Importantly, no assumptions were made about equal numeric or semantic distances between adjacent levels.

The normalization procedure implemented a comprehensive mapping that covered:

- ACTFL base levels and modifiers (-, base, +),
- numeric equivalents using both dot and comma decimal separators,
- equivalent symbolic variants used inconsistently across years.

An excerpt of the resulting mapping illustrates the approach:

- NM-, NM, NM+ $\rightarrow$ ordinal level 1
- IM-, IM, IM+ $\rightarrow$ ordinal level 4
- AH-, AH, AH+ $\rightarrow$ ordinal level 8
- S $\rightarrow$ ordinal level 9

Administrative or missing-test labels were mapped to missing values and excluded from downstream modeling. This procedure yielded two normalized ordinal columns per learner where available: one for the entrance essay and one for the exit essay.

3.4.4 Restructuring assessment records

The normalized assessment table originally stored entrance and exit scores in a wide format, with one row per learner-year. For modeling purposes, this structure was transformed into a long format, where each row corresponds to a single assessed essay instance, characterized by:

- academic year,
- anonymized learner ID,
- test type (Entrance or Exit),
- ordinal proficiency label.

Rows with missing scores (e.g., exams not taken or not scored) were removed. This step ensured that each retained instance represented a valid human-rated essay suitable for supervised learning.

3.4.5 Essay file parsing and alignment

The essay texts themselves were stored as individual plain-text files organized in a hierarchical directory structure by year, assessment type, and prompt. Filenames followed a semi-structured convention encoding the learner's name and test type (e.g., Entry or Exit).

To align essays with assessment records, file paths were programmatically parsed to extract:

- academic year,
- test type (entrance or exit),
- learner name components.

These extracted identifiers were normalized and matched against the anonymization lookup table to recover the corresponding `Student_ID`. Only essays for which a valid anonymized learner ID and a matching normalized score were available were retained. Essays that could not be reliably matched to an assessment record were discarded.

Essay texts were read using UTF-8 encoding with graceful fallback handling to accommodate legacy encodings occasionally present in archival data. No manual correction of learner errors was performed; all spelling, grammatical, and lexical deviations were preserved exactly as produced by the learners.

3.4.6 Final dataset construction

The final output of the preprocessing pipeline is a dataset where each row corresponds to a single essay paired with a single ordinal proficiency label. The dataset contains two primary fields used in modeling:

1. the raw essay text,
2. the corresponding ordinal proficiency score.

An illustrative excerpt of the final structure is shown below (text truncated for readability):

Essay (excerpt)	**Score**
Сообщение. Об этом слов мы думаем каждый день. . .	5
У меня не есть словай. . .	2
Честность – важное качество в жизни. . .	7

Table 3.3: Essay excerpt sample

The inclusion of short text excerpts in this thesis serves illustrative purposes only. No learner-identifiable information is disclosed, and full texts are not reproduced verbatim outside the secure research environment.

This preprocessing pipeline ensures that the resulting dataset is:

- ethically anonymized,

- internally consistent across years,
- aligned with an explicit ordinal interpretation of proficiency,
- suitable for both feature-based and neural modeling approaches.

3.5 Train–validation–test split

A careful data splitting strategy is particularly important in the present study due to the limited sample size ($N = 1,138$) and pronounced class imbalance in the dataset. While a conventional approach in supervised learning is to partition the data into three disjoint subsets (training, validation, and test), such a strategy was deemed inappropriate for this dataset.

As discussed in Section 3.3, the highest proficiency levels are represented by only a small number of essays. In particular, the Advanced High (Class 7) category contains just 19 instances. Applying a standard fixed test split would result in only 3–4 essays per split for this class. If these specific essays happened to be outliers, the evaluation would be misleading.

To mitigate these issues, we adopted a **Stratified K-Fold Cross-Validation** strategy ($K = 5$ for all the evaluated models). In this setup:

- The dataset is partitioned into K disjoint subsets (folds).
- Stratification ensures that the class distribution in every fold mirrors the overall population (e.g., every fold contains proportional representatives of the rare Class 7).
- The model is trained and evaluated K times; in each iteration, $K-1$ folds serve as the training set and the remaining fold serves as the evaluation set.

This choice guarantees that every single essay in the corpus is used for evaluation exactly once, providing a comprehensive assessment of model performance without the variance introduced by random sub-sampling. We report the aggregated metrics (mean and standard deviation) across all folds. While we adopt a stratified K-Fold Cross-Validation strategy ($K = 5$) in this study, we note that some learners contributed both entrance and exit essays. Consequently, random partitioning may allow texts from the same learner to appear in both training and validation folds, potentially introducing a mild data leakage effect. A more conservative approach would be leave-one-author-out cross-validation, ensuring that all essays from the same learner are contained within a single fold. Such a strategy could be considered in future work to obtain stricter generalization estimates.

It is important to emphasize that the primary goal of this thesis is not to obtain the highest possible absolute performance scores. Rather, the objective is to perform a comparative analysis of architectural approaches (e.g., comparing Ordinal Regression against Standard Classification, and Feature Fusion against Text-only models) under low-resource constraints. By keeping the dataset, label distribution, and (where applicable) hyperparameters comparable across experiments, we focus on identifying which class of approach is more suitable for automated proficiency assessment in skewed, small-scale learner corpora.

Within this framing, the adopted splitting strategy provides a reasonable trade-off between statistical stability and methodological rigor. While the reported results may be mildly optimistic compared to a large-scale fully held-out evaluation, they remain representative of the relative strengths and weaknesses of the examined pipelines in practical low-data assessment scenarios.

Chapter 4

Engineered Linguistic Features

4.1 Motivation: complementing transformers with interpretable linguistic metrics

Automated Essay Scoring (AES) and, more broadly, automated language level assessment have historically developed along two largely complementary methodological lines: feature-based statistical models grounded in explicit linguistic theory, and representation-learning approaches based on neural networks. As discussed in the related work chapters, the former constitute not only the earliest systematic attempts at automatic assessment—beyond trivial baselines such as length-only or random classifiers—but also a long-standing standard in both research and production systems.

Early and influential AES systems, such as e-rater and IntelliMetric [1, 9], relied almost exclusively on manually engineered linguistic features capturing surface, lexical, syntactic, and discourse-level properties of learner texts. Despite their relative simplicity, such systems demonstrated that carefully selected feature sets could achieve strong correlations with human judgments, particularly when aligned with established assessment rubrics. Importantly, many legacy systems, and even several modern commercial or institutional tools, continue to rely—at least partially—on feature-based classifiers, attesting to their robustness, transparency, and practical effectiveness.

The renewed interest in feature-based methods within contemporary research is not accidental. Transformer-based models learn dense, high-dimensional representations of language that are powerful but largely opaque: while they implicitly encode information about grammaticality, lexical richness, coherence, and discourse structure, these properties are not directly observable or controllable. In contrast, engineered linguistic features are explicitly defined, linguistically motivated, and interpretable. When used in linear models, ordinal regression, or other relatively simple classifiers, they allow researchers and practitioners to inspect which aspects of language use are being prioritized by the system and how they contribute to the final prediction.

This interpretability is not merely a methodological convenience; it closely mirrors how human assessors conceptualize writing proficiency. Widely used assessment rubrics, such as those employed in IELTS or in university-level foreign language programs (e.g., the Middlebury College writing rubrics for French), explicitly reference dimensions such as grammatical accuracy, range and complexity of syntactic structures, lexical sophistication, cohesion, and overall clarity of organization. These criteria

map naturally onto classical linguistic metrics: sentence length and variability relate to fluency and control, lexical diversity measures approximate vocabulary range, readability indices reflect processing difficulty, and syntactic depth or clause counts serve as proxies for grammatical complexity.

From this perspective, feature-based modeling can be seen as an operationalization of assessment rubrics rather than an abstract optimization problem. For example, IELTS descriptors for higher bands emphasize a "wide range of structures," "flexible and accurate use" of grammar, and "natural and sophisticated control of lexical features." Similarly, Middlebury's rubrics explicitly distinguish levels of grammatical and lexical complexity, fluency, and organizational clarity. Engineered linguistic features provide a principled way to approximate these qualitative descriptors with quantitative signals, enabling models that are not only predictive but also conceptually aligned with human evaluation practices.

Another practical motivation for including engineered features is methodological complementarity. While transformers excel at capturing contextual semantics and long-range dependencies, they do not obviate the value of explicit linguistic indicators. In fact, when the feature space is limited to a few dozen well-chosen metrics, there is often little justification for applying deep architectures on top of them: linear models, generalized linear models, or ordinal classifiers are frequently sufficient and more stable, especially in low-resource or multilingual settings. In such cases, feature-based models serve both as strong baselines and as diagnostic tools for understanding model behavior.

Recent work has explored more advanced, vector-based features derived from sentence or essay embeddings, such as measuring topical relevance to the prompt via cosine similarity or estimating coherence through semantic similarity between adjacent sentences. While theoretically appealing, these approaches introduce assumptions that are highly dependent on the assessment context and the underlying rubric. For instance, a high semantic similarity to the prompt does not necessarily imply higher language proficiency, and originality or creative divergence—valued in some educational contexts—may even reduce such similarity scores. Consequently, the interpretation of embedding-based features becomes substantially more subtle and setting-dependent.

Given these considerations, this thesis adopts a deliberately conservative and theoretically grounded approach. Rather than pursuing highly context-sensitive vector-based metrics, we focus on classical, well-established engineered linguistic features that have a clear history in AES research and a transparent relationship to assessment rubrics. These include sentence-level, token-level, lexical, readability, and syntactic metrics. By doing so, we aim to provide a strong, interpretable baseline that complements transformer-based representations and allows for meaningful analysis of how explicit linguistic properties contribute to automated language level assessment.

In the following sections, we describe the selected feature sets in detail, outline the feature computation pipeline, and analyze their statistical properties and role within the overall modeling framework.

4.2 Features Extracted

To capture multiple dimensions of written language proficiency, a set of 19 engineered linguistic features was extracted from each essay. These features are designed to approximate aspects of human as-

sessment rubrics, combining surface-level statistics with lexical, syntactic, and error-based indicators. Features are grouped into six categories: statistics, lexical properties, part-of-speech distributions, syntactic complexity, readability, and error rates.

4.2.1 The Length-Agnostic Constraint: Exclusion of Volume Metrics

While raw text length is often strongly correlated with proficiency in timed essay exams, relying on volume metrics (e.g., total word count) introduces significant bias. A model reliant on length may incorrectly penalize succinct, high-proficiency writers or reward verbose, low-proficiency writers ("gaming the system"). Furthermore, length is heavily dependent on specific prompt constraints (e.g., "write 200 words" vs. "write 500 words"), limiting the model's generalizability across different testing conditions.

Therefore, adhering to the feedback from domain experts and to ensure the system measures *proficiency* rather than *productivity*, we explicitly exclude all raw volume counts. The feature set is strictly limited to intensive metrics—ratios, densities, and averages—which remain robust regardless of total essay duration.

4.2.2 Sentence- and Token-Level Statistics

To characterize structural properties of the text, sentence- and token-level statistics are computed:

- **Mean and maximum sentence length** (in tokens).
- **Mean and maximum token length** (in characters).

We include mean sentence length as it serves as a proxy for general fluency and syntactic capacity, while maximum length captures the learner's "ceiling"—their ability to construct at least one complex, sustained linguistic unit. Token length specifically targets derivational complexity (e.g., distinguishing simple roots from complex, multi-morphemic academic vocabulary).

Conversely, we explicitly exclude sentence length in characters, as it provides little orthogonal information not already captured by the combination of token length and sentence length in tokens. We also omit minimum values (as learners at all levels produce short sentences for stylistic reasons) and median values, which were found to be statistically redundant with the mean during exploratory analysis.

4.2.3 Lexical Features

Lexical richness is approximated using two complementary metrics, each capturing a different aspect of vocabulary usage:

- **Moving-Average Type-Token Ratio (MATTR):** Calculated with a window of 50 tokens. We explicitly select MATTR over the standard Type-Token Ratio (TTR). Standard TTR is known to decrease mechanically as text length increases; MATTR eliminates this bias, providing a diversity metric that is comparable across essays of varying lengths [6].

- **Lexical Density:** Measured as the proportion of content words (nouns, verbs, adjectives, adverbs) relative to the total alphanumeric word count.

$$\text{Lexical Density} = \frac{N_{\text{content}}}{N_{\text{words}}} \tag{4.1}$$

where N_{words} represents the count of tokenized words excluding punctuation marks. This captures the density of information; higher density distinguishes academic, informational writing from "filler-heavy" conversational speech.

4.2.4 Part-of-Speech Ratios

Relative frequencies of selected part-of-speech categories are included:

- **Noun, Verb, Adjective, and Pronoun Ratios.**
- **Function Word Ratio** (prepositions, conjunctions, particles, auxiliaries, determiners).

These ratios are inspired by functional approaches to register and proficiency analysis and provide insight into grammatical maturity, referential density, and discourse structure. For example, increased noun and adjective usage often accompanies advanced descriptive ability, while excessive pronoun reliance may signal limited lexical resources.

4.2.5 Syntactic Complexity Features

Syntactic sophistication is measured using dependency parsing structures, normalized to remain length-neutral:

- **Clauses per Sentence (normalized):** Unlike raw clause counts, which correlate with volume, this ratio captures the density of subordination and coordination within a single sentence.
- **Average Dependency Tree Depth:** The mean maximum depth of the syntactic tree across all sentences.

$$\text{Avg Tree Depth} = \frac{1}{S}\sum_{i=1}^{S} \max(\text{depth}(t) \mid t \in \text{Sentence}_i) \tag{4.2}$$

where S is the number of sentences and $\text{depth}(t)$ is the distance from the root to token t.

Clause density serves as a proxy for the ability to express multi-propositional content, while tree depth reflects hierarchical syntactic organization. These features are particularly relevant for Russian, where rich morphology and flexible word order allow for structurally complex constructions even in relatively short texts.

4.2.6 Readability Indices

We compute four standard readability formulas adapted for Cyrillic script [13, 8, 17, 4]:

- Automated Readability Index (ARI)
- Flesch Reading Ease
- SMOG Index
- Coleman-Liau Index (CLI)

To ensure these metrics serve as valid proxies for complexity rather than length, we utilize indices that rely on intensive rates. For example, the **Coleman-Liau Index (CLI)** relies on character counts, making it robust to phonetic spelling errors common in L2 data:

$$\text{CLI} = 0.0588L - 0.296S - 15.8 \tag{4.3}$$

where L is the average number of letters per 100 words, and S is the average number of sentences per 100 words.

Although these metrics were originally developed for English, they primarily rely on surface features such as word length and sentence length. In this work, they are used not as absolute readability measures, but as coarse proxies for textual complexity, complementing language-specific features.

4.2.7 Error-Based Features

Finally, proficiency is defined not only by complexity but by accuracy. We include:

- **Spelling Error Rate** (per 100 tokens).
- **Grammatical Error Rate** (per 100 tokens).

These features directly operationalize the "Accuracy" dimension of the ACTFL rubric. By normalizing per 100 tokens, we ensure that the metric reflects the density of errors (a sign of lower control) rather than the total number of errors (which would simply track essay length). Distinct features are maintained for spelling and grammar to differentiate between orthographic deficits (common in heritage learners) and morphosyntactic deficits (common in L2 learners).

4.3 Feature Computation Pipeline

Feature extraction is implemented as a modular Python pipeline combining established open-source NLP tools for Russian. The goal is to ensure reproducibility, linguistic validity, and compatibility with low-resource institutional environments.

Tokenization, sentence segmentation, and POS tagging are performed with **spaCy** [10], while **Stanza** [20] handles dependency parsing. Both tools provide pretrained models for Russian and are widely used in contemporary NLP research. This combination leverages spaCy's fast and robust surface-level processing alongside Stanza's accurate syntactic trees and morphological annotations, ensuring high-quality feature extraction for both token- and sentence-level metrics.

Syntactic tree depth is computed via a depth-first traversal of dependency trees, measuring the longest path from each token to the root. Clause counts are approximated using finite verb detection and selected dependency relations (e.g., `root`, `conj`, `ccomp`, `advcl`), following common practice in computational studies of syntactic complexity.

Lexical and readability metrics are computed using the `textstat` library. While some indices were originally designed for English, they are retained as consistent surface-level complexity indicators applied uniformly across all texts.

Spelling and grammatical error rates are obtained using **LanguageTool** [18] for Russian. While not designed specifically for learner corpora, LanguageTool provides broad coverage of orthographic and morphosyntactic phenomena and is commonly used in educational NLP as a high-precision error detection baseline. To reduce noise unrelated to proficiency, superficial casing-related rules are excluded, focusing instead on orthographic and morphosyntactic errors. Error counts are normalized by token count to allow comparison across essays of different lengths.

All features are computed independently for each essay and assembled into a fixed-length feature vector. Prior to model training, features are standardized using Z-score normalization. While some features exhibit collinearity, this is considered an inherent property of language proficiency rather than a flaw: richer vocabulary, longer sentences, and lower error rates naturally co-occur in human writing. The modelling approaches explored in later chapters are therefore chosen to tolerate and exploit this interdependence, rather than attempting to artificially eliminate it.

4.4 Exploratory Data Analysis of the Feature Space

This section presents an exploratory analysis of the engineered feature space described in the previous sections. The goal of this analysis is threefold: (i) to assess the individual relationship between each feature and the proficiency label, (ii) to inspect how feature distributions evolve across proficiency levels, and (iii) to evaluate internal dependencies between features, including potential multicollinearity and unintended data leakage. Together, these analyses provide empirical justification for the chosen feature set and inform subsequent modeling decisions.

4.4.1 Correlation with Proficiency Level (Pearson)

As an initial step, Pearson correlation coefficients were computed between each feature and the proficiency label. Although the ACTFL-based proficiency scale is ordinal by nature, Pearson correlation provides a widely used and easily interpretable first approximation of linear association. At this stage, the label is therefore treated as a numeric variable.

Figure 4.1 presents the resulting correlation profile. Several distinct patterns emerge:

- **Syntactic Signals:** Syntactic complexity measures are among the strongest predictors. Average dependency tree depth exhibits the highest correlation with proficiency, followed closely by clause-based features. This aligns with SLA theory, confirming that the capacity to embed clauses is a primary hallmark of advanced proficiency.

- **Readability Proxies:** Indices such as the Coleman-Liau Index (CLI) and Automated Readability Index (ARI) show strong positive correlations. Since these metrics rely heavily on word length and sentence length, they effectively act as proxies for lexical and structural sophistication.
- **The Adjective-Pronoun Trade-off:** Part-of-speech ratios support linguistically intuitive interpretations. Adjective and pronoun ratios show stronger correlations (positive and negative, respectively) than noun or verb ratios. This suggests that higher-level learners increasingly enrich their language with modifiers while relying less heavily on simple pronominal references.
- **Error Asymmetry:** Spelling error rate correlates more strongly with proficiency than grammar error rate. This indicates that orthographic accuracy improves consistently across levels, whereas grammatical errors remain a persistent challenge even in more advanced learner texts.

Overall, syntactic features, error rates, readability metrics, and count-based measures tend to outperform simple ratios in terms of linear association with the target label.

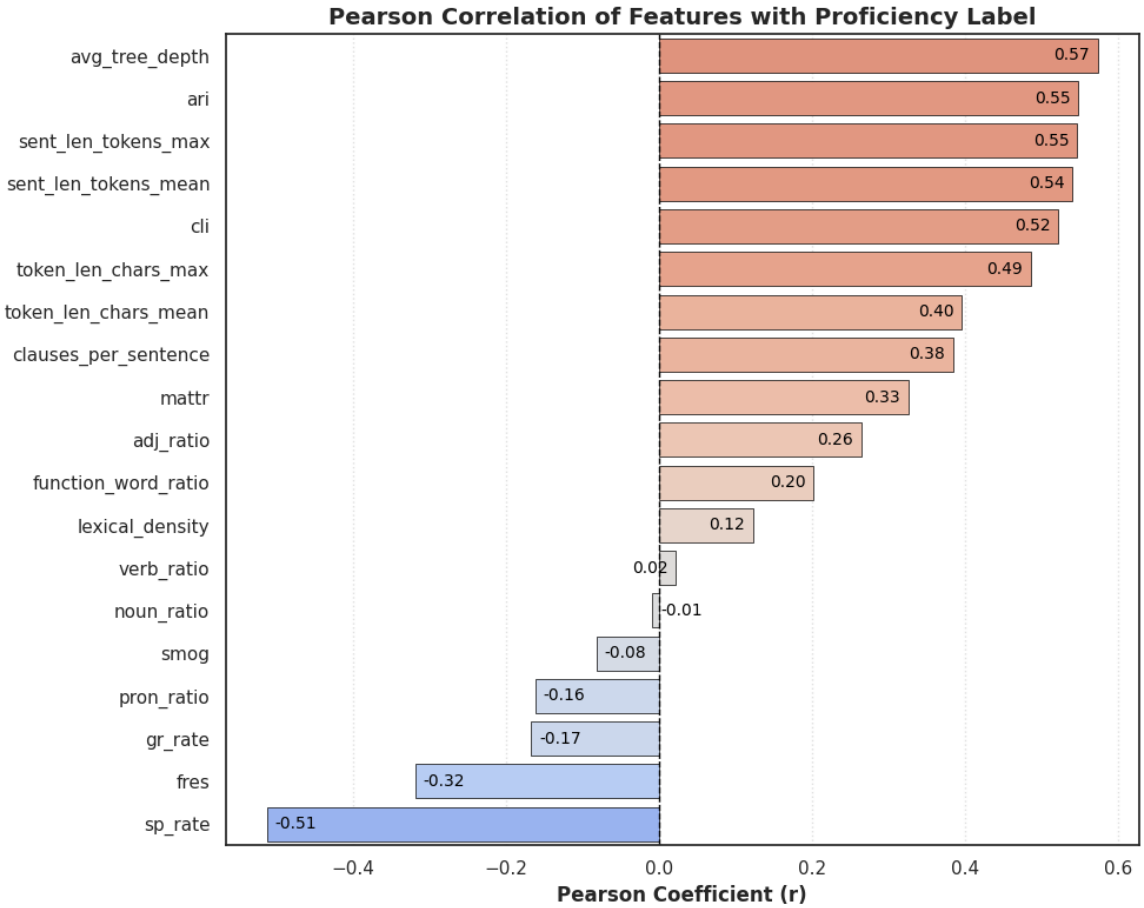

Figure 4.1: Pearson correlation coefficients (r) between extracted linguistic features and proficiency labels. Syntactic features (Tree Depth) and intensive error rates show the strongest linear signals.

