MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 33, NO. 1, 2026
journal homepage: www.mais-journal.ru

ARTIFICIAL INTELLIGENCE

Multi-tier Linguistic Feature Engineering for CEFR Classification: A Comprehensive Analysis of Deterministic and Machine Learning-based Features

T. M. Chikake[1], E. M. Bazanova[1], A. V. Gorizontova[1]

DOI: 10.18255/1818-1015-2026-1-6-29

[1]Moscow Institute of Physics and Technology, Dolgoprudny, Russia

MSC2020: 68T50
Research article
Full text in English

Received December 25, 2025
Revised February 25, 2026
Accepted February 27, 2026

We analyzed 133 linguistic features for automated proficiency classification under the Common European Framework of Reference (CEFR) in a two-tier architecture: deterministic Tier 1 (lexical, morphological, and syntactic measures) and machine-learning-based Tier 2 (semantic coherence, topic structure, cohesion, and error-analysis signals). Experiments were conducted on a corpus of 3,205 learner texts from mixed sources, with triangulated validation against expert-verified Cambridge examination subsets. The materials were collected in 2022–2025 and included a substantial institutional corpus of over 3,000 essays and other writing texts produced by students of Moscow Institute of Physics and Technology (MIPT) studying English as a foreign language and regularly assessed by our AI-powered testing system ISTOK (Intelligent System for Testing General Language Competencies). Feature matrices were standardized after missing-value handling (fold-local median imputation for cross-validation and zero-fill for held-out reporting). In supervised evaluation, the best Tier 1 + 2 model reaches 66.72 % exact accuracy (macro F1 = 0.69) and 94.53 % adjacent accuracy (within one CEFR level) on a 3,198-sample CEFR-labeled benchmark split; an extended comparison including preliminary Tier 3 features achieves 67.50 %. Unsupervised analyses show strong structure for extreme levels (A1 99.5 % purity; C2 82.4 % purity) and moderate alignment with professional Cambridge labels (Adjusted Rand Index = 0.303). We report block ablations and compact subset searches, with strongest signals from morphological complexity and lexical sophistication, and consistent incremental gains from error-based features. The results provide a validated, interpretable feature inventory and practical guidance for feature selection in automated language assessment systems.

Keywords: CEFR classification; linguistic features; multi-tier architecture; automated assessment; feature engineering; natural language processing

INFORMATION ABOUT THE AUTHORS

Chikake, Tendai M. (corresponding author)	ORCID iD: 0000-0002-9512-1256. E-mail: tendaichikake@phystech.edu PhD, Junior Researcher
Bazanova, Elena M.	ORCID iD: 0000-0002-6306-8892. E-mail: bazanova.em@mipt.ru PhD, Associate Professor
Gorizontova, Anna V.	ORCID iD: 0009-0007-4151-3338. E-mail: gorizontova.av@mipt.ru PhD, Associate Professor

Funding: Ministry of Science and Higher Education of the Russian Federation (Agreement No. 075-03-2026-305, January 16, 2026, project "Applied Research on the Implementation of Artificial Intelligence Technologies in Higher Education", project code: FSMG-2025-0086).

For citation: T. M. Chikake, E. M. Bazanova, and A. V. Gorizontova, "Multi-tier linguistic feature engineering for CEFR classification: a comprehensive analysis of deterministic and machine learning-based features", *Modeling and Analysis of Information Systems*, vol. 33, no. 1, pp. 6–29, 2026. DOI: 10.18255/1818-1015-2026-1-6-29.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 33, № 1, 2026
сайт журнала: www.mais-journal.ru

Многоуровневое лингвистическое конструирование признаков для классификации CEFR: комплексный анализ детерминированных и основанных на машинном обучении признаков

Т. М. Чикаке[1], Е. М. Базанова[1], А. В. Горизонтова[1]

DOI: 10.18255/1818-1015-2026-1-6-29

[1]Московский физико-технический институт, Долгопрудный, Россия

УДК 004.912
Научная статья
Полный текст на английском языке

Получена 25 декабря 2025 г.
После доработки 25 февраля 2026 г.
Принята к публикации 27 февраля 2026 г.