4.4.2 Correlation with Proficiency Level (Spearman)

To account for the ordinal nature of the proficiency scale, Spearman rank correlations were additionally computed. Unlike Pearson correlation, Spearman's ρ captures monotonic relationships without

assuming linear spacing between adjacent proficiency levels.

Figure 4.2 shows that the overall ranking of features remains largely consistent with the Pearson analysis, indicating robust monotonic relationships. However, notable differences emerge that highlight the non-linear nature of the data.

In particular, ARI becomes the most strongly associated feature alongside average tree depth under monotonic ranking. Furthermore, the relative importance of `lexical_density` versus `function_word_ratio` is reversed compared to the Pearson setting. These shifts confirm that the relationship between linguistic features and ACTFL levels is **monotonic but non-linear**. The linguistic "distance" traversed from Novice to Intermediate is not necessarily equal to the step from Advanced to Superior, validating the decision to explore ordinal-aware loss functions in Chapter 5.

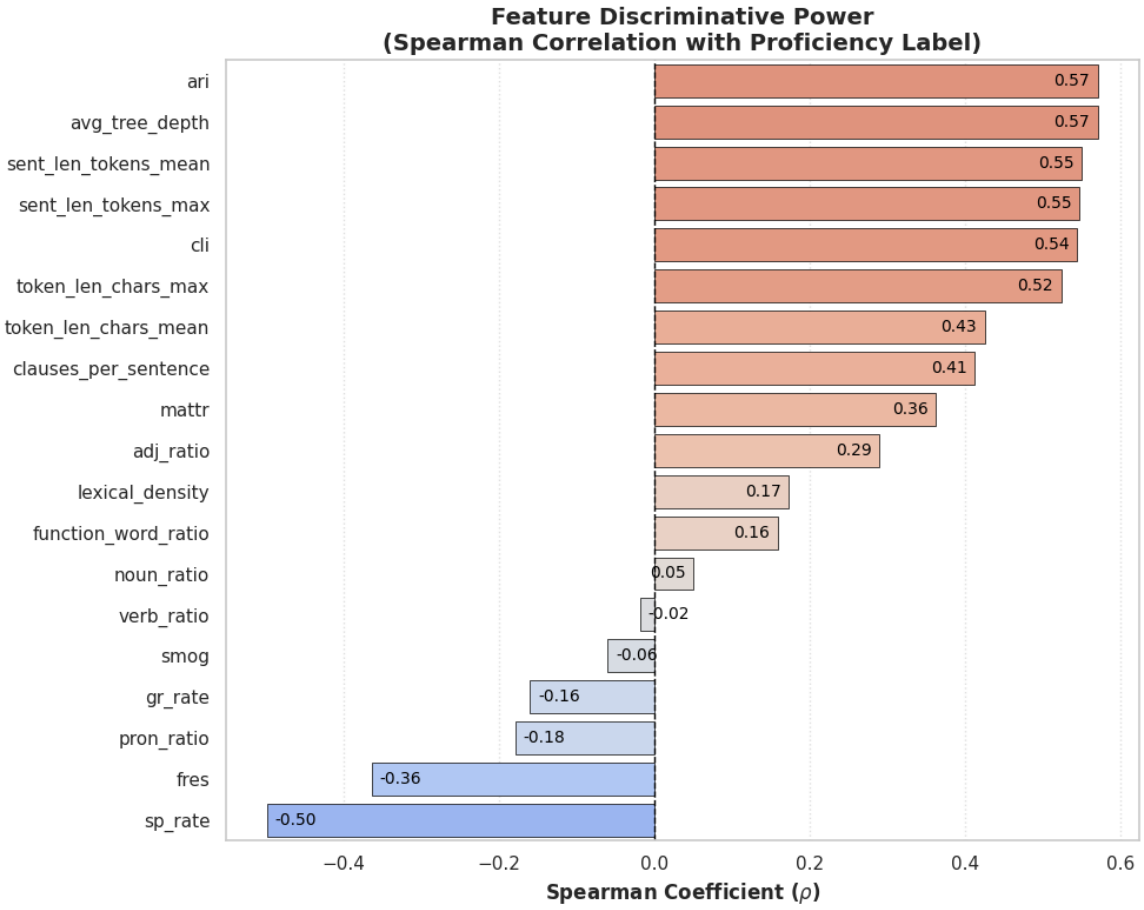

Figure 4.2: Spearman rank correlation coefficients (ρ) showing discriminative power. The ranking remains largely consistent with Pearson, though ARI becomes the dominant feature under monotonic ranking.

4.4.3 Feature Distributions Across Proficiency Levels

To examine how feature values vary across proficiency levels, feature distributions are visualized using boxplots grouped by label (Figure 4.3). This representation enables direct comparison of medians, dispersion, and outliers across proficiency bands.

For many features, the boxplots reveal clear separation between levels, confirming their discriminative potential. A recurring pattern is that the lowest proficiency levels (Novice Mid/High) exhibit the

largest variance, while intermediate and higher levels show more compact distributions. This aligns with the "Novice Paradox": extremely short texts often statistically mimic "perfect" texts (zero errors, standard length) or fail completely, leading to high heterogeneity.

We also observe the Floor Effect in SMOG. The SMOG index exhibits almost zero variance across the dataset, effectively collapsing to its mathematical floor constant (≈ 3.12). Since SMOG relies on the count of polysyllabic words (3+ syllables), its failure here reflects the extreme scarcity of such vocabulary in L2 learner Russian. Consequently, SMOG provides negligible signal for this specific task.

4.4.4 Multicollinearity and Redundancy

To investigate internal dependencies within the feature space, a correlation matrix was computed across all features and visualized as a heatmap (Figure 4.4).

Three distinct clusters of correlation are observable:

- **Surface Complexity Cluster:** Readability metrics (ARI, CLI, FRES) are highly intercorrelated ($|r| > 0.8$), reflecting their shared mathematical reliance on word and sentence length.
- **Syntactic-Length Covariance:** Average tree depth correlates moderately with mean sentence length ($r \approx 0.5$). This is linguistically valid, as longer sentences in Russian typically require deeper dependency structures to remain grammatical.
- **Orthogonality of Grammar Errors:** Notably, grammar error rate (`gr_rate`) shows weak correlations with most other features, including spelling error rate. This suggests that morphosyntactic control is an independent dimension of proficiency—a learner can possess a rich vocabulary (high MATTR) yet still struggle with case endings.

While these correlations indicate some redundancy, they do not present pathological multicollinearity that would prevent model convergence.

4.4.5 Verification of Length Independence

Finally, potential data leakage was assessed by examining correlations between features and raw token count. Since essay length is known to correlate with proficiency in exam settings, it is crucial to ensure that features do not trivially encode text length.

Figure 4.5 demonstrates that strictly intensive metrics—such as `sp_rate`, `lexical_density`, and `noun_ratio`—show negligible correlation with volume ($|r| < 0.2$). While features like `avg_tree_depth` show moderate positive correlation ($r \approx 0.4$), this is attributable to **natural co-variation due to proficiency**: higher proficiency causes *both* longer essays and deeper syntax.

Crucially, no feature displays near-perfect dependence on essay length. These results confirm that the engineered feature set captures qualitative aspects of linguistic proficiency rather than acting as a proxy for productivity.

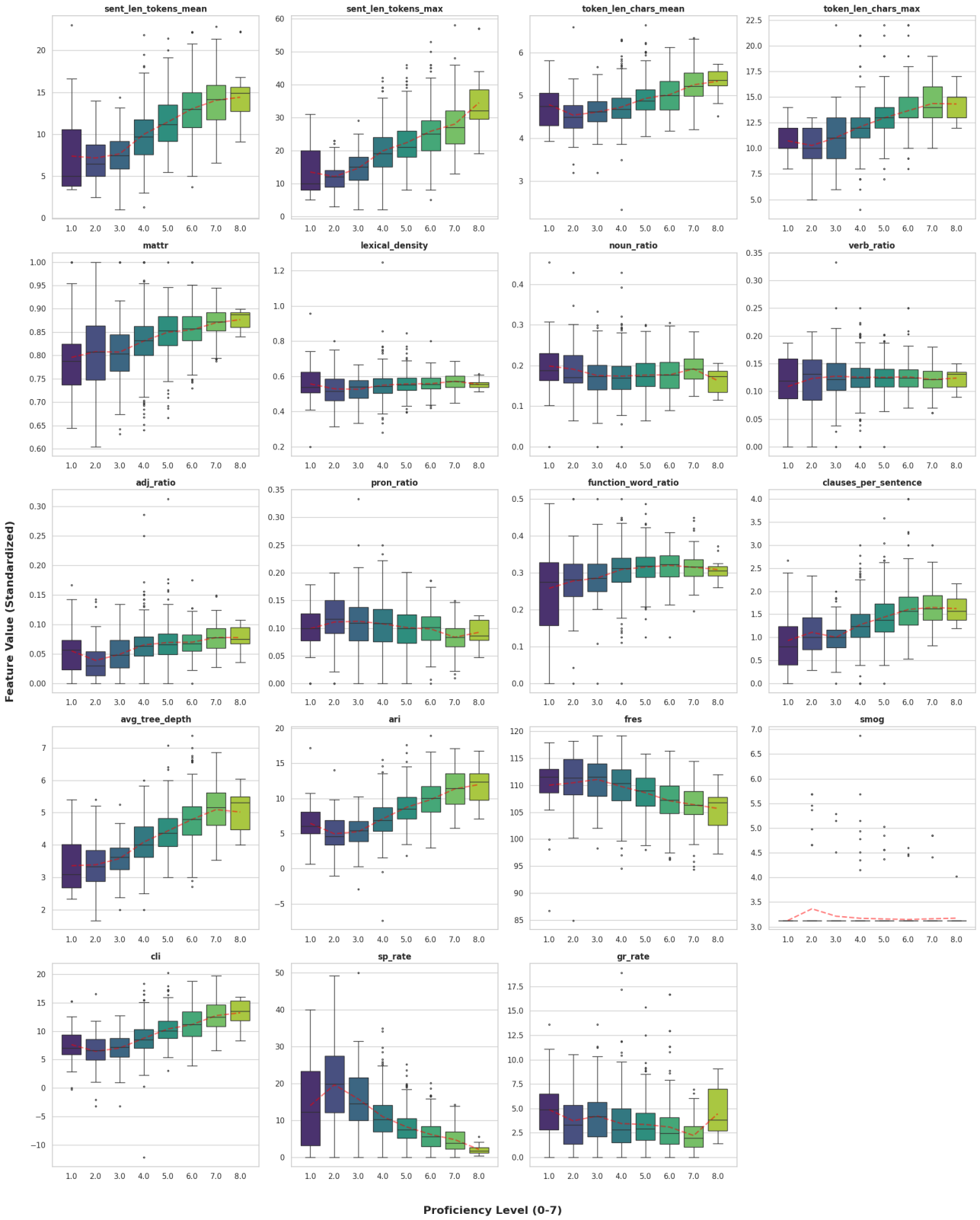

Figure 4.3: Distribution of standardized feature values across proficiency levels (0-7). The red dashed line indicates the mean trend. Note the floor effect in SMOG and the high variance in Novice levels for surface metrics.

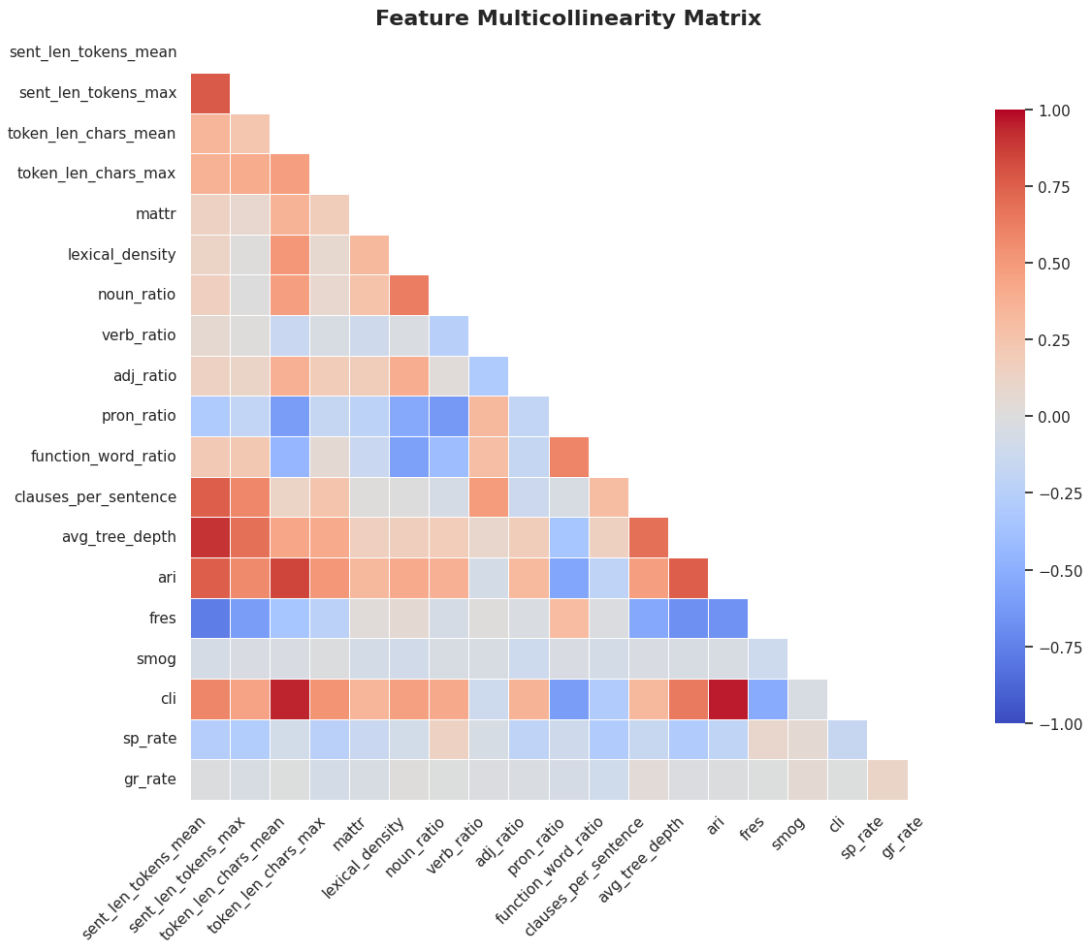

Figure 4.4: Feature-to-feature correlation heatmap. Strong red blocks indicate high multicollinearity, particularly among readability indices (ARI, FRES, CLI).

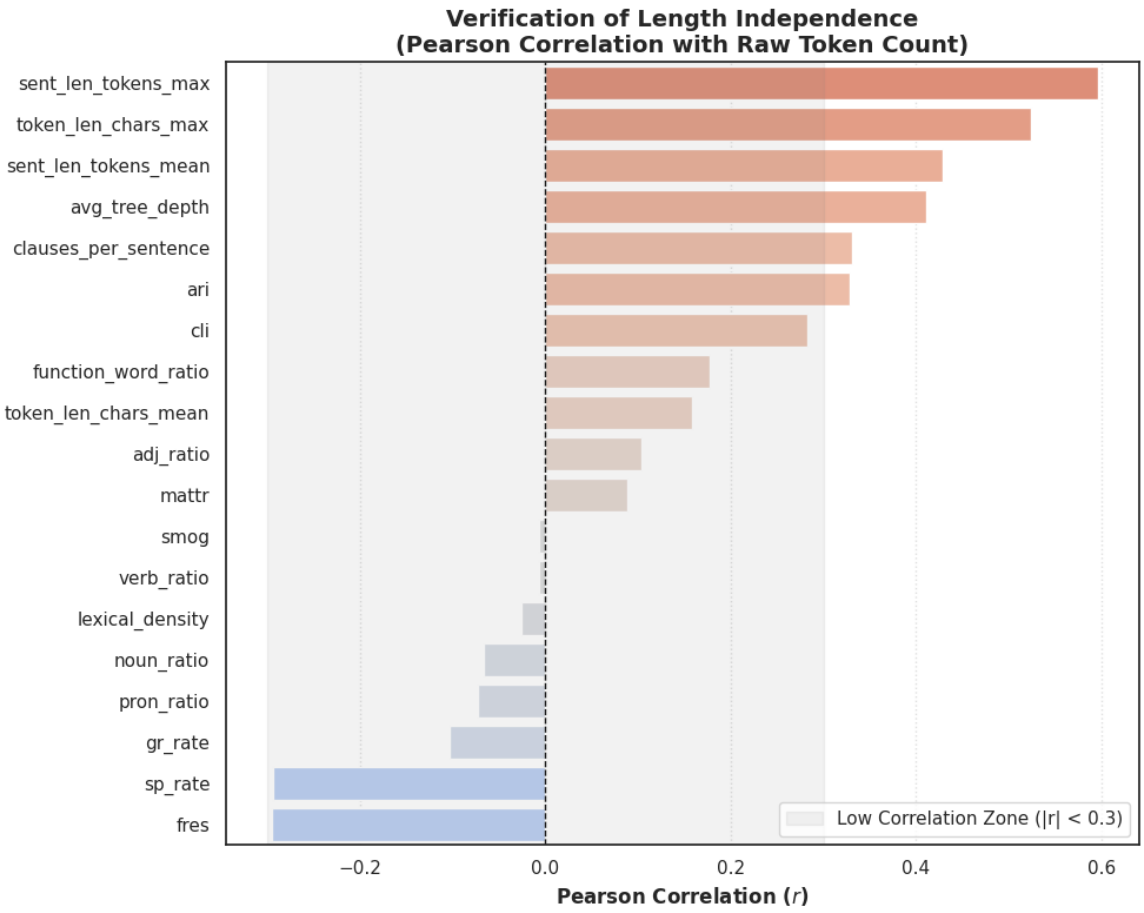

Figure 4.5: Correlation of engineered features with raw essay token count. The majority of intensive features fall within the low-correlation safety zone ($|r| < 0.3$), verifying the length-agnostic design.

4.5 Feature Normalization and Preprocessing

The extracted linguistic feature set exhibits substantial heterogeneity in terms of numerical ranges and physical units. For example, readability metrics such as the Flesch Reading Ease Score (FRES) typically range between 0 and 100 (and may become negative for extremely complex texts), while part-of-speech ratios are bounded within the interval $[0,1]$. Similarly, average dependency tree depth varies within a narrow single-digit range (e.g., 2.5 to 6.0), whereas spelling and grammatical error rates are computed per 100 tokens and may take on comparatively larger values.

4.5.1 Standardization Strategy

Although neural networks are theoretically capable of learning appropriate weights for inputs with heterogeneous scales—for instance, by assigning smaller coefficients to features with larger magnitudes—using raw feature values presents practical challenges in limited-data settings. Large discrepancies in feature scales can lead to an ill-conditioned optimization landscape, resulting in unstable gradients, slow convergence, or inefficient use of early training epochs.

Given the modest size of the training corpus (see Chapter 3), it is undesirable for the model to expend its limited learning capacity on compensating for scale differences rather than on capturing discriminative linguistic patterns. For this reason, all continuous engineered features are standardized using **Z-score normalization** prior to model training.

Each feature value x is transformed as follows:

$$z = \frac{x - \mu}{\sigma} \tag{4.4}$$

where μ and σ denote the mean and standard deviation computed exclusively on the **training set**. These statistics are stored and subsequently applied to the validation and test sets to ensure consistency and to prevent data leakage. This procedure corresponds to standard practice in both classical machine learning and neural modeling pipelines.

4.5.2 Implications for Hybrid Modeling

Feature normalization is not merely a preprocessing convenience but a structural prerequisite for the hybrid modeling strategies explored in Chapter 5. In the proposed architectures, the standardized linguistic feature vector is concatenated with the dense semantic embedding produced by a Transformer encoder (e.g., the `[CLS]` representation from XLM-RoBERTa).

Transformer embeddings use Layer Normalization, which constrains their values. Typically, embeddings cluster between -2 and 2. Concatenating these embeddings with unnormalized linguistic features (e.g., a raw FRES score of 60.5) would cause the latter to dominate the linear transformations in downstream layers, effectively overwhelming the semantic signal.

Standardization ensures **numerical compatibility** between heterogeneous feature sources, allowing both engineered linguistic metrics and neural representations to contribute meaningfully and pro-

portionally during optimization. This alignment is particularly important in low-resource scenarios, where stable and efficient training dynamics are critical for generalization.

4.6 Summary

This chapter detailed the construction and validation of a set of 19 linguistically motivated features designed to capture multidimensional aspects of L2 Russian proficiency.

Our exploratory analysis confirms that these features provide strong, discriminative signals. Syntactic complexity and error density emerge as primary indicators of proficiency, while lexical richness and part-of-speech distributions offer complementary information. The analysis also highlighted the non-linear nature of the ACTFL scale and the noisy distributional properties of Novice-level texts, suggesting that simple linear models may be insufficient for optimal performance.

Having established a robust and interpretable feature set, the following chapter will formalize the modeling approaches used in this study. We will define the baseline models that utilize these engineered features and introduce the neural architectures that will serve as the core of our experimental comparison.

Chapter 5

Modelling Approaches

Having established the theoretical framework in Chapter 2 and validated the dataset and feature set in Chapters 3 and 4, this chapter details the specific predictive architectures employed to estimate essay proficiency.

The modeling strategy is designed to answer three hierarchical research questions:

1. **Baseline Performance:** How do standard machine learning paradigms—specifically nominal classification and linear regression—perform when applied to the engineered feature set?
2. **The Ordinal Advantage:** Can performance be improved by explicitly modeling the rank-ordered nature of ACTFL levels (via CORAL/CORN) and augmenting the feature space with lexical profiling?
3. **The Neural Gap:** To what extent do deep contextual embeddings (XLM-RoBERTa) outperform rigorous feature-based models, and can their combination (Hybrid) yield superior robustness?

This chapter is structured by complexity: we begin with statistical baselines (Section 5.1), progress to advanced ordinal feature-based models (Section 5.2), and conclude with deep neural architectures and fusion strategies (Sections 5.3 and 5.4). This hierarchical progression allows us to systematically evaluate the contribution of each modeling innovation, from standard feature-based methods to complex deep architectures, ensuring that performance gains can be directly attributed to specific design choices.

5.1 Baselines and Benchmarks

To rigorously evaluate the utility of increasingly complex modeling strategies, it is essential to establish strong lower-bound benchmarks. We therefore employ three categories of baselines: *dummy baselines* (purely statistical heuristics), *nominal baselines* (standard multi-class classification), and *metric baselines* (linear regression). Collectively, these models serve as a *control group*, representing the level of performance achievable using conventional techniques before introducing task-specific ordinal constraints or deep contextual representations.

This structured progression enables a principled comparison: any observed improvement in later sections can be explicitly attributed to modeling choices such as the incorporation of ordinal loss functions, contextual embeddings, or hybrid feature fusion.

5.1.1 Problem Formulation and Notation

Before defining the individual baseline models, we formalize the essay proficiency assessment task. Let

$$\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{N}$$

denote a dataset of N learner essays, where $x_i \in \mathbb{R}^d$ represents the feature vector of the i-th essay (with $d = 19$ for the base engineered linguistic feature set), or alternatively a sequence of tokens in the case of neural models. The target variable y_i is drawn from an ordered label set

$$\mathscr{Y} = \{r_0, r_1, \ldots, r_{K-1}\},$$

corresponding to discrete ACTFL proficiency levels ranging from *Novice Low* to *Superior*.

Crucially, these labels satisfy a strict ordinal constraint:

$$r_0 \prec r_1 \prec \cdots \prec r_{K-1}.$$

However, the *psychometric distance* between adjacent levels r_k and r_{k+1} is unknown and likely non-uniform. This distinction lies at the core of the modeling challenge. As demonstrated in the following sections, baseline approaches typically fail to capture this ordinal nuance: nominal classifiers ignore the ordering altogether, while metric regressors assume equal spacing between proficiency levels—an assumption that is difficult to justify from a second language acquisition perspective.

5.1.2 Dummy Baselines

Dummy baselines ignore the linguistic content of the essays entirely and rely solely on the label statistics observed in the training set, denoted by $P(Y_{\text{train}})$. Effectively, these models are *blind* to the input representation x_i and produce constant predictions independent of the essay text.

Although often omitted in simpler studies, such heuristics play a crucial role in experimental validation. They define the *null hypothesis* of the task: any sophisticated model—whether feature-based or neural—must significantly outperform these trivial strategies to demonstrate that it has extracted meaningful linguistic signal rather than merely exploiting dataset imbalance.

We implement three statistical heuristics, each corresponding to an optimal solution under a specific loss function.

Majority Class (Mode) Predictor

The Majority Class classifier (also known as the Zero-Rule algorithm) always predicts the most frequent proficiency level observed in the training set:

$$\hat{y}_{\text{mode}} = \arg\max_{k \in \mathcal{Y}} \sum_{i=1}^{N} \mathbb{I}(y_i = k), \tag{5.1}$$

where $\mathbb{I}(y_i = k)$ is an indicator function that returns 1 when essay i belongs to proficiency level k, and 0 otherwise. In other words, the summation simply counts how many essays in the training set are labeled as level k.

In our dataset, which exhibits a skew toward intermediate proficiency, this baseline consistently predicts *Intermediate Mid* (Label 4) for all essays.

Theoretical Expectation. This strategy theoretically maximizes **Accuracy** in highly imbalanced label distributions. However, it provides no discriminative capacity and fails to capture any meaningful ranking information. Since the Quadratic Weighted Kappa (QWK) metric explicitly corrects for chance agreement, and a constant prediction corresponds to chance-level agreement with the modal class, we expect this baseline to achieve a QWK score close to 0.0.

Mean and Median Regressors

The remaining dummy baselines treat ordinal proficiency labels as continuous numerical values and predict a global measure of central tendency computed from the training partition.

Mean Regressor. The Mean Regressor predicts the arithmetic mean of the training labels,

$$\mu_y = \frac{1}{N} \sum_{i=1}^{N} y_i,$$

for all test instances. Because the proficiency labels are discrete, the final prediction is obtained via rounding:

$$\hat{y}_{\text{mean}} = \text{round}(\mu_y).$$

Theoretical Expectation. From a statistical standpoint, the arithmetic mean is the unique minimizer of the Mean Squared Error (MSE). Consequently, this baseline establishes a strong lower bound for RMSE performance and may rival more complex models when the signal-to-noise ratio in the linguistic features is low.

Median Regressor. The Median Regressor predicts the median value $\tilde{y}$ of the training label distribution for every input instance. This provides a robust constant baseline that is less sensitive to extreme values than the mean predictor.

Theoretical Expectation. The median is the robust minimizer of the Mean Absolute Error (MAE). In the presence of outliers—such as rare *Superior* or *Novice Low* essays—it provides a more stable estimate of central tendency than the mean.

While these regressors are incapable of ranking individual essays—since they assign the same score to all inputs—their error values (MAE and RMSE) define a meaningful *performance floor*. As discussed in Chapter 6, any trained model whose RMSE exceeds that of the Mean Regressor can be considered to have failed to extract signal beyond a trivial estimate based solely on dataset-level statistics.

5.1.3 The Nominal Baseline: Cross-Entropy Classification

To evaluate the impact of ignoring the inherent ordering of proficiency levels, we train a standard **Multinomial Logistic (Softmax) Regression** classifier. This model represents a nominal approach, treating the proficiency scale as a set of independent categories rather than an ordered progression.

Formally, the model treats the target variable y as a nominal class label $c \in \{0, \dots, K-1\}$. Given the standardized feature vector $x_i \in \mathbb{R}^d$, the model learns a separate weight vector w_c and bias b_c for each class. The probability that essay i belongs to proficiency level c is computed via the Softmax function:

$$P(y_i = c|x_i) = \frac{e^{w_c^T x_i + b_c}}{\sum_{j=0}^{K-1} e^{w_j^T x_i + b_j}} \tag{5.2}$$

The model is optimized by minimizing the Cross-Entropy (or Negative Log-Likelihood) loss:

$$\mathscr{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=0}^{K-1} \mathbb{I}(y_i = c) \log P(y_i = c|x_i) \tag{5.3}$$

The model parameters $\{w_c, b_c\}_{c=0}^{K-1}$ are estimated by minimizing this loss with respect to the weights and biases:

$$\{w_c^*, b_c^*\}_{c=0}^{K-1} = \arg\min_{\{w_c, b_c\}} \mathscr{L}_{CE}.$$

Theoretical Limitation: Ordinal Ignorance

In the context of AES, the fundamental weakness of this baseline is its ordinal ignorance. The Cross-Entropy loss function imposes a uniform penalty structure: the cost of misclassifying a *Novice* essay as *Intermediate* is mathematically identical to misclassifying it as *Superior*, provided the predicted probability mass is the same.

Consequently, while this model may achieve competitive *Accuracy* (exact matches), we anticipate suboptimal performance on the Quadratic Weighted Kappa (QWK) metric. Because QWK penalizes errors quadratically according to the distance between true and predicted ranks, a loss function that treats all misclassifications as equidistant is theoretically misaligned with the evaluation objective.

5.1.4 The Metric Baseline: Linear Regression

Linear Regression serves as a widely used historical baseline in Automated Essay Scoring, historically popularized by foundational systems such as ETS's *e-rater*. In contrast to the nominal baseline, this

approach explicitly models the ranking information but introduces a strong parametric assumption: it treats the discrete proficiency labels as continuous values lying on a metric scale with equal intervals.

The model predicts a continuous raw score $\hat{y}_{raw} \in \mathbb{R}$ via a linear projection of the feature vector:

$$\hat{y}_{raw} = w^T x + b \tag{5.4}$$

Parameters are estimated by minimizing the Mean Squared Error (MSE), which penalizes large deviations more heavily than small ones:

$$\mathscr{L}_{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_{raw})^2 \tag{5.5}$$

Inference and Rounding

Since the ACTFL proficiency scale consists of discrete integers, the continuous output of the regression model must be mapped back to the valid label set $\mathscr{Y}$. We apply a clipping and rounding operation during inference:

$$\hat{y}_{final} = \text{round}(\min(\max(\hat{y}_{raw}, 0), K-1)) \tag{5.6}$$

Theoretical Limitation: The Equal-Interval Assumption

While Linear Regression is often robust to noise in small datasets and captures the general direction of proficiency (e.g., higher syntactic complexity correlates with higher scores), it suffers from the "Equal-Interval Assumption." The model assumes that the psychometric distance between *Novice Low* (0) and *Novice Mid* (1) is identical to the distance between *Advanced High* (6) and *Superior* (7).

In second language acquisition, this linearity is rarely observed; progress at lower levels is often rapid, while the distinction between upper-advanced levels relies on subtle nuance. By imposing a linear constraint on a non-linear ordinal phenomenon, this baseline is expected to struggle with discriminability at the extreme ends of the scale, though it typically outperforms nominal classification by virtue of respecting the global order.

5.2 Advanced Feature-Based Approaches

Having established the performance baselines with standard baselines, we now advance the modeling strategy in two directions: first, by augmenting the feature space to explicitly capture lexical sophistication, and second, by adopting loss functions specifically designed for ordinal data.

This section details the construction of the extended feature set and defines the methodologies used to exploit it, ranging from an upgraded metric regression to specialized ordinal logistic architectures.

5.2.1 Extended Feature Extraction: Lexical Profiling

The core feature set described in Chapter 4 ($d = 19$) primarily captures morphosyntactic complexity and error density. However, a close reading of the ACTFL guidelines suggests that *Lexical Sophistication*—the range and precision of vocabulary—is a primary discriminator at higher proficiency levels. To explicitly model this dimension, we augment the input vector with five additional indicators derived from Common European Framework of Reference (CEFR) graded vocabulary lists.

Using the Russian proficiency vocabulary lists provided by the University of Helsinki (Levels A1 through C1), we calculate the **Lexical Coverage Ratio** for each essay. Critically, to avoid multicollinearity (where a C1 word is also counted as a B2 word), we employ a *disjoint* assignment strategy: a lemma is assigned strictly to the lowest level list in which it appears.

While ACTFL and CEFR represent distinct proficiency frameworks, CEFR-graded vocabulary lists provide a practical and widely used proxy for modeling lexical sophistication across proficiency levels.

Formally, let L_T be the set of unique lemmas in essay T, and V_k be the disjoint set of vocabulary items specific to CEFR level $k \in \{A1, \ldots, C1\}$. The coverage ratio R_k is calculated as:

$$R_k = \frac{|L_T \cap V_k|}{|L_T|} \tag{5.7}$$

This operation yields five new continuous features: `ratio_A1` through `ratio_C1`. The final feature vector x for the advanced models thus has dimensionality $d = 24$.

Validation of Lexical Features

To verify the discriminatory power of these proposed lexical features, we analyzed their distribution across proficiency levels. As illustrated in Figure 5.1, the coverage of high-proficiency vocabulary (B2/C1) exhibits a clear positive monotonicity with the ground-truth labels. The distinction is particularly pronounced between *Advanced* (Labels 5-6) and *Superior* (Label 7) essays, a boundary where morphosyntactic features often plateau. Conversely, reliance on A1 vocabulary shows a consistent downward trend as proficiency increases.

Furthermore, to ensure that these ratios are not merely proxies for essay length (e.g., the hypothesis that longer essays statistically accumulate more rare words by chance), we computed the Pearson correlation between each new feature and the raw token count (Figure 5.2). The results confirm that the vocabulary ratios fall within the "independence zone" ($|r| < 0.3$). This indicates that the features capture the *density* of sophisticated vocabulary rather than simple verbosity, satisfying the length-independence constraint established in Chapter 4.

5.2.2 Intermediate Benchmark: Linear Regression on Extended Features

Before employing complex ordinal loss functions, we first isolate the contribution of the new lexical features. We define an intermediate benchmark by re-training the **Linear Regression** model (originally defined in Section 5.1.4) on the augmented 24-dimensional feature vector.

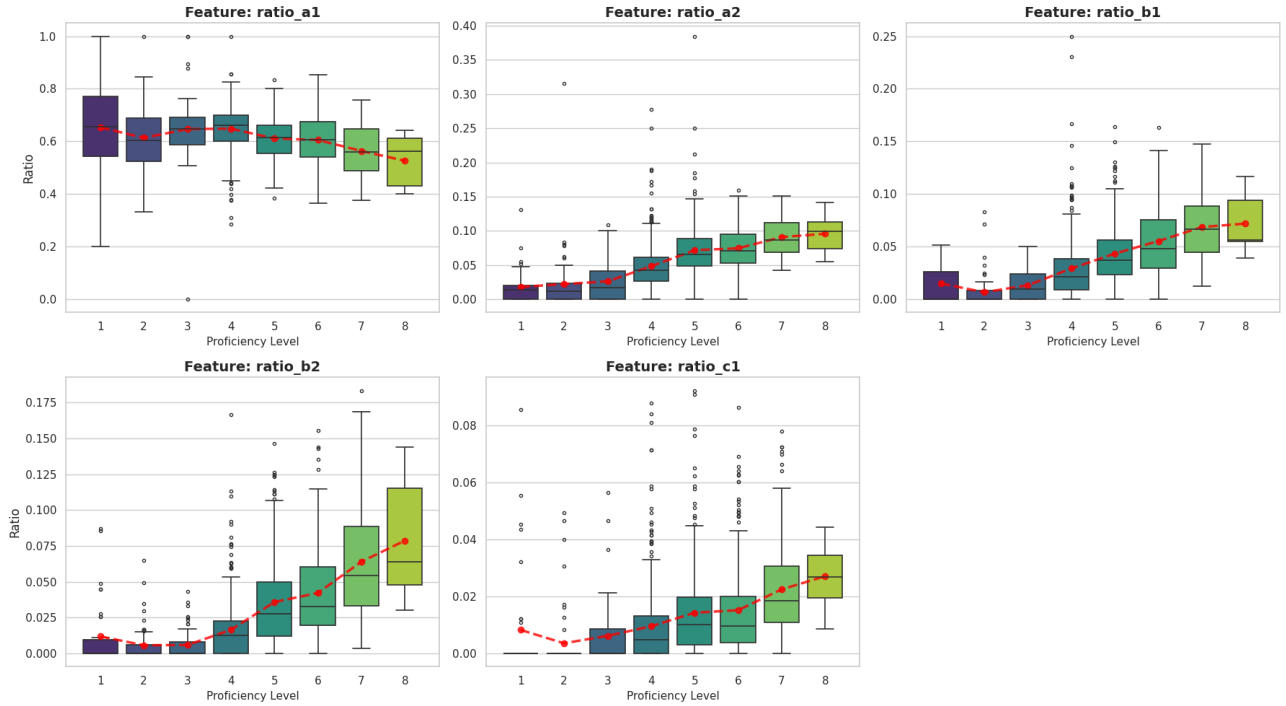

Figure 5.1: Distribution of CEFR-based vocabulary coverage ratios across proficiency levels. Higher proficiency levels are characterized by a significantly denser usage of B2 and C1 lemmas, validating the features' monotonicity.

By comparing this model against the original 19-feature baseline, we can explicitly disentangle the performance gains attributed to *data enrichment* (the addition of lexical signals) versus those attributed to *methodological changes* (the use of ordinal loss functions in subsequent sections). This model minimizes the same MSE objective and uses the same rounding inference strategy, ensuring a direct comparison.

5.2.3 Ordinal Logistic Regression (Feature-CORAL)

To address the limitations of both nominal classification (which ignores order) and metric regression (which assumes equidistance), we implement a **Rank-Consistent Ordinal Logistic Regression** model. This approach is conceptually grounded in the *Consistent Rank Logits* (CORAL) framework proposed by [3].

While the term "Regression" is used because the target variable has a natural ordering, the underlying mechanics rely on a decomposition of the problem into a series of binary classification subtasks. By constraining these subtasks to share a common weight vector, the model learns a latent proficiency score while dynamically adjusting the decision boundaries between levels.

Task Decomposition: Rank-Monotonic Encoding

Standard classification encodes a label $y = k$ as a "one-hot" vector (e.g., $[0,0,1,0]$). However, this representation destroys the relationship between classes. Feature-CORAL instead employs a **rank-monotonic** (or "thermometer") encoding scheme.

For a dataset with K proficiency levels (where $K = 8$ for ACTFL levels 0–7), we transform the single integer label y_i into a binary vector $r_i \in \{0,1\}^{K-1}$. Each element $r_i^{(k)}$ represents the answer to

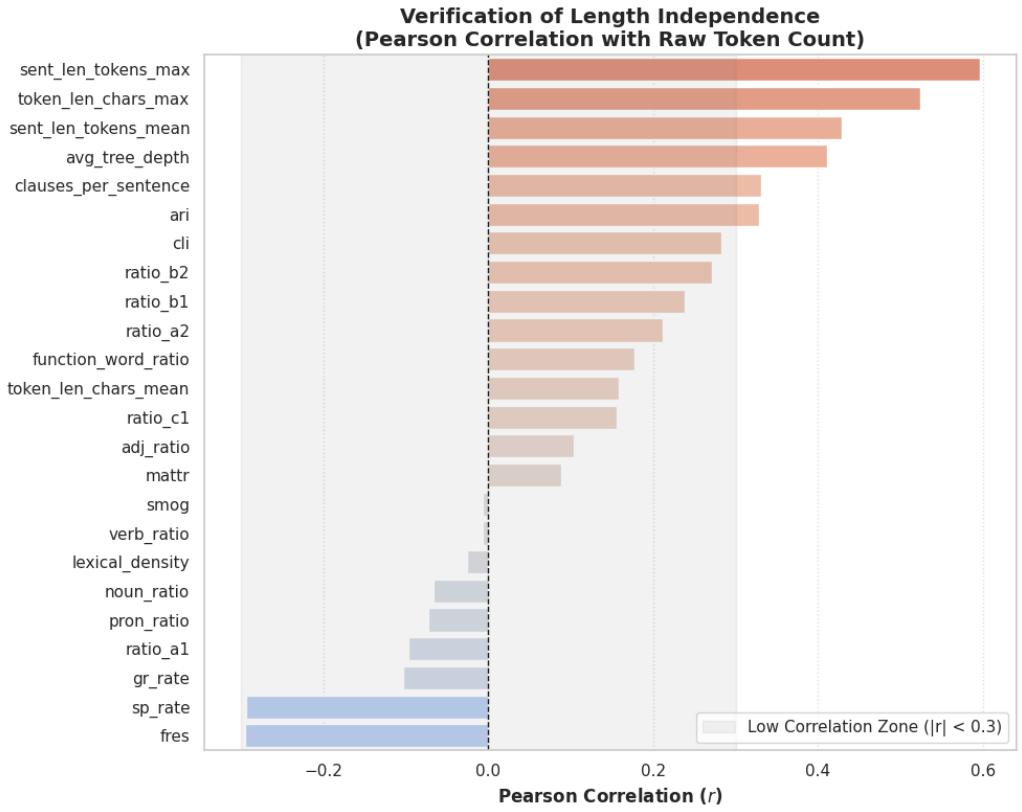

Figure 5.2: Pearson correlation between feature values and raw essay length. The lexical ratios (green bars) show weak correlation ($|r| < 0.3$), confirming their independence from text length.

the boolean question: *"Is the proficiency of essay i strictly greater than level k?"*

$$r_i^{(k)} = \begin{cases} 1 & \text{if } y_i > k \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } k \in \{0, 1, \ldots, K-2\} \tag{5.8}$$

For example, an *Intermediate Mid* essay (Label 4) is encoded not as a single category, but as passing the first four thresholds while failing the subsequent ones:

$$\text{Label } 4 \implies [1,1,1,1,0,0,0]$$

This encoding transforms the ordinal problem into $K-1$ simultaneous Bernoulli trials.

Model Architecture and The Consistency Guarantee

The critical innovation of CORAL, distinguishing it from simply training $K-1$ separate logistic regressions, is the **weight sharing constraint**.