Представлен анализ 133 лингвистических признаков для автоматической классификации уровня владения языком по шкале Common European Framework of Reference (CEFR) в двухуровневой архитектуре: детерминированные признаки Tier 1 (лексические, морфологические и синтаксические показатели) и признаки Tier 2, основанные на методах машинного обучения (семантическая связность, тематическая структура, когезия и сигналы ошибок). Эксперименты выполнены на корпусе из 3,205 текстов обучающихся из разнородных источников; валидация проводилась с триангуляцией по экспертно верифицированным подмножествам экзаменационных данных Cambridge. Материалы были собраны в 2022–2025 годах и включают существенный институциональный корпус из более чем 3,000 эссе и других письменных работ студентов Московского физико-технического института (МФТИ), изучающих английский язык как иностранный; уровень владения языком у них регулярно оценивается нашей интеллектуальной системой тестирования ISTOK (Intelligent System for Testing General Language Competencies). Признаковые матрицы стандартизировались после обработки пропусков (медианная импутация внутри фолда для кросс-валидации и заполнение нулями для отчётных экспериментов на отложенной выборке). В задаче контролируемой классификации лучшая модель Tier 1 + 2 достигает 66.72 % точности (макро F1 = 0.69) и 94.53 % смежной точности (ошибка не более чем на один уровень CEFR) на эталонном разбиении для 3,198 текстов с метками CEFR; расширенное сравнение с предварительными признаками Tier 3 достигает 67.50 %. Неконтролируемый анализ выявляет структуру пространства признаков: для крайних уровней наблюдаются кластеры высокой чистоты (A1 99.5 %; C2 82.4 %), а на экспертных данных Cambridge получено умеренное согласование с профессиональными оценками (Adjusted Rand Index = 0.303). Приведены результаты блочных абляций и поиска компактных подмножеств; наибольшую информативность обеспечивают морфологическая сложность и лексическая продвинутость, а признаки ошибок дают устойчивый дополнительный выигрыш.

Ключевые слова: классификация CEFR; лингвистические признаки; многоуровневая архитектура; автоматизированная оценка; инженерия признаков; обработка естественного языка

ИНФОРМАЦИЯ ОБ АВТОРАХ

Чикаке, Тендай Мапунгвана (автор для корреспонденции)	ORCID iD: 0000-0002-9512-1256. E-mail: tendaichikake@phystech.edu Канд. тех. наук, младший научный сотрудник
Базанова, Елена Михайловна	ORCID iD: 0000-0002-6306-8892. E-mail: bazanova.em@mipt.ru Канд. пед. наук, доцент
Горизонтова, Анна Всеволодовна	ORCID iD: 0009-0007-4151-3338. E-mail: gorizontova.av@mipt.ru Канд. ист. наук, доцент

Финансирование: Министерство науки и высшего образования Российской Федерации (соглашение № 075-03-2026-305 от 16 января 2026 г., проект «Прикладные исследования по внедрению технологий искусственного интеллекта в высшем образовании», код проекта: ФСМГ-2025-0086).

Для цитирования: T. M. Chikake, E. M. Bazanova, and A. V. Gorizontova, “Multi-tier linguistic feature engineering for CEFR classification: a comprehensive analysis of deterministic and machine learning-based features”, *Modeling and Analysis of Information Systems*, vol. 33, no. 1, pp. 6–29, 2026. DOI: 10.18255/1818-1015-2026-1-6-29.

Introduction

Automated assessment of language proficiency has emerged as a critical research area driven by increasing demand for scalable, objective evaluation of second language competence across educational and professional contexts. The Common European Framework of Reference for Languages (CEFR) provides a standardized framework for proficiency assessment, defining six levels from basic (A1) to proficient user (C2) that guide language education and certification worldwide [1]. However, automated classification of texts according to CEFR levels remains challenging due to the complex, multidimensional nature of language proficiency and the subtle linguistic distinctions that characterize developmental progressions across proficiency levels.

Traditional approaches to automated language assessment have relied primarily on surface-level features such as text length, basic readability metrics, and frequency-based vocabulary analysis [2]. While these approaches provide useful baselines, they fail to capture the rich linguistic complexity that distinguishes proficiency levels, as vocabulary frequency alone cannot reliably discriminate between adjacent CEFR levels at any point of the proficiency scale. Recent advances in computational linguistics and machine learning present opportunities for more sophisticated feature engineering that can capture morphological sophistication, grammatical range and accuracy, syntactic complexity, semantic coherence, and discourse organization patterns that characterize different proficiency levels.

Comprehensive feature engineering for CEFR classification requires systematic analysis across multiple dimensions of linguistic competence. Deterministic linguistic features provide reliable, interpretable measures of lexical sophistication, morphological complexity, and syntactic structure that align with theories of language development. Machine learning-based features can capture semantic patterns, discourse coherence, and topic organization that are difficult to quantify through rule-based approaches alone. However, much prior work studies narrow feature subsets in isolation, leaving gaps in our understanding of complementarity, redundancy, and practical trade-offs for deployment.