We model the probability of exceeding rank k using a standard logistic sigmoid function $\sigma(\cdot)$. However, the weight vector w projecting the input features x onto the proficiency scale is *shared* across all tasks. Only the bias term b_k (the threshold) is specific to each task k:

$$P(y_i > k | x_i) = \sigma(z_k) = \sigma(w^T x_i + b_k) \tag{5.9}$$

Geometric Interpretation: Parallel Hyperplanes. Geometrically, the term $S = w^T x_i$ represents the essay's "Latent Proficiency Score"—a scalar value indicating overall writing quality derived from the linguistic features. The bias terms $\{b_0, \ldots, b_{K-2}\}$ act as cut-points along this proficiency axis.

Because w is fixed, the decision boundaries for all ranks are defined by the hyperplanes $w^T x + b_k = 0$. Since these hyperplanes share the same normal vector w, they are strictly **parallel**. This geometric constraint provides a theoretical guarantee of *Rank Monotonicity*. Since the sigmoid function is monotonically increasing, and the input score $w^T x$ is identical for all tasks, the relative ordering of probabilities depends solely on the biases. As the model trains, it tends to learn ordered biases ($b_0 > b_1 > \ldots$), encouraging that:

$$P(y > 0|x) \geq P(y > 1|x) \geq \cdots \geq P(y > K-2|x)$$

This prevents the logical inconsistencies often seen in independent classifiers (e.g., a model predicting an essay is "Standard" but also "Superior").

Optimization Objective

The model is trained to minimize the aggregate error across all binary subtasks. The loss function $\mathscr{L}$ is defined as the sum of binary cross-entropies for each of the $K-1$ thresholds:

$$\mathscr{L} = -\sum_{i=1}^{N} \sum_{k=0}^{K-2} \lambda^{(k)} \left[r_i^{(k)} \log(\hat{p}_{i,k}) + (1 - r_i^{(k)}) \log(1 - \hat{p}_{i,k}) \right] \tag{5.10}$$

where $\hat{p}_{i,k} = \sigma(w^T x_i + b_k)$. While the importance weights $\lambda^{(k)}$ can be tuned to emphasize specific boundaries, we set $\lambda^{(k)} = 1$ for all k, treating all proficiency transitions as equally important during optimization.

This formulation provides a distinct advantage over the Metric Baseline (Linear Regression). In Linear Regression, the distance between thresholds is fixed (the distance between 1 and 2 is equal to 2 and 3). In Feature-CORAL, the bias parameters $\{b_k\}$ are learnable. This allows the model to adaptively learn that the "linguistic distance" between *Advanced* and *Superior* is much larger than between *Novice* and *Intermediate*, approximating the non-linear psychometric structure of the ACTFL scale directly from the data.

Inference

To obtain a final predicted integer label $\hat{y}_i$ during testing, we sum the predicted probabilities for all binary tasks. Because the probabilities represent the likelihood of exceeding each successive threshold, their sum approximates the expected rank:

$$\hat{y}_i = \text{round}\left(\sum_{k=0}^{K-2} P(y_i > k|x_i) \right) \tag{5.11}$$

This summation method is robust; even if the probabilities near the decision boundary are uncertain (e.g., $0.6, 0.4$), the aggregation yields a stable estimate of the proficiency level.

5.2.4 Conditional Ordinal Regression (Feature-CORN)

While Feature-CORAL ensures rank consistency through a rigid weight-sharing constraint, this architecture theoretically limits the model's expressiveness. By forcing the weight vector w to be identical for all thresholds, CORAL assumes that the direction of proficiency is constant across the entire scale—implying that the linguistic features distinguishing *Novice* from *Intermediate* are identical (and weighted identically) to those distinguishing *Advanced* from *Superior*.

To relax this assumption while maintaining ordinal consistency, we implement the **Conditional Ordinal Regression (CORN)** framework proposed by [21]. Unlike CORAL, CORN learns distinct weight vectors for each threshold, deriving consistency not from geometry, but from the chain rule of probability.

Probabilistic Decomposition: The Chain Rule

Instead of asking independent boolean questions (as in nominal classification) or parallel boolean questions (as in CORAL), CORN models the *conditional* probability of exceeding a rank.

We decompose the ordinal target into a sequence of nested binary tasks. For a specific level k, the classifier asks: *"Given that this essay has already achieved at least level k, what is the probability that it also exceeds level k?"*

Formally, let C_k be the event that the true label $y > k$. The unconditional probability of exceeding rank k is computed using the chain rule of probability:

$$P(y > k|x) = \prod_{j=0}^{k} P(y > j|y > j-1, x) \tag{5.12}$$

(with the base case $P(y > -1) = 1$).

This formulation effectively turns the proficiency scale into a "tournament" or a "ladder." To reach the *Superior* level, an essay must successively pass the conditional tests for all preceding levels.

Architecture: Independent Conditional Heads

Because the tasks are framed conditionally, we are no longer bound to share weights. We define $K-1$ independent binary classifiers (logistic regressions), each with its own weight vector w_k and bias b_k:

$$P(y > k|y > k-1, x) = \sigma(w_k^T x + b_k) \tag{5.13}$$

This independence is the primary theoretical advantage of Feature-CORN over Feature-CORAL. It allows the model to learn **rank-specific feature importance**. For example, the classifier at the lower end of the scale ($k = 0$) might assign high positive weights to basic sentence length and token counts, while the classifier at the high end ($k = 6$) might ignore length entirely and focus exclusively on lexical sophistication ratios (e.g., ratio_C1) or subordination indices.

Optimization: Conditional Training Subsets

To train these conditional heads, CORN utilizes a data subsetting strategy. The classifier for level k is trained *only* on the subset of data that is relevant to that specific decision boundary.

Let $\mathscr{D}$ be the full training set. For each task k, we construct a conditional training set $\mathscr{D}_k$:

$$\mathscr{D}_k = \{(x_i, y_i) \in \mathscr{D} \mid y_i > k-1\} \tag{5.14}$$

The binary target $t_i^{(k)}$ for this subset is 1 if $y_i > k$ and 0 otherwise.

This approach inherently handles the **imbalance** problem found in essay scoring datasets.

- When training the distinction between *Advanced* and *Superior* ($k = 6$), the model explicitly excludes all *Novice* and *Intermediate* essays (where $y < 6$).
- This prevents the vast majority of low-scoring essays from overwhelming the gradient for the high-level classifiers, allowing the model to focus on the subtle differences between the top ranks.

The total loss is the sum of binary cross-entropies over these conditional subsets:

$$\mathscr{L} = -\sum_{k=0}^{K-2} \sum_{(x_i,y_i)\in\mathscr{D}_k} \left[t_i^{(k)} \log \hat{p}_{i,k} + (1 - t_i^{(k)}) \log(1 - \hat{p}_{i,k}) \right] \tag{5.15}$$

Inference

During inference, a test input x is passed through all $K-1$ independent classifiers. We reconstruct the unconditional cumulative probabilities using the product rule derived in Eq. (16):

$$P(y > k) = P(y > 0 | y > -1) \times P(y > 1 | y > 0) \times \cdots \times P(y > k | y > k-1) \tag{5.16}$$

Once the cumulative probabilities $P(y > k)$ are reconstructed for all k, we employ the same robust summation strategy used in CORAL to derive the final integer prediction:

$$\hat{y}_i = \text{round}\left(\sum_{k=0}^{K-2} P(y_i > k) \right) \tag{5.17}$$

By combining the flexibility of independent weights with the theoretical rigor of conditional probability, Feature-CORN aims to maximize performance on the imbalanced tails of the proficiency distribution.

Theoretical Trade-off: Flexibility vs. Efficiency

While CORN resolves the consistency issue via the chain rule rather than weight constraints, this flexibility comes at a cost in terms of parameter efficiency. In Feature-CORAL, the single shared weight vector w is updated by gradients from every essay in the dataset, providing a strong regularization effect that is beneficial for smaller datasets.

In contrast, Feature-CORN effectively splits the data into nested subsets. The classifier responsible for the highest-level distinction (e.g., *Advanced High* vs. *Superior*) is trained on a significantly smaller partition of the data. Consequently, while CORN has lower bias (it can model level-specific features), it potentially suffers from higher variance and a greater risk of overfitting on the sparsely populated tails of the proficiency distribution.

5.2.5 Summary of Feature-Based Approaches

The progression from standard linear regression to advanced ordinal architectures (CORAL and CORN) on the augmented feature set provides the foundation for the deep learning experiments described in the next section. Our methodological choices up to this point reveal three critical hypotheses regarding the modeling of L2 proficiency.

First, the theoretical superiority of ordinal loss functions. By replacing the rigid equidistance assumption of linear regression with the learnable thresholds of Feature-CORAL and Feature-CORN, we aim to minimize absolute error (MAE) and better capture the non-uniform psychometric distances between ACTFL levels. As detailed in Chapter 7, these ordinal architectures are expected to yield decision boundaries that align more closely with human rater intuition than standard metric baselines.

Second, the stability-plasticity trade-off. While Feature-CORN theoretically offers greater expressivity by uncoupling the weight vectors for each rank, this flexibility imposes a higher demand on data quantity per level. Given the dataset size ($N \approx 1,100$), we anticipate that the shared-weight constraint of Feature-CORAL may offer a necessary regularization effect, potentially yielding more robust generalization across cross-validation folds compared to the parameter-heavy CORN architecture.

Third, the limitations of engineered features. Despite the methodological sophistication of ordinal loss functions, feature-based models are inherently bounded by the information capacity of the input vector. While our 24 metrics capture distinct morphological and lexical signals, they lack the contextual awareness required to detect semantic coherence, argumentation quality, or idiomatic usage—nuances that differentiate high-proficiency writing.

We hypothesize that these handcrafted features will eventually reach a saturation plateau, where further methodological refinements (e.g., swapping regression for classification) yield diminishing returns. This limitation motivates the transition to the second half of our modeling strategy: the use of Deep Contextual Embeddings (XLM-RoBERTa), described in the following section.

5.3 Transformer-Based Models

While the feature-based approaches described in the previous sections offer high interpretability and stability, they are fundamentally limited by the information capacity of the engineered metrics. A count of "subordinate clauses" can quantify syntactic complexity, but it cannot capture semantic coherence, argumentation quality, or the subtleties of idiomatic usage that distinguish *Advanced* from *Superior* proficiency.

To address this gap, we transition to the second paradigm of our modeling strategy: **deep representation learning**. Instead of manually defining linguistic features, we employ a pre-trained Transformer architecture to learn a high-dimensional, contextual representation of the essay text directly from raw tokens.

5.3.1 From Attention to Contextual Embeddings

The foundation of modern NLP lies in the Transformer architecture introduced by Vaswani et al. [24] ("Attention Is All You Need"). Unlike previous recurrent models (RNNs/LSTMs) that processed text sequentially, Transformers utilize a *Self-Attention* mechanism (Query, Key, Value matrices) to process the entire sequence in parallel. This allows every token in an essay to attend to every other token, capturing long-range dependencies and context regardless of distance.

Building on this, BERT (Bidirectional Encoder Representations from Transformers) [7] introduced the concept of pre-training on massive corpora using a *Masked Language Modeling* (MLM) objective. The model learns to predict hidden tokens based on their context.

For document classification tasks like Automated Essay Scoring, we rely on a design choice inherent to BERT-like models: the **Classification Token ([CLS])**. In a standard Transformer encoder, every input token corresponds to a vector in the output layer. However, the architecture reserves a special token at the start of the sequence (denoted as `<s>` or `[CLS]`). This token does not carry intrinsic semantic meaning (like the word "apple" or "run" would); instead, through the layers of self-attention, it aggregates information from all other tokens in the sequence. By the final layer, the embedding of this token serves as a single vector representation of the entire text, effectively summarizing the essay's semantic and syntactic content into a fixed-size vector $h \in \mathbb{R}^d$.

This vector h_{CLS} serves as the input to our downstream proficiency scoring heads (Linear Regression, Classification, or Ordinal Regression), allowing us to leverage a powerful pre-trained encoder as a feature extractor.

5.3.2 Backbone Architecture: XLM-RoBERTa

Given the target language (Russian) and the limited size of our annotated dataset ($N \approx 1,100$), training a deep neural network from scratch is infeasible. We therefore rely on the *Transfer Learning* paradigm, fine-tuning a model that has already learned the structure of the Russian language from massive unlabelled corpora.

We select **XLM-RoBERTa (XLM-R)** [5] as our backbone encoder. XLM-R is a multilingual

variant of the Robustly Optimized BERT Pretraining Approach (RoBERTa). While not the absolute largest model in existence, it represents a widely adopted, open-source state-of-the-art baseline that balances performance with computational efficiency. Furthermore, its multilingual nature ensures that our scoring pipeline remains extensible to other L2 languages in the future with minimal architectural changes.

We specifically choose XLM-R over the older multilingual BERT (mBERT) for two critical reasons:

- **Larger Vocabulary & SentencePiece:** XLM-R uses a vocabulary of 250k subwords (compared to 110k in mBERT). This is particularly advantageous for Russian, a highly inflected language where a single root can produce dozens of morphological variants. XLM-R employs the SentencePiece tokenizer, which effectively breaks these complex words into constituent sub-lexical units (e.g., separating case endings from roots), allowing the model to generalize across word forms it has never seen before.
- **Training Objective:** Unlike the original BERT, which used a Next Sentence Prediction (NSP) task (predicting if two sentences follow each other), XLM-R is optimized purely on *Masked Language Modeling* (MLM). Research has shown that removing the NSP objective and training on longer sequences with dynamic masking yields more robust sentence-level representations for downstream classification tasks.

Input Representation and Pooling

We utilize the `xlm-roberta-base` configuration, which consists of 12 Transformer layers, 12 attention heads, and a hidden dimension size of $d = 768$.

The input essay is first tokenized into subwords and truncated or padded to a fixed maximum length of $L = 512$ tokens. The input sequence x is constructed as:

$$x_{input} = [\texttt{<s>}, w_1, w_2, \ldots, w_N, \texttt{</s>}] \tag{5.18}$$

where `<s>` is the special beginning-of-sequence token and `</s>` is the end-of-sequence token.

This sequence is passed through the 12 Transformer layers. To obtain the final representation for the essay, we perform a pooling operation by extracting the contextual embedding of the `<s>` token from the output of the final layer ($H_{last} \in \mathbb{R}^{L \times 768}$):

$$h_{CLS} = \text{Pool}(H_{last}) = H_{last}[0] \in \mathbb{R}^{768} \tag{5.19}$$

This vector h_{CLS} contains the aggregate semantic information of the essay. To mitigate overfitting during fine-tuning on our small dataset, we apply strong Dropout ($p = 0.1$) to this embedding before passing it to the specific prediction heads described in the following sections.

5.3.3 Deep Ordinal Heads

To rigorously test the "Ordinal Advantage" hypothesis in the deep learning setting, we attach four distinct scoring architectures to the output of the XLM-R encoder. While the backbone encoder ($h_{CLS} \in \mathbb{R}^{768}$) remains constant across all experiments, the choice of the prediction head determines how the model interprets the geometry of the proficiency space.

Mathematically, this transition from feature-based to deep learning models involves replacing the static, interpretable feature vector $x_{feats} \in \mathbb{R}^{24}$ with the learned, high-dimensional deep embedding h_{CLS}. The following subsections detail how each head transforms this latent representation into a final proficiency score.

Deep Metric Baseline (Linear Regression)

The simplest approach to Deep AES is to treat the proficiency score as a continuous variable. We attach a single linear layer to the Transformer output:

$$\hat{y} = w^T h_{CLS} + b \tag{5.20}$$

where $w \in \mathbb{R}^{768}$ is a learnable weight vector and $b \in \mathbb{R}$ is a bias term. The model is trained to minimize the Mean Squared Error (MSE) loss:

$$\mathscr{L}_{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{5.21}$$

During inference, the continuous output $\hat{y}$ is rounded to the nearest integer and clipped to the valid range $[0,7]$.

Theoretical Implication: This architecture assumes that the semantic distance between any two adjacent proficiency levels is identical (e.g., the linguistic "distance" from Novice Low to Novice Mid is equal to the distance from Advanced High to Superior). As discussed in Section 5.1.1, this assumption of equidistance rarely holds in second language acquisition, potentially limiting the model's ability to model non-linear jumps in proficiency.

Deep Nominal Baseline (Cross-Entropy)

This is the standard architecture for text classification tasks using BERT-like models. The 768-dimensional embedding is projected into $K = 8$ independent logits, one for each proficiency class:

$$z = W h_{CLS} + b \tag{5.22}$$

where $W \in \mathbb{R}^{8 \times 768}$ and $b \in \mathbb{R}^8$. These logits are normalized via the Softmax function to produce a probability distribution over the classes:

$$P(y = k | h_{CLS}) = \frac{e^{z_k}}{\sum_{j=0}^{K-1} e^{z_j}} \tag{5.23}$$

The model is optimized using the Cross-Entropy loss.

Theoretical Implication: While powerful, the nominal baseline suffers from the "disjoint class" problem. It treats the proficiency levels as independent categories, ignoring their inherent order. To the loss function, a prediction of *Novice Low* (Class 0) for a *Superior* (Class 7) essay is penalized no differently than a prediction of *Advanced High* (Class 6). This forfeits valuable ordinal information that could guide the encoder during training.

Deep CORAL (Rank-Consistent Ordinal Regression)

To reintroduce ordinality without the rigid assumptions of linear regression, we attach the CORAL head to the Transformer. Unlike the nominal model which uses a matrix W, Deep CORAL projects the embedding h_{CLS} onto a single shared weight vector $w \in \mathbb{R}^{768}$, effectively collapsing the high-dimensional representation onto a single "proficiency axis."

The model learns $K-1$ distinct bias units $\{b_1, b_2, \ldots, b_{K-1}\}$ which serve as the thresholds (cut-offs) between proficiency levels along this axis. The probability that an essay exceeds rank k is given by:

$$P(y > k | h_{CLS}) = \sigma(w^T h_{CLS} + b_k) \tag{5.24}$$

where σ is the sigmoid function.

Theoretical Implication: Deep CORAL acts as a regularizer. By forcing all binary decisions (e.g., "Is it > Novice?" and "Is it > Advanced?") to share the same weight vector w, the model forces the Transformer to organize the 768-dimensional latent space such that essay proficiency aligns along a single linear direction. This is particularly beneficial in low-resource settings, as the parameters are shared across all ranks, reducing the risk of overfitting.

Deep CORN (Conditional Ordinal Regression)

Finally, we implement the Deep CORN architecture, which represents the most flexible ordinal approach. Unlike CORAL, which enforces a single proficiency axis, CORN constructs an ensemble of $K-1$ independent binary classifiers on top of the shared encoder.

For each rank k, the model learns a unique projection vector $w_k \in \mathbb{R}^{768}$:

$$P(y > k | y > k-1, h_{CLS}) = \sigma(w_k^T h_{CLS} + b_k) \tag{5.25}$$

Critically, this conditional probability formulation allows the model to utilize different subspaces of the 768-dimensional embedding for different difficulty levels.

Theoretical Implication: This architecture hypothesizes that the linguistic features distinguishing low-level essays are fundamentally different from those distinguishing high-level essays. For instance, the transition from *Novice* to *Intermediate* might be driven by vocabulary size (ex. dimension d_{12} of the embedding), while the transition from *Advanced* to *Superior* might be driven by discourse coherence (ex. dimension d_{345}).

During training, we employ the conditional data subsetting strategy. The binary head for rank k is trained *only* on essays that are known to be at least rank $k-1$. This means that the gradients

backpropagated into the Transformer backbone for the “Advanced vs. Superior” distinction are derived exclusively from high-proficiency data, effectively creating a specialized "Curriculum Learning" effect within the neural network.

5.4 Hybrid Approaches

While the feature-based and Transformer-based paradigms represent distinct methodological approaches, they are not mutually exclusive. In fact, we hypothesize that they capture complementary aspects of linguistic proficiency.

Transformer models like XLM-R excel at capturing semantic coherence, topical relevance, and complex contextual dependencies—features that are difficult to encode manually. However, due to their subword tokenization strategies (SentencePiece) and Masked Language Modeling objectives, these models are explicitly designed to be *robust* to noise. In the context of language assessment, this robustness is a double-edged sword: the model may successfully “understand” a sentence full of spelling errors and grammatical slips, thereby failing to penalize the candidate for mechanical inaccuracies that are critical for determining lower CEFR levels.

Conversely, our engineered features (Chapter 4) are explicitly designed to capture these mechanical and syntactic dimensions (e.g., lexical diversity ratios, error densities, readability scores). While they lack semantic “understanding” (low recall), they should offer high precision in quantifying structural quality.

Selection Strategy: To maximize the effectiveness of the hybrid approach, we do not test every possible combination of models. Instead, we adopt a selective strategy: we integrate the best-performing feature set with the best-performing Transformer architecture, as determined by the validation results in Sections 7.1 ML feature-based baselines, and 7.2 Transformer-based classification results.

This ensures that our fusion mechanisms are applied to the strongest possible baselines.

To leverage the strengths of both paradigms, we investigate two hybrid strategies: **Early Fusion** (Feature Concatenation) and **Late Fusion** (Ensembling).

5.4.1 Feature Fusion (Early Fusion)

In the Early Fusion approach, we inject the linguistic knowledge encoded in our manual features directly into the deep learning pipeline, allowing the prediction head to learn non-linear interactions between the semantic embedding and the explicit metrics.

Mathematically, we extract the context vector $h_{CLS} \in \mathbb{R}^{768}$ from the final layer of the fine-tuned XLM-R model. Simultaneously, we compute the vector of handcrafted features $x_{feats} \in \mathbb{R}^{24}$ for the same text. Critical to this approach is **Z-score normalization**: since the Transformer embeddings typically follow a unit-normal distribution (due to LayerNorm operations), we must standardize x_{feats} to have zero mean and unit variance to prevent numerical instability or gradient dominance during training.

We construct a composite representation h_{hybrid} via concatenation:

$$h_{hybrid} = h_{CLS} \oplus x_{normalized} \in \mathbb{R}^{768+24} \tag{5.26}$$

This extended vector of dimension $d = 792$ serves as the input to our scoring heads (Nominal, CORAL, or CORN). This architecture allows the model to utilize the deep semantic features while "attending" to explicit signals like text length or error rate when the semantic signal is ambiguous.

5.4.2 Ensembling (Late Fusion)

The Late Fusion strategy treats the best-performing Feature-Based Ordinal model (e.g., Feature-CORAL) and the best Deep Learning model (e.g., Deep CORAL) as independent experts. We assume that the errors produced by these two distinct model classes are essentially uncorrelated—that is, the instances where the engineered linguistic features fail are likely different from the instances where the contextual embedding fails.

We perform a weighted linear combination of their predictions. Let $\hat{y}_{DL}$ be the continuous score predicted by the XLM-R model and $\hat{y}_{Feat}$ be the score predicted by the Feature-Based Ordinal model. The final ensemble prediction $\hat{y}_{ensemble}$ is given by:

$$\hat{y}_{ensemble} = \alpha \cdot \hat{y}_{DL} + (1 - \alpha) \cdot \hat{y}_{Feat} \tag{5.27}$$

where $\alpha \in [0, 1]$ is a learnable scalar weight.

Unlike Early Fusion, which requires joint training (and thus risks the deep network overfitting the manual features or ignoring them entirely), Late Fusion optimizes the combination parameter α on a held-out validation set after the individual models have converged. This ensures that we retain the global optimum of the Transformer while nudging its predictions using the rigid, rule-based perspective of the engineered ordinal model.

5.5 Summary of Experimental Design

To ensure a comprehensive evaluation of the proposed methodology, we structured our experiments as a progressive hierarchy of models. This design allows us to isolate the contribution of each component—input representation (Engineered Features vs. Deep Embeddings) and learning objective (Metric vs. Nominal vs. Ordinal)—to the final grading performance.

Our experimental suite is divided into three distinct families:

1. **Shallow Baselines (Feature-Based):** These models assess the limits of explicit linguistic engineering. They test whether traditional complexity metrics (Chapter 4) are sufficient to capture proficiency when paired with standard machine learning algorithms. We evaluate the full spectrum of loss functions here, from Linear Regression (Metric) to CORAL/CORN (Ordinal).

2. **Deep Baselines (Transformer-Based):** These models evaluate the power of latent semantic representations. By keeping the encoder constant (XLM-R) and swapping the prediction head,

we isolate the impact of the loss function in the deep learning paradigm. This allows us to directly compare, for example, a Deep Linear Regressor against a Deep Ordinal Network.

3. **Hybrid Systems:** Finally, we evaluate whether explicit linguistic features and deep latent representations are complementary. We test this via both architecture-level fusion (concatenation) and prediction-level fusion (ensembling).

Table 5.1 details the complete list of model configurations developed and evaluated in this thesis.

Model Paradigm	**Specific Architecture**	**Input Data**	**Loss / Objective**
Zero-Rule	Dummy Classifier	N/A	Majority Class / Mean / Median
Feature-Based	Metric Baseline (LinReg)	Engineered Features ($x \in \mathbb{R}^{24}$)	MSE (Regression)
	Nominal Baseline (LogReg)	Engineered Features ($x \in \mathbb{R}^{24}$)	Cross-Entropy (Classif.)
	Feature-CORAL	Engineered Features ($x \in \mathbb{R}^{24}$)	CORAL Loss (Ordinal)
	Feature-CORN	Engineered Features ($x \in \mathbb{R}^{24}$)	CORN Loss (Conditional)
Deep Learning	Deep Metric (XLM-R)	Raw Tokens ($L = 512$)	MSE (Regression)
	Deep Nominal (XLM-R)	Raw Tokens ($L = 512$)	Cross-Entropy (Classif.)
	Deep CORAL (XLM-R)	Raw Tokens ($L = 512$)	CORAL Loss (Ordinal)
	Deep CORN (XLM-R)	Raw Tokens ($L = 512$)	CORN Loss (Conditional)
Hybrid	Early Fusion (Concatenation)	Tokens $\oplus$ Features	CORAL Loss
	Late Fusion (Ensemble)	Model Predictions ($\hat{y}_1, \hat{y}_2$)	Weighted Linear Combination

Table 5.1: Complete catalog of experimental configurations. The study progresses from interpretable feature-based models to high-capacity deep ordinal networks and hybrid ensembles.

All models were evaluated using the same 5-Fold Stratified Cross-Validation scheme to ensure statistical robustness. For the Deep Learning and Hybrid models, we employed a standardized training regimen (AdamW optimizer, 2×10^{-5} learning rate, early stopping based on QWK) to ensure that performance differences are attributable to the architecture rather than hyperparameter discrepancies.

Chapter 6

Training and Evaluation

Having defined the model architectures in Chapter 5, this chapter details the experimental protocols employed to train and evaluate them. A rigorous evaluation framework is critical for Automated Essay Scoring (AES), particularly when comparing disparate paradigms such as feature-based engineering and deep representation learning.

This chapter specifies the computational environment, the hyperparameter configurations for both shallow and deep models, and the specific loss functions used to enforce ordinal constraints. Furthermore, it provides a formal derivation of the evaluation metrics selected to assess proficiency scoring, with a particular focus on the Quadratic Weighted Kappa (QWK), the standard metric for the field.

6.1 Training Setup and Hyperparameters

To ensure the reproducibility of our results and fair comparison between model families, we established a standardized training environment. All experiments were implemented using the `PyTorch` deep learning framework and the Hugging Face `transformers` library for the neural components, while `scikit-learn` was utilized for statistical baselines.

6.1.1 Computational Environment

The computationally intensive experiments (specifically the fine-tuning of XLM-RoBERTa) were conducted on high-performance GPUs (NVIDIA Tesla T4) via the Google Colab environment. To mitigate non-deterministic behavior inherent in GPU training, we enforced strict reproducibility protocols by fixing the random seeds for the CPU generator (`numpy`, `random`) and the GPU generator (`torch.cuda`) to a constant value ($SEED = 42$) at the initialization of every cross-validation fold.

6.1.2 Hyperparameter Configuration

Hyperparameters were selected based on empirical preliminary searches and standard configurations for low-resource fine-tuning.

Feature-Based Models

For the statistical baselines described in Sections 5.1 and 5.2, we focused on regularized linear models to prevent overfitting on the limited dataset ($N \approx 1,100$).

- **Linear Regression (Metric Baseline):** We applied **Ridge Regression** (Linear Regression with L_2 regularization). We utilized the standard regularization strength ($\alpha = 1.0$) to constrain coefficient magnitude and mitigate multicollinearity among the linguistic features.
- **Feature-CORAL (Neural):** The feature-based ordinal model was trained as a shallow neural network using the **Adam** optimizer with a learning rate of 0.01 and a weight decay of 0.001 to ensure convergence on the 24-dimensional input vector.
- **Normalization:** As emphasized in Chapter 4, all input features were Z-score normalized (zero mean, unit variance) prior to training to ensure numerical stability.
- **Full Fine-tuning:** All parameters of the XLM-R encoder were updated during training (i.e., full fine-tuning), rather than freezing the backbone or employing parameter-efficient adaptation methods.

Transformer-Based Models (XLM-R)

For the Deep Learning experiments (Section 5.3), we adopted a fine-tuning strategy designed to preserve the pre-trained multilingual knowledge while adapting the model to the specific ordinal ranking task.

Optimization: We utilized the **AdamW** optimizer (Adam with Weight Decay), which serves as the standard for fine-tuning Transformer models.

- **Learning Rate:** We employed a conservative learning rate range of 1×10^{-5} to 2×10^{-5}, which is standard practice when fully fine-tuning Transformer models, in order to mitigate potential catastrophic forgetting of the pre-trained multilingual representations.
- **Scheduler:** A linear learning rate decay was applied with 0 warmup steps, gradually reducing the rate to 0 by the end of training.
- **Batch Size:** Due to GPU memory constraints, we utilized a batch size of 8.

Regularization & Early Stopping: Given the relatively small dataset size, overfitting was a primary concern. In addition to the inherent dropout ($p = 0.1$) within the XLM-R encoder layers, we applied weight decay ($\lambda = 0.01$) through the AdamW optimizer and employed Early Stopping based on validation QWK:

- **Patience:** Training terminates if the validation metric (QWK) does not improve for 3 consecutive epochs.
- **Max Epochs:** An upper limit of **6 epochs** was set, as the pre-trained models typically converged rapidly to optimal performance.

- **Best Model Recovery:** The trainer was configured to load the weights from the epoch that achieved the highest validation QWK, ensuring the final model is optimal for the target metric.

6.2 Loss Functions for Ordinal Classification

While the architecture (Chapter 5) defines the information flow, the loss function defines the optimization landscape. To test the "Ordinal Advantage" hypothesis, we compared four distinct loss objectives:

1. **Cross-Entropy (Nominal Baseline):** The standard negative log-likelihood loss. It treats ranks as independent categories, penalizing all errors equally regardless of distance.

$$\mathscr{L}_{CE} = -\sum_{c=0}^{K-1} y_c \log(p_c)$$

2. **Mean Squared Error (Metric Baseline):** Used for regression heads. It enforces a metric distance penalty but strictly assumes equidistance between proficiency levels.

$$\mathscr{L}_{MSE} = \frac{1}{N}\sum (y_{true} - y_{pred})^2$$

3. **CORAL Loss (Ordinal):** We utilized the implementation from the `coral-pytorch` library. The loss is computed as the summation of binary cross-entropies across the $K-1$ rank thresholds.

$$\mathscr{L}_{CORAL} = \sum_{k=1}^{K-1} \left[\log(\sigma(f_k(x))) \cdot r^{(k)} + \log(1 - \sigma(f_k(x))) \cdot (1 - r^{(k)}) \right]$$

   Crucially, this loss penalizes rank-inconsistency, forcing the cumulative probabilities to remain monotonic.

4. **CORN Loss (Conditional):** The Conditional Ordinal Regression loss aggregates the binary cross-entropy of conditional subsets S_k (where the true label $y > k-1$).

$$\mathscr{L}_{CORN} = -\sum_{k=1}^{K-1} \sum_{x_i \in S_k} \left[t_i^{(k)} \log(p_i^{(k)}) + (1 - t_i^{(k)}) \log(1 - p_i^{(k)}) \right]$$

   This allows the model to optimize specific decision boundaries (e.g., *Advanced* vs. *Superior*) independently, weighting features differently for each transition.

6.3 Evaluation Metrics

Evaluating AES systems requires metrics that align with human perception of grading quality. In a proficiency scale, errors are not binary; misclassifying a student by one level is a minor variance, while misclassifying them by four levels is a system failure. Therefore, we employ a suite of metrics with varying penalty structures.

6.3.1 Quadratic Weighted Kappa (QWK)

The primary evaluation metric for this thesis is Cohen's Kappa with quadratic weights (κ_w), often referred to as QWK. It is the *de facto* standard in the AES literature (used by ETS and in Kaggle competitions) because it penalizes disagreements quadratically based on the distance between ratings.

Formally, let an $N \times N$ matrix O (Observed) represent the confusion matrix where $O_{i,j}$ corresponds to the count of essays receiving human score i and system score j. Let matrix E (Expected) represent the expected agreement by chance.

An $N \times N$ weight matrix W is constructed to penalize the distance between scores:

$$W_{i,j} = \frac{(i-j)^2}{(N-1)^2} \tag{6.1}$$

where N is the number of classes (8 in our case). The QWK score is then computed as:

$$\kappa_w = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}} \tag{6.2}$$

Interpretation: A score of $\kappa_w = 1.0$ indicates perfect agreement, while 0.0 indicates agreement equivalent to random chance. QWK is particularly robust for our skewed dataset because it strictly penalizes large outliers, ensuring that the model does not achieve artificially high scores simply by guessing the majority class.

6.3.2 Regression Metrics (MAE, MSE, RMSE)

To assess the precision of the model's latent predictions in terms of "proficiency levels," we employ standard regression metrics.

Mean Absolute Error (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{6.3}$$

MAE is highly interpretable in the context of ACTFL levels. An MAE of 0.5 implies that, on average, the system's prediction deviates from the human rating by half a proficiency level.

Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{6.4}$$

RMSE assigns higher weight to large errors. A model with low MAE but high RMSE suggests that while it is generally accurate, it occasionally makes catastrophic errors (e.g., predicting Level 0 for a Level 7 essay).

6.3.3 Accuracy

We report standard Accuracy (percentage of exact matches) strictly for comparison with legacy systems. In ordinal quantification, Accuracy is often misleading; a model that predicts *Level* 3 for a *Level* 4 essay is useful, but Accuracy treats it as a total failure equivalent to predicting *Level* 0. Therefore, Accuracy is not used for model selection.

6.4 Continuous vs. Discrete Prediction Outputs

A methodological challenge arises because our ground truth labels are discrete integers $y \in \{0, \dots, 7\}$, while many of our models (Linear Regression, CORAL, Ensembles) produce continuous scalar outputs.

To map these continuous values to the ACTFL scale for evaluation (QWK calculation), we apply a clipping and rounding strategy:

$$\hat{y}_{final} = \text{round}(\min(\max(\hat{y}_{raw}, 0), 7)) \tag{6.5}$$

This ensures that predictions are valid integers within the bounds of the 8-class scale. However, for the purpose of **Error Analysis** (Chapter 7), we often retain the continuous raw logits to visualize the model's uncertainty near decision boundaries.

6.5 Model Selection Criteria

Given the multi-metric evaluation landscape, a single criterion must be selected to determine the "best" model during the hyperparameter tuning and checkpointing phases.

We strictly adhere to the following selection rule: **Models are selected based on the maximization of Validation QWK.**

During the 5-Fold Cross-Validation process, the checkpoint saved for each fold is the one that achieved the highest QWK score on the validation hold-out set, irrespective of the training loss. This distinction is crucial because a model might minimize Cross-Entropy loss by becoming very confident about the wrong class, whereas QWK directly optimizes for the ordinal agreement we seek. QWK itself is non-differentiable and unsuitable as a direct training objective; therefore, it is used strictly for validation-based model selection rather than gradient-based optimization. Consequently, all results reported in Chapter 7 represent the performance of the QWK-optimal checkpoints aggregated across the 5 folds.

6.5.1 Cross-Validation Protocol

All experiments were evaluated using stratified 5-fold cross-validation at the essay level. For each fold, the dataset was partitioned into training (80%) and validation (20%) splits, ensuring that the

distribution of ACTFL proficiency levels was preserved across folds. No essay appears in more than one split within a fold.

Chapter 7

Experiments and Results

This chapter presents the empirical findings of the study, systematically evaluating the performance of hand-crafted feature-based models, deep representation learning approaches, and hybrid fusion strategies for Automated Essay Scoring (AES) in Russian. We begin by establishing reference points using naive and statistical baselines, proceed to benchmarking feature-based and Transformer-based ordinal models, and conclude with an analysis of fusion techniques. Finally, we conduct a detailed error analysis to elucidate the linguistic behaviors underlying the observed performance differences.

7.1 Experimental Overview

Building on the architectures defined in Chapter 5 and the training protocols in Chapter 6, we evaluated a comprehensive hierarchy of modeling configurations. To ensure precise terminological distinction, we categorize our experiments into Reference Baselines (Naive and Statistical Baselines) and Experimental Models (Enhanced Feature-Based, Deep, and Hybrid).

Table 7.1 summarizes the key characteristics of each experimental condition.

The results are organized hierarchically: Section 7.2 reports the performance of the reference baselines; Section 7.3 evaluates the advanced feature-based models; Section 7.4 benchmarks the deep learning architectures; and Section 7.5 analyzes the impact of fusion strategies. Finally, Section 7.7 provides a granular error analysis to explain the linguistic drivers of model performance.

7.2 Reference Baselines

Before interpreting the performance of complex neural architectures, it is necessary to establish the lower bounds of performance using non-learning and simple statistical approaches.

7.2.1 Naive Baselines

To quantify the difficulty of the task, we deployed three "dummy" predictors introduced in Chapter 6. These models ignore the input text entirely and rely solely on the label distribution of the training set.

Table 7.2 summarizes their performance. As expected, all naive strategies yielded a QWK of approximately 0.0, confirming that the dataset contains non-trivial linguistic signal that cannot be

Category	Model Name	Input Representation	Objective Function
Naive Baselines	Dummy Mean	N/A	Predict Mean Score
	Dummy Median	N/A	Predict Median Score
	Dummy Mode	N/A	Predict Majority Class
Statistical Baselines	Linear Regression	19 Basic Features	MSE (Ridge)
	Logistic Regression	19 Basic Features	Cross-Entropy
Feature-Based Models	Linear Regression	24 Extended Features	MSE (Ridge)
	Feature-CORAL	24 Extended Features	CORAL (Ordinal)
	Feature-CORN	24 Extended Features	CORN (Conditional)
Deep Learning	XLM-R (Nominal)	Raw Tokens ($L = 512$)	Cross-Entropy
	XLM-R + CORAL	Raw Tokens ($L = 512$)	CORAL (Ordinal)
	XLM-R + CORN	Raw Tokens ($L = 512$)	CORN (Conditional)
Hybrid	Early Fusion	Tokens $\oplus$ Features	CORAL (Ordinal)
	Late Fusion	Ensemble Predictions	Weighted Linear Combination

Table 7.1: Recap of experimental configurations. We distinguish between Naive/Statistical reference baselines (using 19 basic features) and the primary experimental models (using the full 24-feature set or Deep Learning).

approximated by simple central tendency.

Model	Strategy	QWK	MAE
Dummy Mean	Predict μ_{train}	0.000	1.192
Dummy Median	Predict Med_{train}	0.000	1.292
Dummy Mode	Predict Majority Class	0.000	1.535

Table 7.2: Performance of Naive Baselines (5-Fold CV)

This result is structurally significant. A QWK of zero indicates that agreement with human raters is no better than random chance given the class marginals. Any model achieving a positive QWK is therefore demonstrating successfully learned signal extraction.

7.2.2 Statistical Baselines (19 Features)

Moving beyond naive guessing, we evaluated two standard statistical models trained on the core set of 19 linguistic features (excluding the experimental lexical profiling metrics). This step isolates the predictive power of the fundamental linguistic signals—such as syntactic depth and error density.

We compared two paradigms:

- **Linear Regression (Ridge):** Treats proficiency as a continuous variable, minimizing Mean Squared Error (MSE).
- **Logistic Regression (Nominal):** Treats proficiency levels as independent categories, minimizing Cross-Entropy loss.

Table 7.3 presents the performance of these models averaged across the 5 folds.

Analysis: The jump in performance compared to the naive baselines ($0.0 \rightarrow 0.74$) confirms that the core linguistic feature set contains strong, discriminative proficiency signals. Crucially, Linear

Model	QWK	MAE	RMSE
Logistic Regression (Classification)	0.7129	0.7302	1.1007
Linear Regression (Ridge)	**0.7443**	**0.7278**	**0.9522**

Table 7.3: Performance of Statistical Baselines (19 Features)

Regression outperforms Logistic Regression (+0.03 QWK). This highlights the importance of modeling the *order* of proficiency levels. By treating the label as a continuous value, the regression model heavily penalizes predictions that are far from the truth, naturally aligning with the QWK metric.

7.3 Feature-Based Models (24 Features)

Having established benchmarks on the core linguistic set (19 features), we proceeded to evaluate the full feature vector. As detailed in Chapter 4, this experimental condition augments the structural features with 5 Lexical Profiling features. These features quantify the ratio of vocabulary belonging to CEFR levels A1 through C1, capturing semantic sophistication.

7.3.1 Linear Regression: The Impact of Lexical Profiling

To assess whether adding lexical profiling features improves model precision, we first retrain the linear regression model using the full 24-feature set to assess the marginal contribution of the new lexical metrics.

- **Statistical Baseline (19 Features):** QWK = 0.7443, MAE = 0.7278
- **Extended Linear Model (24 Features):** QWK = 0.7596, MAE = 0.7077

Result: The addition of lexical profiling yielded a substantial performance boost (+0.015 QWK, -0.020 MAE). This confirms that vocabulary distribution provides orthogonal information to syntactic complexity, allowing even a simple linear model to reach a QWK of ≈ 0.76.

7.3.2 Ordinal Neural Models (CORAL vs. CORN)

To test if performance could be further improved by removing the linearity assumption, we trained neural networks using the CORAL (Consistent Rank Logits) and CORN (Conditional Ordinal Regression) objective functions on the same 24 standardized features.

Table 7.4 presents the comparative results.

Model	QWK	MAE	RMSE
Linear Regression (Ridge)	0.7596	0.7077	0.9191
Feature-CORAL	0.7591	**0.6678**	0.9800
Feature-CORN	**0.7614**	0.6722	0.9910

Table 7.4: Performance of Extended Feature-Based Models (24 Features)

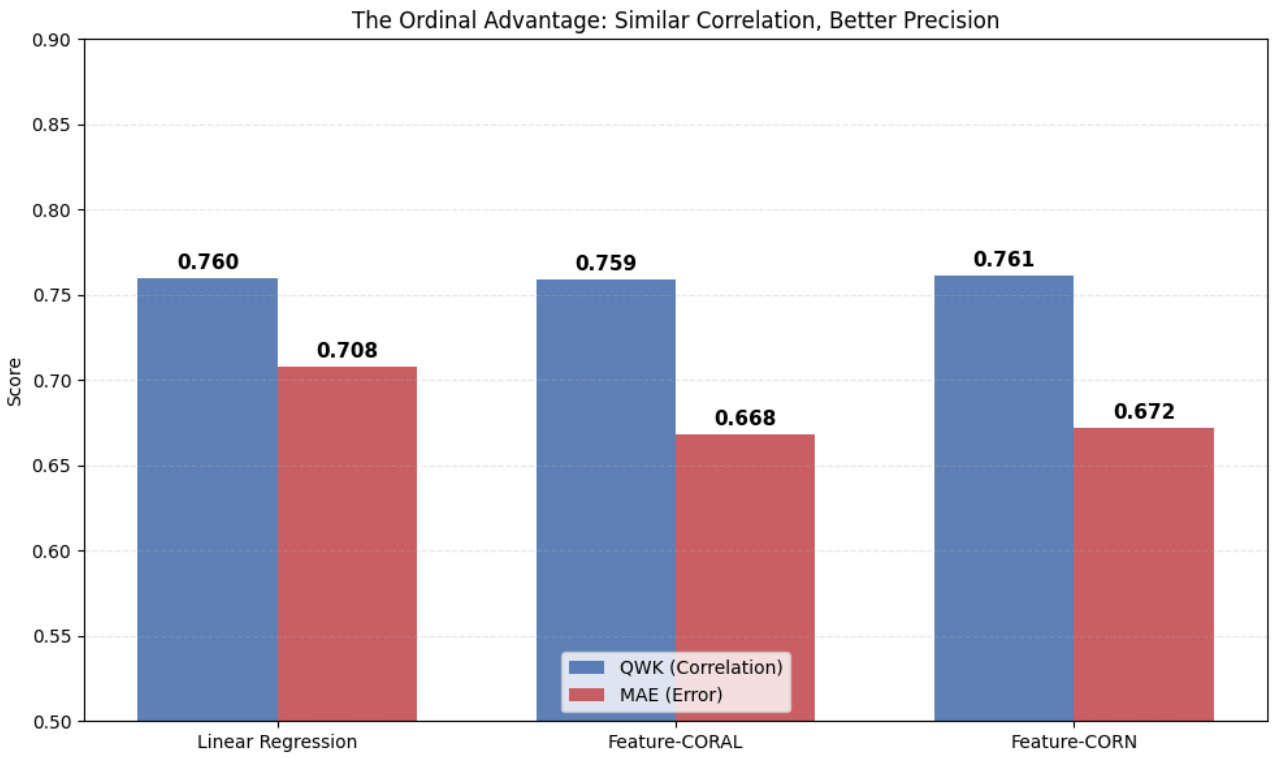

Figure 7.1: Comparative performance of linear vs. ordinal models. While QWK (rank correlation) is similar across all models, the ordinal architectures (CORAL/CORN) achieve substantially lower Mean Absolute Error, indicating higher precision in exact class prediction.

Analysis of the "Ordinal Advantage": In terms of rank correlation (QWK), the ordinal models performed comparably to the linear regression (≈ 0.76). However, a critical difference emerges in the **Mean Absolute Error (MAE)**. Both Feature-CORAL (0.668) and Feature-CORN (0.672) substantially outperformed the Linear Regression (0.708) in precision. This indicates that while the Linear model captures the global ranking trend, the Ordinal models are better at placing essays into their exact proficiency bins.

7.3.3 Threshold Analysis (The Equidistance Fallacy)

To understand why the ordinal models achieve lower error, we inspected the learned thresholds of the Feature-CORAL model. Unlike Linear Regression, which enforces equal spacing between grades, CORAL learns distinct cut-offs (b_k) on the latent proficiency scale.

As illustrated in Figure 7.2, the learned thresholds are highly non-uniform. By adjusting the "width" of each proficiency class, the model reduces the average prediction error by approximately 0.04 points, a non-trivial improvement on an 8-point scale.

7.3.4 Comparison: Shared Weights vs. Conditional Probabilities

Comparing the two ordinal approaches, Feature-CORN (0.761) marginally outperformed Feature-CORAL (0.759) in QWK. This aligns with our hypothesis that the conditional independence of CORN allows for more flexible decision boundaries on skewed datasets.

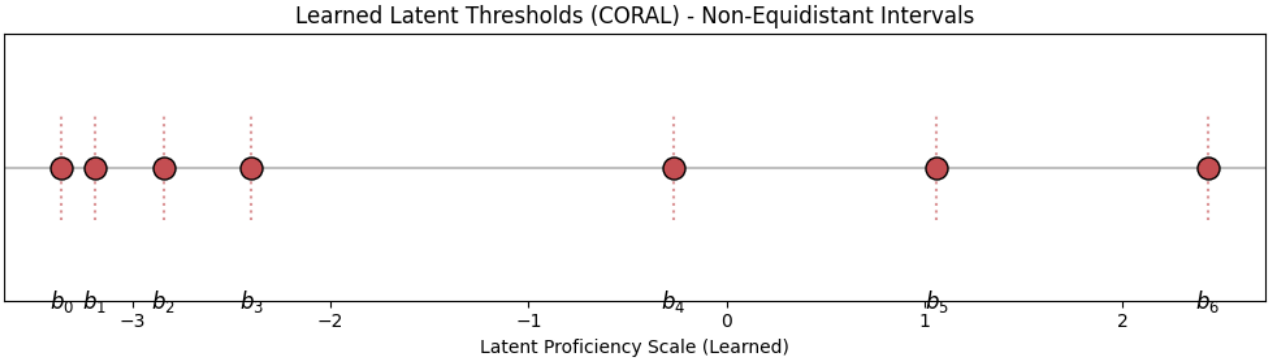

Figure 7.2: Visualization of learned class boundaries in Feature-CORAL. The non-uniform spacing between thresholds (e.g., the wider gap for Novice levels) confirms that the ACTFL proficiency scale is not linear in the feature space, explaining the superior MAE of the ordinal approach.

Figure 7.3 visualizes the weights of the independent binary classifiers in CORN. The heatmap reveals a "feature collage" rather than a uniform progression: distinct linguistic dimensions activate at different proficiency boundaries. For instance, orthographic control appears critical for the initial transition out of the Novice stage, whereas morphosyntactic accuracy becomes the primary differentiator at higher intermediate levels. This granular adaptability likely contributes to why CORN achieves the highest precision among feature-based models.

7.4 Transformer-Based Classification

To assess the contribution of learned contextual representations, we fine-tuned **XLM-RoBERTa-base** (XLM-R) using four distinct output heads: Linear Regression (MSE), Cross-Entropy (Nominal), Deep CORN, and Deep CORAL.

7.4.1 Performance Hierarchy

Table 7.5 summarizes the 5-fold cross-validation performance.

Model Head	QWK	MAE	RMSE	Accuracy
Linear Regression (MSE)	0.8317	0.6046	0.8873	0.4780
Deep CORN (Ordinal)	0.8314	0.5791	0.8624	0.4973
Cross-Entropy (Nominal)	0.8476	0.5317	0.8122	0.5298
Deep CORAL (Ordinal)	**0.8677**	**0.4956**	**0.7663**	**0.5466**

Table 7.5: Performance of XLM-R Models (5-Fold Cross-Validation)

7.4.2 Analysis of Results

The Ordinal Supremacy (CORAL)

The **Deep CORAL** architecture emerged as the definitive state-of-the-art model, achieving a QWK of **0.868** and a Mean Absolute Error (MAE) of **0.496**. This MAE result is particularly notable: it implies that on average, the model's prediction is within half a proficiency point of the human rater.

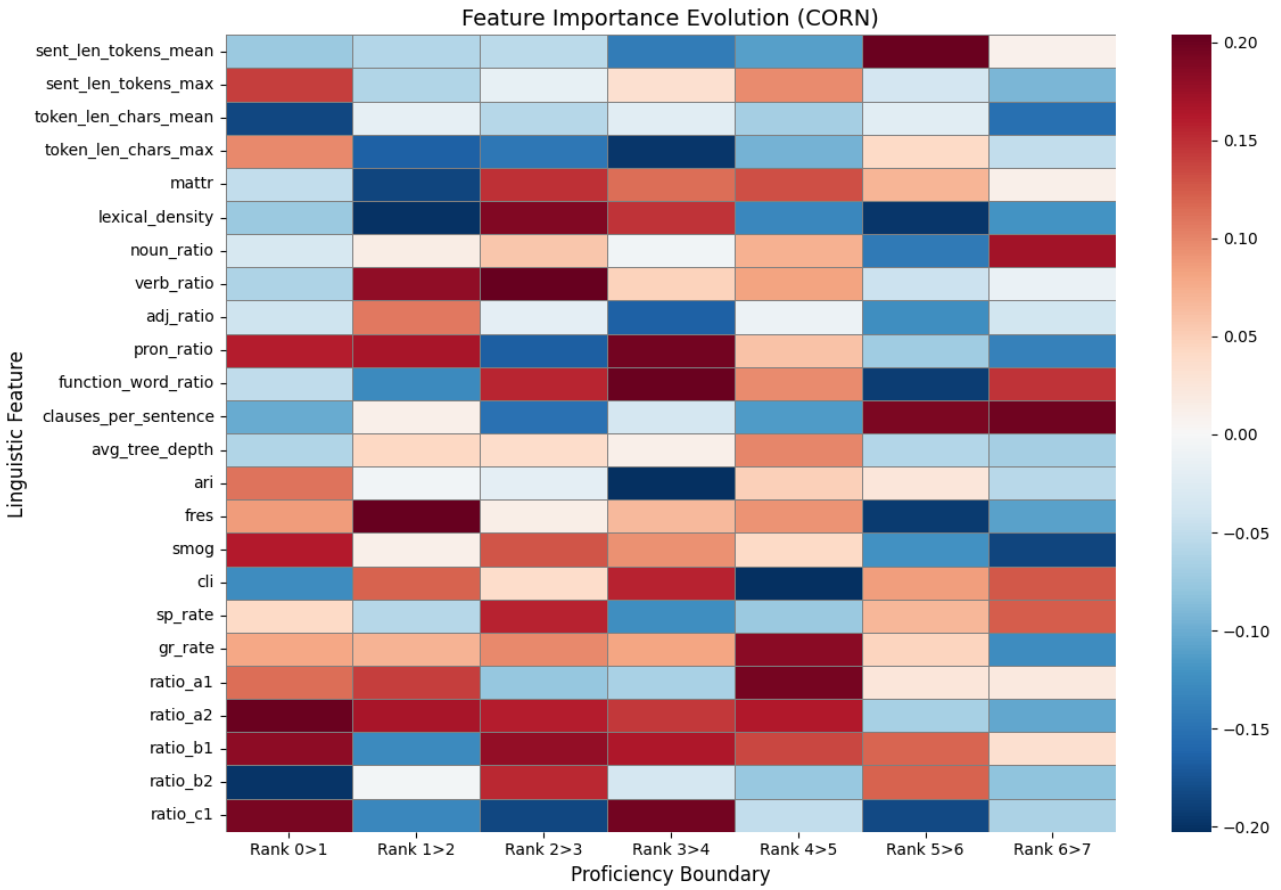

Figure 7.3: Feature Importance Evolution in Feature-CORN. Unlike CORAL, which enforces a single weight vector, CORN reveals a mosaic of feature dependencies. Note how 'Spelling Rate' (sp_rate) emerges as a critical discriminator at the Novice/Intermediate boundary (Rank 2→3), while 'Grammar Errors' (gr_rate) gain prominence at the Intermediate Mid/High boundary (Rank 4→5).

By enforcing a shared weight vector w across all proficiency thresholds (b_k), CORAL effectively combines the semantic power of the Transformer with the structural regularization of ordinal regression. This allows it to outperform the nominal Cross-Entropy baseline ($+0.02$ QWK, -0.036 MAE), proving that explicit modeling of rank order provides a tangible benefit even when using high-capacity deep encoders.

Nominal vs. Ordinal (Cross-Entropy vs. CORN)

Interestingly, the standard Cross-Entropy model (0.848) outperformed Deep CORN (0.831). This reinforces the "Stability-Plasticity" hypothesis:

- **Cross-Entropy** is a robust, low-variance objective. The XLM-R encoder is powerful enough to implicitly cluster adjacent proficiency levels, allowing a nominal classifier to perform well without explicit ordinal constraints.
- **Deep CORN**, which trains $K-1$ independent binary classifiers, appears to suffer from data fragmentation on this dataset ($N \approx 1,100$). Without the shared-weight constraint of CORAL, the independent classifiers for rare classes (e.g., Level 0 or 7) likely failed to generalize as effectively as the monolithic Cross-Entropy head.

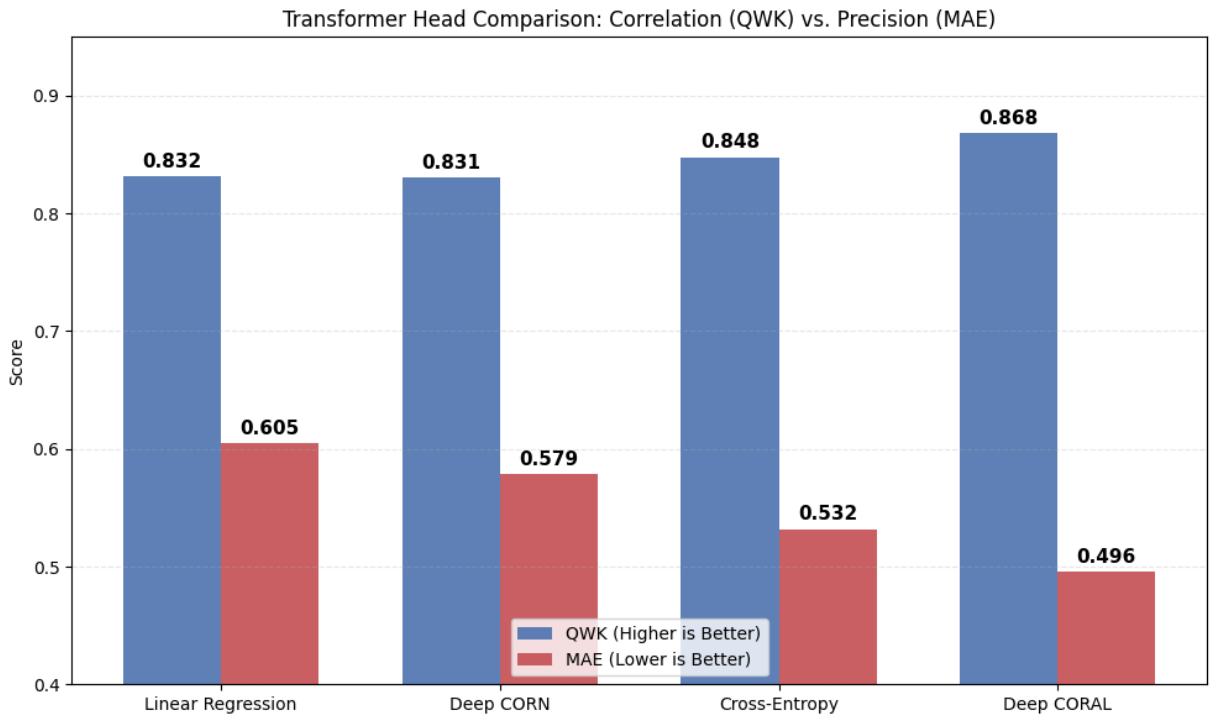

Figure 7.4: Comparative performance of Transformer heads. Deep CORAL simultaneously achieves highest rank correlation (QWK 0.868) and lowest absolute error (MAE 0.496). Note that while Linear Regression achieves respectable correlation, its precision (MAE) is markedly worse than the classification-based heads.

7.4.3 Training Dynamics

Figure 7.5 illustrates the validation QWK trajectories for the representative 5th fold.

The learning curves confirm the stability of the ordinal approach. Deep CORAL does not merely memorize the training data; it steadily improves its generalization metric (QWK) on the validation set, peaking at Epoch 5. The absence of significant oscillation suggests that the conservative learning rate ($1e^{-5}$) successfully mitigated the risk of catastrophic forgetting often associated with fine-tuning on small datasets.

Having established that Deep CORAL achieves the highest precision, we next investigate whether combining explicit linguistic features can further improve performance via fusion strategies.

7.5 Fusion Strategies

A central research question of this thesis was whether explicit linguistic features could complement deep contextual representations to improve predictive performance. To investigate this, we explored two fusion paradigms: Early Fusion (feature concatenation) and Late Fusion (model ensembling).

7.5.1 Early Fusion: The Redundancy Hypothesis

In the Early Fusion setup, the 24-dimensional feature vector was concatenated with the 768-dimensional `[CLS]` embedding produced by XLM-R before being passed to a CORAL classification head.

- **Deep CORAL (Reference):** QWK = 0.868, MAE = 0.496

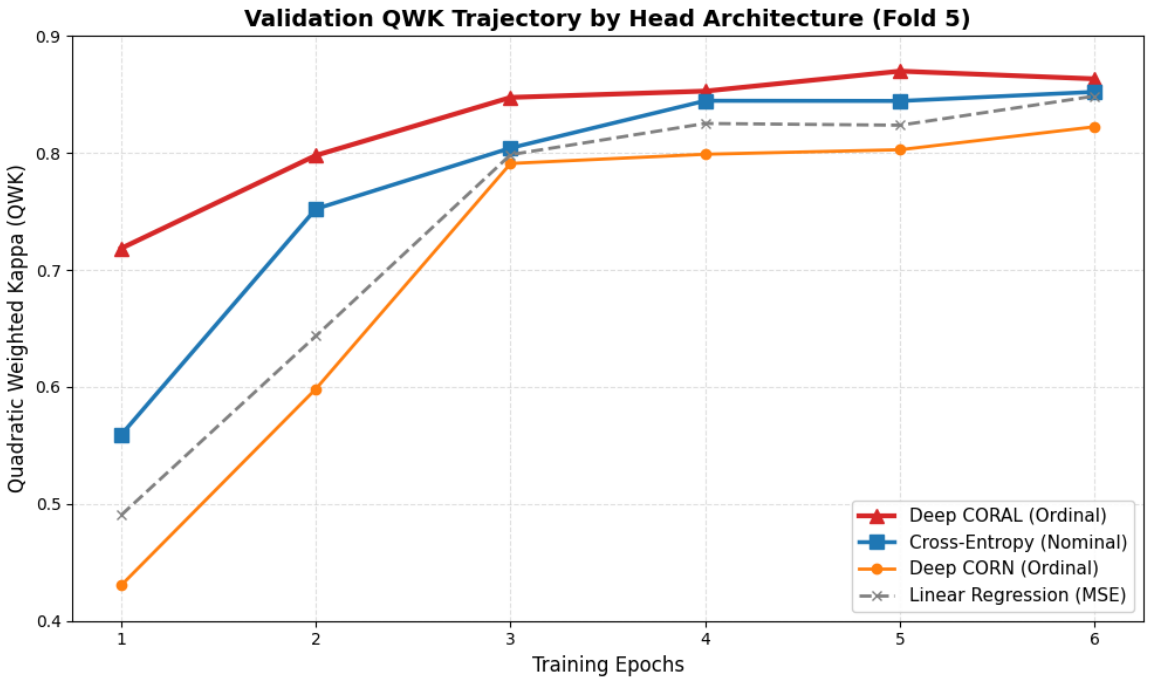

Figure 7.5: Training dynamics of XLM-R heads (Fold 5). Deep CORAL (Red) exhibits robust convergence, steadily improving until Epoch 5. Cross-Entropy (Blue) follows a similar trajectory but plateaus slightly lower. Linear Regression (Gray) shows high variance and slower adaptation, confirming that the MSE loss surface is less compatible with the semantic embedding space.

- **Early Fusion (XLM-R + Features):** QWK = 0.835, MAE = 0.592

Contrary to the hypothesis that explicit features would provide "guidance" to the network, the inclusion of the feature vector led to a notable performance degradation (-0.033 QWK). This outcome supports the **Redundancy Hypothesis**: the pre-trained Transformer has likely already internalized latent representations of surface-level properties (such as sentence length and lexical complexity). Explicitly reintroducing these signals as a separate input stream appears to introduce noise or optimization difficulties, diluting the cleaner gradient signal from the high-dimensional embeddings.

7.5.2 Late Fusion: The Limits of Ensembling

In the Late Fusion setting, we combined the predictions of the best feature-based model (Feature-CORN) and the best deep model (Deep CORAL) using a weighted average:

$$\hat{y}_{\text{final}} = \alpha \cdot \hat{y}_{\text{deep}} + (1 - \alpha) \cdot \hat{y}_{\text{feat}}$$

A grid search for α on the validation sets yielded an optimal value of $\alpha = 0.95$.

- **Feature-CORN (Reference):** QWK = 0.761
- **Deep CORAL (Reference):** QWK = 0.868
- **Late Fusion (Ensemble):** QWK = 0.861

The ensemble failed to surpass the single best deep model. The extremely high weight assigned to the deep model ($\alpha = 0.95$) indicates that the feature-based model offers virtually no corrective signal that the deep model has not already captured. Blending the weaker predictor (0.76) with the stronger one (0.87) resulted in a slight dilution of performance rather than a synergistic improvement.

7.6 Summary of Model Comparisons

Figure 7.6 and Table 7.6 summarize the progression of performance across all experimental conditions.

Category	Model Configuration	QWK	MAE	RMSE
Baselines	Logistic Regression (19 Feats)	0.713	0.730	1.101
	Linear Regression (19 Feats)	0.744	0.728	0.952
Feature-Based	Linear Regression (24 Feats)	0.760	0.708	0.919
	Feature-CORAL	0.759	0.668	0.980
	Feature-CORN	0.761	0.672	0.991
Deep Learning	XLM-R + Linear Regression (MSE)	0.832	0.605	0.887
	XLM-R + Deep CORN	0.831	0.579	0.862
	XLM-R + Cross-Entropy	0.848	0.532	0.812
	XLM-R + Deep CORAL	**0.868**	**0.496**	**0.766**
Hybrid Fusion	Early Fusion (Concat + CORAL)	0.835	0.592	0.893
	Late Fusion (Ensemble $\alpha = 0.95$)	0.861	0.519	0.845

Table 7.6: Comprehensive Leaderboard of All Experimental Configurations (5-Fold CV). Models are ranked by QWK within their respective categories. The global best performance is highlighted in bold.

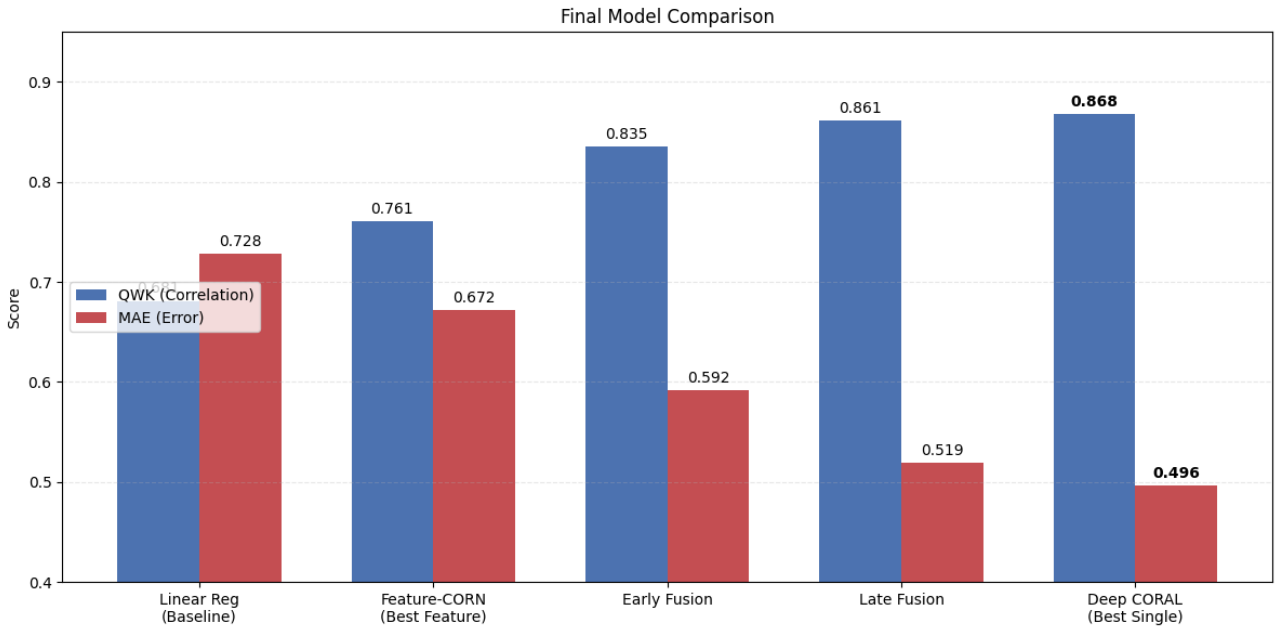

Figure 7.6: Final Model Comparison. The progression clearly demonstrates the superiority of Deep Representation Learning. While feature engineering provides a strong baseline, the Deep CORAL model achieves the state-of-the-art, with fusion strategies failing to provide additional gains.

The empirical evidence points to a definitive conclusion: **Deep Ordinal Learning (XLM-R + CORAL) is the superior approach**, suggesting that manual feature engineering offers limited additional benefit for this task under the current experimental setup. The model achieves a correlation of nearly 0.87 with human raters and an average absolute error of less than 0.5 points.

7.7 Error Analysis

To contextualize the quantitative metrics, we conducted a qualitative analysis of the Deep CORAL model's prediction behaviors. While metrics such as QWK provide a global performance summary, they obscure the specific linguistic phenomena that challenge the model.

7.7.1 Confusion Matrix Analysis

Figure 7.7 presents the confusion matrix for the Deep CORAL model on a representative validation fold.

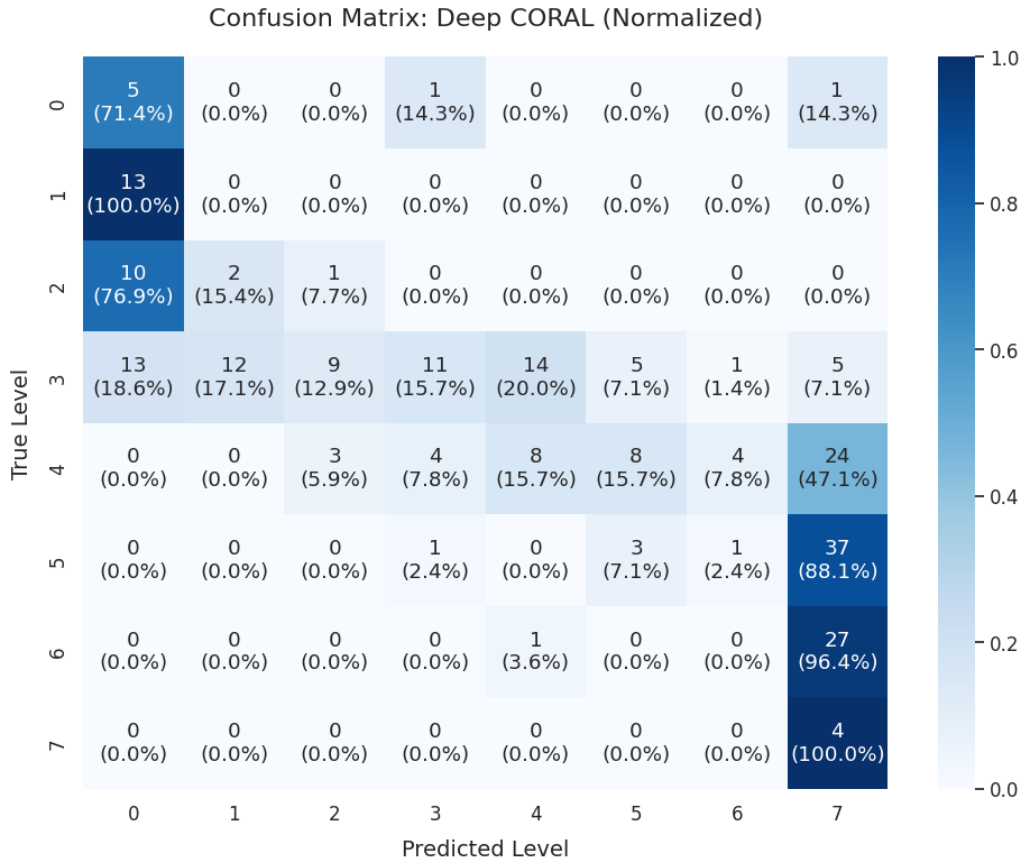

Figure 7.7: Confusion Matrix for Deep CORAL (8 Levels). The strong diagonal concentration suggests the model's reliability for the majority classes (Levels 3–5). However, a pattern of "Tail Instability" is visible: the model occasionally predicts extreme values (Level 0 or Level 7) for essays that belong to the intermediate range, indicating a sensitivity to specific features that trigger high-confidence outlier predictions.

Visual inspection confirms that the majority of errors are **adjacent misclassifications** (e.g., predicting *Intermediate Mid* when the ground truth is *Intermediate High*). The model effectively utilizes

the ordinal structure of the loss function to penalize distant predictions, resulting in a clean diagonal structure for the intermediate levels.

However, a distinct pattern of "over-reaction" emerges at the proficiency extremes. The model occasionally acts with high confidence on the edges of the scale, classifying Intermediate essays as either completely broken (Level 0) or highly advanced (Level 7). To understand this behavior, we examined specific "catastrophic" outliers where $|y_{\text{true}} - y_{\text{pred}}| \geq 3$.

7.7.2 Qualitative Analysis of Severe Errors

These failures reveal distinct divergences between the deep model's reliance on statistical patterns (token probability, lexical sophistication) and the human rater's reliance on communicative effectiveness.

Case 1: The "Unnatural Fluency" (Pragmatic Blindness). *Essay ID 560 (True: 3 [Intermediate Low] → Pred: **7 [Advanced High]**). Excerpt:* "Я боюсь будущего нашего поколения... будет трудно чтобы знать, говорил ли кто-то счастливо или грустно." *Analysis:* Formally, this text is sophisticated: the spelling is perfect, and the syntax is complex (subordinate clauses). The Deep CORAL model, detecting high-probability token sequences and correct morphology, predicted **Advanced High**. However, the human rater penalized the essay for unnatural phrasing—likely a direct calque (literal translation) from English—which lacked the idiomatic flow of a native speaker. The model successfully measured grammatical accuracy but failed to detect the lack of pragmatic authenticity.

Case 2: The "Lexical Hallucination" (Semantic Over-Estimation). *Group: True Intermediate (4/5) → Pred: **7 [Advanced High]**. Excerpt:* "Сейчас, мы живём в технологическим мире... И в этом виде мира есть плусы и минуси... Легче связываться с людям вокруг мир..." *Analysis:* The student employs high-register vocabulary ("technological world") and explicit discourse markers ("pros and cons"), signaling an Advanced-level argumentative structure. The model likely attended to these semantic features and predicted **Level 7**. However, the human rater heavily penalized the text for basic morphological errors ("с людям" instead of "с людьми") and phonetic spelling ("технологическим"). This suggests the Transformer model prioritizes semantic complexity (the "what") over morphological precision (the "how"), whereas human raters require both for an Advanced score.

Case 3: The "Morphological Panic" (Formal Under-Estimation). *Group: True Novice/Intermediate (2/3) → Pred: **0 [Novice Mid]**. Excerpt:* "Привет Дима... Я не вижу твои тёмные волосы за три года... Я хочу летать в Россию чтобы видеть тебья. Как ты чувустея себя?" *Analysis:* This text is communicative and functional—the student successfully asks questions and conveys meaning, satisfying the ACTFL criteria for Intermediate Low. However, the text is riddled with severe spelling and segmentation errors ("тебья", "чувустея"). We hypothesize that the sub-word tokenizer breaks these malformed words into rare, low-probability tokens, causing the model to output a high perplexity score akin to random noise. Consequently, the model predicts **Level 0**, failing to act as a "sympathetic reader" (a key tenet of human assessment) that overlooks formal errors in favor of communicative intent.

7.7.3 Post-hoc Granularity Analysis

A critical question for practical deployment is whether these errors are "local" (confusing adjacent levels) or "catastrophic" (confusing entirely different proficiency bands). To investigate this, we mapped the model's fine-grained 8-class predictions into three coarse proficiency categories aligned with the major ACTFL levels: *Novice* (Levels 0–2), *Intermediate* (Levels 3–5), and *Advanced* (Levels 6–7).

Table 7.7 compares the performance metrics between the fine-grained and coarse-grained evaluations.

Granularity	Accuracy	QWK	MAE	RMSE
Fine-Grained (8 Classes)	0.5175	0.8335	0.5482	0.8609
Coarse-Grained (3 Classes)	**0.6140**	0.6180*	**0.3904**	**0.6318**

Table 7.7: Performance Comparison: Fine-Grained (8 Levels) vs. Coarse-Grained (3 Levels)
*Note: QWK naturally decreases when the number of classes is reduced due to the heavier penalty weight assigned to off-diagonal errors in a smaller matrix.

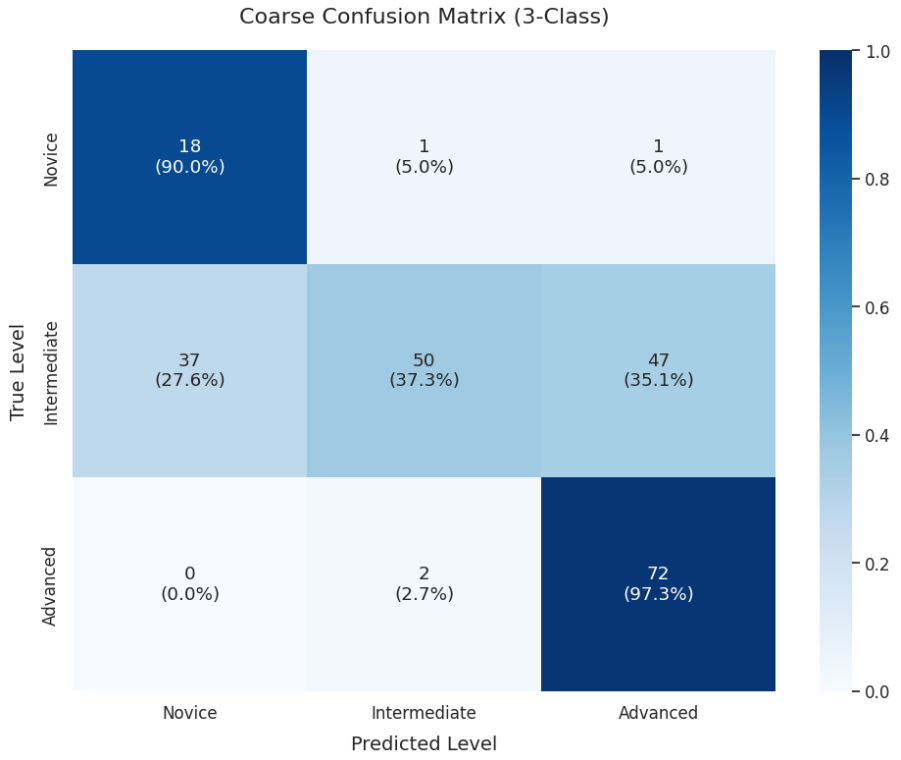

Figure 7.8: Coarse-Grained Confusion Matrix. When predictions are aggregated into major ACTFL categories, the diagonal stability improves significantly. The model rarely confuses *Novice* essays with *Advanced* ones, indicating that the majority of errors in the fine-grained model are "near-misses" within the same broad proficiency band.

The results confirm that the system's reliability increases when the granularity requirement is relaxed. The Accuracy improves by nearly 10 percentage points (to 61.4%), and the Mean Absolute Error (MAE) drops to 0.39. This implies that on average, the model's prediction is less than "half a category" away from the true label. The Coarse Confusion Matrix (Figure 7.8) visually confirms this stability: errors are heavily concentrated in the adjacent cells (e.g., *Intermediate* predicted as

Advanced), with virtually no "cross-spectrum" confusion between *Novice* and *Advanced*. This finding supports the "Tiered Deployment" strategy discussed in Chapter 8, validating the model's utility for broad placement testing.

7.8 Impact of Dataset Imbalance

A critical limitation of this study, inherent to the AES domain, is the imbalance of the proficiency distribution. The extreme classes—**Novice Mid (0)** and **Advanced High (7)**—constitute less than 5% of the training data.

7.8.1 Majority Stability vs. Tail Instability

Contrary to the common expectation that models trained on imbalanced data will simply "gravitate to the mean" (predicting only majority classes), we observed a more complex behavior characterized by tail instability:

- **Stable Core (Levels 3–5):** For the intermediate proficiency levels where training data is abundant, the model behaves conservatively, with errors tightly clustered around the diagonal.
- **Volatile Extremes (Levels 0 & 7):** The model struggles to learn robust boundaries for the rare classes. For *Advanced High (7)*, the model often under-predicts (bias towards the mean). However, for *Novice Mid (0)*, the model exhibits high variance, occasionally predicting *Advanced High (7)*. This suggests that due to the scarcity of "Novice" examples, the model has not fully learned to distinguish "simplicity due to lack of ability" from "conciseness," leading to hallucinations of high proficiency for short, error-free texts.

7.8.2 The Protective Role of Ordinal Loss

Despite this volatility at the extremes, the **Deep CORAL** architecture mitigated the overall impact better than the Nominal Cross-Entropy baseline. Because CORAL learns a cumulative threshold ($P(y > k)$), the decision boundary for Level 0 is linked to Level 1. The model does not need to see hundreds of Level 0 examples to learn the global ranking direction. This structural constraint allowed the model to maintain a state-of-the-art performance within this experimental setting (QWK **0.868**) even without synthetic oversampling, suggesting that ordinal inductive biases are a data-efficient strategy for imbalanced regression tasks.

Chapter 8

Discussion

The experimental results presented in Chapter 7 provide a comprehensive empirical evaluation of Automated Essay Scoring (AES) methodologies for Russian as a Second Language. Beyond raw performance metrics, these findings offer insights into the nature of linguistic proficiency, the capabilities of deep representation learning, and the challenges of modeling the ACTFL scale as an ordered construct.

This chapter synthesizes these findings to address the core research questions of the thesis. We first analyze the comparative efficacy of explicit feature engineering and latent neural representations through the lens of **construct validity**—the degree to which the models capture the intended construct of proficiency rather than surface-level correlates. We then discuss the theoretical implications of the observed ordinal advantage, interpret the learned decision boundaries of the CORAL architecture, and examine the empirical results of hybrid fusion strategies. Finally, we contextualize the findings within the linguistic landscape of Russian and outline implications for L2 assessment systems.

8.1 Insights from Feature-Based and Transformer-Based Approaches

The performance gap between the best feature-based model (Feature-CORN, QWK 0.761) and the best deep learning model (XLM-R + CORAL, QWK 0.868) is substantial ($+0.107$). In AES research, a 0.1 improvement in weighted Kappa typically represents a meaningful practical difference, bridging the gap between experimental prototypes and operationally viable systems. This divergence provides empirical evidence relevant to Research Question 1 regarding the comparative effectiveness of engineered features and pre-trained Transformer representations.

8.1.1 The Validity and Limits of Explicit Linguistic Signals

The feature-based models significantly outperformed naive baselines ($QWK \approx 0.0$), indicating that L2 proficiency can be partially quantified through surface-level linguistic metrics. A linear combination of 24 interpretable features achieved a correlation of approximately 0.76 with human raters, supporting the **inter-supportive hypothesis**: proficiency manifests simultaneously across multiple observable dimensions. Learners who employ more advanced vocabulary (captured by the lexical features) are statistically likely to produce longer sentences (Syntactic Complexity) and make fewer morphologi-

cal errors (Error Rate). This multicollinearity allows linear models to approximate proficiency even without "understanding" the text.

Proxy Measurement and Construct Under-Representation

However, the plateau in feature-based performance suggests an inherent limitation: manual features function as *proxies* rather than direct measurements of proficiency.

- **Length vs. Quality:** Syntactic features captured sentence length and tree depth. While advanced essays are often longer, length alone does not guarantee discourse quality. As discussed in the error analysis (Section 7.7), repetitive clause chaining can inflate complexity metrics without reflecting genuine proficiency.
- **Diversity vs. Appropriateness:** Lexical rarity measures quantify frequency bands but do not assess contextual appropriateness. A rare lexical item used inaccurately may increase a diversity score while reducing communicative effectiveness.

Thus, explicit features capture aspects of linguistic **form**, but they under-represent aspects of **function**, including discourse coherence and pragmatic appropriateness. This limitation constrains their construct validity.

8.1.2 Latent Representations and Holistic Modeling

The improvement achieved by XLM-R suggests that pre-trained Transformer representations capture additional dimensions of the proficiency construct. The feature-based ceiling effect indicates that explicit metrics encode important but incomplete signals.

Modeling Pragmatic and Discourse Competence

Self-attention enables modeling of long-range dependencies and token interactions beyond sentence-level aggregation. This appears particularly relevant for:

1. **Collocational Accuracy:** Russian fluency often depends on conventional verb–noun pairings (e.g., "принимать решение"). Count-based features treat tokens independently, whereas contextual embeddings encode distributional regularities of multi-word units.
2. **Discourse Coherence:** Feature aggregation typically operates at the sentence level. Transformer architectures, by contrast, condition each token representation on the full input sequence, enabling implicit modeling of logical progression and global cohesion.

These findings suggest that latent representations capture information not fully accessible through manually engineered surface statistics. This directly addresses Research Question 1: Transformer-based models provide broader construct coverage than feature-only approaches in this dataset.

8.1.3 The Interpretability Trade-Off

Despite improved predictive performance, Transformer models introduce reduced transparency.

- **Feature-Based Models:** Linear models provide interpretable coefficients, enabling targeted pedagogical feedback (e.g., "Your lexical diversity is below the B2 average").
- **Transformer Models:** Predictions are derived from high-dimensional latent representations that are not directly interpretable without auxiliary explainability techniques.

This trade-off has deployment implications. High-stakes settings may prioritize predictive reliability, whereas formative contexts may require explainability mechanisms or hybrid reporting strategies.

8.2 Ordinal Classification Advantages for ACTFL Scoring

A central methodological objective of this thesis was to evaluate whether ordinal regression better reflects the ACTFL proficiency construct than nominal classification or linear regression. The results consistently indicate that modeling proficiency as an ordered scale yields superior performance across architectures, addressing Research Question 2.

8.2.1 The Equidistance Assumption in Linear Regression

Linear regression minimizes Mean Squared Error under an implicit assumption of equidistance between adjacent levels. However, SLA theory and ACTFL descriptors describe proficiency development as non-linear.

- **Rapid Early Development:** Novice-level progression often involves the rapid acquisition of high-frequency vocabulary and basic syntactic templates.
- **The Intermediate Plateau:** Progression toward Advanced levels typically slows down, requiring a fundamental restructuring of discourse-level competence and pragmatic control.

A single linear mapping imposes a linear fit on this non-linear reality, inevitably misrepresenting the unequal spacing between levels. The higher MAE observed for regression models (0.708) is consistent with this structural misalignment.

8.2.2 Nominal Classification and Distance Agnosticism

Standard cross-entropy classification ignores ordinal distance entirely. All misclassifications incur equal structural penalty regardless of magnitude. Although the XLM-R nominal model achieved strong performance ($QWK = 0.848$), it does not explicitly optimize for distance-weighted agreement.

Qualitative inspection suggested that ordinal models produce more localized probability distributions concentrated around adjacent levels, reducing the risk of "catastrophic" outliers (e.g., predicting Level 0 for a Level 7 essay) which are penalized heavily by the QWK metric.

8.2.3 Alignment of Ordinal Loss with Proficiency Structure

The CORAL and CORN architectures model cumulative probabilities $P(y > k)$, respecting order without assuming equal spacing. The learnable thresholds (b_k) allow the model to learn flexible spacing between adjacent levels in the latent space. This provides an inductive bias consistent with the ACTFL hierarchy.

The superior performance of ordinal deep models ($QWK = 0.868$) supports the hypothesis that ordinal objectives better align with the topological structure of proficiency scales.

8.2.4 Information Sharing Under Class Imbalance

In imbalanced datasets, rare classes (like Level 0 and Level 7) suffer from limited signal. Ordinal formulations link decision boundaries across levels: to classify a text as Level 7, the model must implicitly classify it as $>$ Level 6. This allows gradient information from all samples to shape the global ordering, partially explaining the stability of CORAL performance on underrepresented levels.

8.3 Interpretation of Learned Thresholds (CORAL)

The learned thresholds in the Feature-CORAL model reflect latent transition points between levels. Their non-uniform spacing suggests that the ACTFL scale is not linear in feature space.

Two interpretations are plausible:

1. **Linguistic Distinctiveness:** Larger latent gaps may correspond to transitions requiring qualitatively different linguistic competencies (e.g., the jump from sentence-level Novice production to paragraph-level Intermediate discourse).

2. **Imbalance Calibration:** Threshold spacing may partially reflect the model's adaptation to class frequency. A wide gap between thresholds acts as a learned prior, requiring stronger evidence to push a prediction into a rare class.

While causal attribution cannot be established definitively, the observed non-uniformity reinforces the argument against simple linear modeling of proficiency.

8.4 Strengths and Limitations of the Combined Model

The fusion experiments yielded a scientifically important negative result: hybridization did not improve performance. Both Early Fusion (concatenation) and Late Fusion (ensembling) failed to surpass the single best Deep Learning model. This addresses Research Question 3 regarding the added value of hybridization.

8.4.1 Redundancy of Explicit Features

The near-zero ensemble contribution of feature-based predictions ($\alpha \approx 0.95$ favoring the deep model) suggests that most predictive signal captured by manual features was already encoded in the Transformer representations. This is consistent with the **Redundancy Hypothesis**: a sufficiently large pre-trained model implicitly learns to track surface-level statistics like sentence length and lexical variety, reducing the marginal utility of external counters.

8.4.2 Early Fusion and Representation Mismatch

The performance degradation in Early Fusion (-0.033 QWK) likely stems from a representational incompatibility between dense, contextual embeddings and low-dimensional handcrafted features. The gradient descent process may struggle to balance the learning rates for these two disparate input streams, effectively treating the manual features as noise rather than signal.

8.4.3 Engineering Implications

Given negligible performance improvement and increased architectural complexity, a pure Transformer-based approach appears empirically justified for high-accuracy deployment within the constraints of this study. Feature-based models remain valuable strictly for interpretability and scenarios where computational resources are constrained.

8.5 Implications for Russian L2 Assessment

This study contributes to AES research beyond English by examining Russian, a morphologically rich language with free word order.

8.5.1 Morphology as a Signal

Grammar error rate emerged as a consistent predictor. In Russian, morphological endings encode syntactic relations, and incorrect inflection can disrupt intelligibility. The success of XLM-R suggests that its sub-word tokenizer (SentencePiece) effectively handles this morphology by associating specific suffixes (e.g., "стол-ом") with syntactic contexts.

8.5.2 Free Word Order and Self-Attention

Self-attention mechanisms are naturally compatible with the long-distance dependencies characteristic of Russian syntax (e.g., separation of subject and verb). This may partially explain the strong deep model performance relative to surface-level syntactic metrics which do not account for word order.

8.5.3 Granularity and Deployment: A Tiered Strategy

A practical finding derived from our error analysis is that AES reliability depends heavily on the required granularity. While the model occasionally confuses adjacent 8-point levels (e.g., *Intermediate Low* vs. *Mid*), the aggregation of the confusion matrix reveals that "cross-category" errors are rare.

When we map predictions to coarse categories (*Novice*, *Intermediate*, *Advanced*), the system's reliability increases significantly (Coarse Accuracy $> 61\%$, MAE ≈ 0.39). This suggests a tiered deployment strategy:

1. **Tier 1 (Automated Placement):** Use the Deep CORAL model to place students into broad course levels (e.g., Russian 101 vs. 201). The high coarse accuracy ensures minimal misplacement between major proficiency bands.

2. **Tier 2 (Formative Feedback):** For granular grading, use the system as a "second opinion" or formative tool, while relying on human raters for high-stakes certification at the sub-level boundaries.

8.6 Methodological Limitations

Several limitations constrain the generalizability of these findings.

- **Dataset Scope:** The dataset is limited in size ($N \approx 1,100$) and exhibits class imbalance, particularly at the extreme proficiency levels (Novice Mid and Advanced High).

- **Label Reliability:** ACTFL ratings are human judgments and contain inherent subjectivity. The model's "errors" may in some cases reflect label noise rather than predictive failure.

- **Single Language:** All experiments were conducted on Russian L2 essays. While the architecture is language-agnostic, the specific feature importance (e.g., morphology) may not transfer directly to analytic languages like English or Chinese.

- **Prompt Constraints:** Essays were drawn from a specific academic context, potentially restricting conclusions about cross-topic generalization.

These limitations do not invalidate the results but define their interpretive boundaries. Future work should evaluate ordinal modeling under larger, multi-language conditions to further validate the "Ordinal Advantage."

Taken together, these findings suggest that proficiency assessment in morphologically rich languages benefits from contextualized representations and ordinal learning objectives, while hybridization with surface features offers limited marginal gains under strong pre-trained encoders.

Chapter 9

Conclusion and Future Work

This thesis addressed a central gap in Automated Essay Scoring (AES): the limited availability of robust, deep learning-based assessment systems for morphologically rich languages such as Russian, particularly within resource-constrained educational contexts. By systematically benchmarking hand-crafted linguistic features against deep representation learning, and by rigorously evaluating ordinal versus nominal modeling strategies, this study provides a structured empirical foundation for Russian L2 proficiency modeling.

Importantly, the research was conducted not on idealized or synthetic corpora, but on the imbalanced and institutionally grounded dataset of the Middlebury Russian School. The results suggest that strong predictive performance is achievable even with modest datasets ($N \approx 1,100$), provided that the modeling architecture—especially the loss function—is aligned with the ordered structure of the proficiency construct.

9.1 Summary of Contributions

The contributions of this thesis can be grouped into three interrelated domains: empirical, methodological, and architectural.

9.1.1 Empirical: A Benchmark for Russian L2 AES

This study establishes a comprehensive empirical benchmark for AES in Russian, a language characterized by fusional morphology and flexible word order. Feature-based models relying on syntactic depth, lexical frequency, and error-rate metrics achieved substantial agreement with human raters ($QWK \approx 0.76$), demonstrating that surface-level linguistic signals capture meaningful aspects of proficiency.

However, performance plateaued under purely engineered representations. Fine-tuning a multilingual Transformer (XLM-R) increased agreement to $QWK = 0.868$, indicating that contextualized representations capture additional dimensions of the construct not fully accessible through manual aggregation. These findings suggest that deep transfer learning can substantially mitigate data scarcity constraints in small institutional datasets, allowing pre-trained models to adapt effectively to learner language.

9.1.2 Methodological: The Ordinal Imperative

A central methodological finding of this thesis is the consistent advantage of ordinal regression over both nominal classification and linear regression.

The ACTFL proficiency scale is hierarchical and non-equidistant. Linear regression imposes an assumption of equal spacing between adjacent levels, which does not reflect observed learning trajectories (e.g., the well-documented Intermediate plateau). Nominal cross-entropy classification, in contrast, ignores ordinal distance entirely, treating all misclassifications as structurally equivalent.

By implementing the CORAL (Consistent Rank Logits) framework, this thesis introduced an inductive bias that respects the ordered nature of proficiency. Linking adjacent decision boundaries allowed the model to learn a coherent ranking direction across levels, improving stability under class imbalance. Within the constraints of this dataset, ordinal modeling yielded the strongest and most consistent performance across architectures.

9.1.3 Architectural: The Limited Utility of Hybridization

Fusion experiments were motivated by the hypothesis that explicit linguistic features might complement deep embeddings. However, neither early nor late fusion surpassed the best Transformer-based ordinal model.

These results provide empirical support—within this experimental setting—for the **Redundancy Hypothesis**: fine-tuned Transformer representations appear to encode much of the syntactic and lexical information traditionally captured through manual feature engineering. While feature-based models retain interpretability advantages, their predictive contribution under hybridization was minimal.

In direct response to the research questions posed in Chapter 1:

- **RQ1:** Transformer-based models substantially outperform feature-based approaches for Russian L2 AES.
- **RQ2:** Ordinal objectives provide measurable and consistent gains over nominal and linear formulations.
- **RQ3:** Hybrid feature–embedding fusion does not yield additional improvements when strong pre-trained encoders are used.

9.2 Practical Relevance for Automated Scoring

Beyond methodological contributions, the findings carry practical implications for language programs operating under realistic constraints.

9.2.1 A Blueprint for Language Schools

Many institutions possess datasets similar in scale and distribution to the one used in this study: several thousand historically graded essays with imbalanced proficiency levels. The present results suggest that such data may be sufficient to train a reliable AES system for coarse-grained placement.

When predictions are aggregated into broad ACTFL categories (Novice, Intermediate, Advanced), model accuracy improves significantly (to $> 60\%$), while the Mean Absolute Error drops to 0.39. This indicates that while the model may struggle with fine-grained sub-level distinctions (e.g., Low vs. Mid), it is highly reliable at placing students into the correct major proficiency band. This level of reliability supports the feasibility of automated **Tier 1** placement, potentially reducing instructor workload in large programs.

9.2.2 A Two-Track Deployment Strategy

The results also motivate a dual deployment framework that satisfies the "Qualitative Validity" criterion established in the Introduction:

1. **Evaluator (Deep Ordinal Model):** For placement or summative contexts where predictive accuracy is paramount, the XLM-R + CORAL architecture offers the highest reliability and may function as a calibrated second rater.

2. **Tutor (Interpretable Feature Model):** For formative assessment, feature-based regression provides transparent feedback signals (e.g., lexical diversity, syntactic complexity), enabling actionable pedagogical guidance even if the overall score correlation is slightly lower.

Such a two-track framework reconciles predictive performance with interpretability considerations, aligning technical capability with educational priorities.

9.3 Future Directions

While this thesis establishes a structured empirical baseline, several extensions would strengthen and generalize the findings.

9.3.1 Data Expansion and Stricter Validation

The primary constraint of this study was dataset size ($N = 1,126$). Future work may explore:

- **Synthetic Data Augmentation:** Controlled generation of essays for underrepresented levels using large language models could help stabilize decision boundaries at proficiency extremes. Careful filtering and human verification would be essential to avoid distributional artifacts.

- **Stricter Generalization Protocols:** Leave-One-Prompt-Out or Leave-One-Year-Out validation would provide stronger evidence of topic-level robustness and reduce the risk of prompt-specific overfitting.

9.3.2 Domain-Adaptive Pre-training

An intermediate phase of Domain-Adaptive Pre-training (DAPT), continuing masked language modeling on a corpus of learner Russian, could better align pre-trained encoders with interlanguage characteristics. Such adaptation may improve sensitivity to morphosyntactic error patterns that distinguish adjacent proficiency levels.

9.3.3 Parameter-Efficient Fine-Tuning

In this thesis, the XLM-R encoder was fully fine-tuned, which is computationally expensive. Future work should explore **parameter-efficient adaptation methods**, such as *Low-Rank Adaptation (LoRA)*. These techniques freeze the majority of pre-trained parameters and learn a small set of task-specific weights. This approach would be particularly valuable for educational institutions with limited hardware, enabling the deployment of large foundation models on consumer-grade GPUs while potentially reducing overfitting on small datasets.

9.3.4 Multi-Task Learning

Future systems may benefit from Multi-Task Learning (MTL) frameworks that jointly predict proficiency scores and auxiliary linguistic tasks, such as grammatical error detection. Sharing representations across tasks could encourage the encoder to attend more closely to morphosyntactic distinctions, potentially enhancing discrimination at critical level boundaries where formal accuracy is a key discriminator.

Final Remarks:

This thesis demonstrates that reliable Automated Essay Scoring for L2 Russian is achievable within realistic institutional constraints. By aligning modeling assumptions with the ordinal structure of proficiency and leveraging contextualized Transformer representations, it is possible to construct systems that are both empirically strong and pedagogically relevant. While further validation and scaling are required, the findings presented here contribute a principled foundation for future work at the intersection of artificial intelligence and language education.

Bibliography

[1] Yigal Attali and Jill Burstein. Automated essay scoring with e-rater v.2. In *Proceedings of the Annual Meeting of the International Association for Educational Assessment*, 2006.

[2] Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, Martin Chodorow, Lisa Braden-Harder, and Mary Dee Harris. Automated scoring using a hybrid feature identification technique. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 206–210, 1998.

[3] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020.

[4] Meri Coleman and T. L. Liau. Computer readability formulas. Technical report, Human Communication Research, 1975.

[5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics, 2020.

[6] Michael A. Covington and James D. McFall. Cutting the gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100, 2010.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[8] Rudolf Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221, 1948.

[9] Peter W Foltz, Darrell Laham, and Thomas K Landauer. Automated essay scoring using semantic similarity. *Assessing Writing*, 11(3):115–132, 2010.

[10] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. In *Proceedings of the 56th*

Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 1–7, 2017.

[11] Olesya Kisselev, Andrey Klimov, and Mikhail Kopotev. Syntactic complexity measures as indices of language proficiency in writing: Focus on heritage learners of russian. *Heritage Language Journal*, 18(3):1–30, 2021.

[12] Olesya Kisselev, Andrey Klimov, and Mikhail Kopotev. Syntactic complexity measures as linguistic correlates of proficiency level in learner russian. In Agnieszka Leńko-Szymańska and Sandra Götz, editors, *Complexity, Accuracy and Fluency: Learner Corpus Research*, pages 51–80. John Benjamins, 2022.

[13] O Kolak and Others. Automated readability index. *Advances in Psychology*, 71:121–132, 1990.

[14] Mikhail Kopotev, Olesya Kisselev, and Daria Kormacheva. Exploring collocational complexity in L2 russian: A corpus-driven contrastive analysis. *Russian Language Journal*, 73(2), 2023.

[15] Thomas K Landauer, Peter W Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

[16] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019.

[17] G.H. McLaughlin. SMOG grading: A new readability formula. *Journal of Reading*, 12:639–646, 1969.

[18] Eric Nyberg. Languagetool: Open-source grammar, style and spell checker, 2020.

[19] Ellis B Page. The imminence of grading essays by computer. *Phi Delta Kappan*, 47(5):238–243, 1966.

[20] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jamie Bolton, and Christopher D Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, 2020.

[21] Xintong Shi, Wenzhi Cao, and Sebastian Raschka. Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *Pattern Recognition*, 113:107809, 2021.

[22] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1882–1891, 2016.

[23] Masaki Uto, Yikuan Xie, and Maomi Ueno. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 6077–6088, 2020.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

[25] Yongjie Wang, Chuan Wang, Ruobing Li, and Hui Lin. On the use of BERT for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 3418–3429, 2022.

Appendix A

Reproducibility and Implementation Details

To ensure the transparency and reproducibility of the experimental results presented in this thesis, this appendix provides the specific hyperparameter configurations used for the final models and details on code availability.

A.1 Code Availability and Repository Structure

The complete source code for this project is hosted on GitHub to support the reproducibility of our findings. The repository is organized as a sequential research pipeline:

`https://github.com/nikgorbachev/russian-l2-proficiency-scoring`

- `notebooks/01_preprocessing/`: Scripts for text normalization and initialization of the *Stanza* and *spaCy* pipelines.
- `notebooks/02_feature_extraction/`: Implementation of the 24 linguistic features used for syntactic and lexical profiling.
- `notebooks/03_baselines/`: Dummy baselines and Reference statistical models (Linear and Logistic Regression).
- `notebooks/04_feature-based_models/`: Linear Regression on Extended Feature Set (24 featurs) and Ordinal models implementations (Feature-CORAL/CORN).
- `notebooks/05_deep_models/`: Fine-tuning notebooks for the XLM-R architectures.
- `notebooks/06_fusion/`: Implementation and evaluation of Early and Late fusion strategies.
- `notebooks/07_visualisations/`: Scripts for generating the confusion matrices and error distribution plots reported in the results.

A.2 Hyperparameter Configurations

A.2.1 Deep Learning Models (Transformer-Based)

All Transformer-based models (XLM-R with Nominal, CORAL, and CORN heads, as well as Early Fusion) were trained using the Hugging Face `transformers` library. While most settings were kept constant to ensure a rigorous comparison, the Learning Rate was tuned specifically for the ordinal architectures, which proved more sensitive to optimization magnitude. Table A.1 details these settings.

Hyperparameter	Value
Base Model	`xlm-roberta-base`
Optimizer	AdamW
Learning Rate	2×10^{-5} (Nominal & Linear Regression) 1×10^{-5} (CORAL, CORN & Early Fusion)
Batch Size	8
Gradient Accumulation Steps	1 (Effective Batch Size = 8)
Weight Decay	0.01
Max Sequence Length	512 tokens
Num Epochs	6
Early Stopping Patience	3 epochs
Scheduler	Linear Decay (Default)
Seed	42 (fixed for all folds)

Table A.1: Hyperparameter configuration for XLM-R based experiments. Note that the ordinal models (CORAL/CORN) required a lower learning rate ($1e^{-5}$) for stable convergence compared to the standard nominal baselines ($2e^{-5}$).

A.2.2 Feature-Based and Statistical Baselines

The non-neural baselines were implemented using `scikit-learn` and `PyTorch`.

- **Linear/Logistic Regression:** We utilized standard Ridge (L2) Regularization with an inverse regularization strength of $C = 1.0$ and the L-BFGS solver for optimization.
- **Feature-CORAL/CORN (Non-Deep):** These models employed a simple Multi-Layer Perceptron (MLP) architecture projecting the 24-dimensional feature vector to the ordinal heads. Due to the low dimensionality of the input, they were trained with a significantly higher learning rate (1×10^{-2}) compared to the deep models, and a batch size of 32.

A.3 Software Dependencies

The experimental environment was built on the Python 3.12 ecosystem, utilizing a combination of transformer-based architectures and linguistically-informed NLP pipelines. The core library versions

used in the final evaluation are:

- **Python:** 3.12.3
- **PyTorch:** 2.9.0 (with CUDA 12.6 support)
- **Transformers (Hugging Face):** 5.0.0
- **Stanza:** 1.11.0 (for high-accuracy Russian morphological and dependency parsing)
- **spaCy:** 3.8.11 (utilizing the `ru_core_news_md` model for lexical and POS-based features)
- **Pydantic:** 2.12.3 (for robust data validation within the feature extraction pipeline)

The integration of *Stanza* was specifically chosen for the extraction of syntactic complexity features (e.g., Mean Dependency Distance) due to its state-of-the-art performance on the `syntagrus` Universal Dependencies treebank for Russian. *spaCy* was utilized as a secondary high-speed pipeline for broader lexical profiling and part-of-speech distribution metrics.

A.4 Implementation Notes: NLP Pipelines

To capture the multidimensional nature of Russian proficiency, two distinct NLP engines were utilized:

1. **Stanza (Stanford NLP):** We employed the Stanza `ru` (Russian) model using the `syntagrus` package. This was used for deep dependency parsing and lemmatization. As Russian is morphologically rich, Stanza's character-level language models (*charlm*) provided the necessary precision for calculating Average Tree Depth and complex morphological ratios.
2. **spaCy (Explosion AI):** The `ru_core_news_md` (Medium) model was used for efficient tokenization and Named Entity Recognition (NER). It served as the primary engine for calculating length-based metrics and lexical diversity indices (MATTR), where high-throughput processing was required.

A.5 Disclosure of Generative AI Use

In accordance with the University of Bologna's policy on the responsible use of Artificial Intelligence, I declare that Generative AI tools were utilized during the preparation of this thesis. Specifically:

- **Tools Used:** Google Gemini 3 and OpenAI GPT-5.2.
- **Scope of Use:** These tools were employed for:
  - *Code Assistance:* Debugging, refactoring, and generating boilerplate code for Python scripts.
  - *Text Editing:* Proofreading, linguistic polishing, and improving the flow of English prose.

- *Ideation:* Assisting in the structuring of the initial table of contents and refining the formatting of LaTeX tables.

The author maintains full responsibility for the accuracy, originality, and scientific integrity of the content presented in this document. All AI-generated suggestions were critically reviewed and verified against primary sources and experimental data.