This paper addresses these limitations by presenting a comprehensive analysis of a multi-tier feature architecture encompassing deterministic linguistic measures (Tier 1) and machine learning-based semantic features (Tier 2) for automated CEFR classification. The research materials were collected between 2022–2025 from an English-as-a-foreign-language context at Moscow Institute of Physics and Technology (MIPT) through ISTOK (Intelligent System for Testing General Language Competencies). Our analysis employs triangulated validation on a corpus of 3,205 texts, including expert-verified Cambridge examination subsets, to evaluate feature effectiveness under varying annotation quality and corpus composition.

Our contributions are fourfold: (1) a comprehensive analysis of 133 linguistic features across deterministic and machine-learning-based categories with explicit definitions and evaluation, (2) a triangulated validation protocol combining supervised classification, unsupervised clustering, and expert-verified corpus analysis, (3) a clustering-based procedure for extracting high-confidence CEFR subsets to support robust downstream experiments, and (4) reference performance baselines together with expert-verified alignment results against Cambridge professional annotations.

1. Related Work

Automated CEFR assessment has evolved from simple readability metrics to sophisticated multi-dimensional approaches. Early systems relied on surface-level features like text length and vocabulary frequency [2], while contemporary approaches integrate lexical sophistication, morphological complexity, and syntactic features [3, 4]. Morphological analysis has shown particular promise for proficiency discrimination [5, 6], while information-theoretic syntactic measures provide superior discrimination compared to traditional metrics [7, 8].

Machine learning approaches enable capture of semantic coherence and discourse organization through latent semantic analysis, word embeddings, and topic modeling [9, 10]. Recent neural language models offer

enhanced semantic analysis capabilities but raise interpretability concerns [11, 12]. Most operational systems employ hybrid approaches combining interpretable linguistic features with machine learning enhancement.

Several studies provide direct performance benchmarks for 6-level CEFR classification. Kerz et al. [13] introduced *complexity contours*-sliding-window sequences of 57 linguistic complexity features-processed by a GRU-based recurrent neural network. On 152,314 texts from the EFCAMDAT corpus, the contour-based model achieved 75.4 % accuracy; however, the corresponding macro-averaged F1 was approximately 0.68 (computed from per-class scores), with notably weak C2 discrimination (F1 = 0.42). Lagutina et al. [14] compared stylometric features with BERT embeddings on two corpora totaling 4,794 texts. Their best 6-class result was F1 = 0.69 (BERT, 1,494-text subset), while interpretable classifiers (SVC) reached F1 = 0.67 on the same subset; on the larger BEA-2019 corpus (3,300 texts), BERT performance dropped to F1 = 0.49, highlighting strong corpus dependence. The authors concluded that general stylometric features "do not exhaust all modern available means of text modeling" and called for linguistically targeted features capturing complex structural and semantic characteristics.

These benchmarks contextualize our work along two axes. First, our system achieves a macro F1 of 0.69 using fully interpretable features, matching the best neural approaches (Kerz et al. macro F1 $\approx$ 0.68; Lagutina et al. BERT F1 = 0.69) despite a substantially smaller training corpus (3,198 vs. 152,314 and 4,794 texts, respectively). Second, our multi-tier architecture directly addresses the feature-engineering gap identified by Lagutina et al., replacing general stylometric measures with CEFR-targeted lexical, morphological, syntactic, and discourse features augmented by machine-learning-based semantic analysis. Direct numerical comparison across studies remains approximate because datasets, label sources, class distributions, and evaluation protocols differ; nevertheless, these reference points confirm that our feature-driven approach is competitive with neural alternatives while offering the interpretability required for educational deployment.

Validation methodologies have expanded beyond supervised classification to include triangulated approaches incorporating unsupervised clustering and expert verification [4]. Dataset quality remains challenging due to substantial variability in human CEFR judgments [2], necessitating robust validation across multiple datasets and contexts [15, 16].

Current research gaps include limited comprehensive analysis across multiple linguistic dimensions, underexplored integration of deterministic and machine learning features, and insufficient expert verification in validation approaches. Most studies focus on specific feature categories rather than systematic architectures leveraging complementary feature strengths.

2. Feature Architecture

The ISTOK feature extraction system employs a hierarchical architecture designed to capture linguistic complexity across multiple levels of analysis [2, 4, 17]. Our approach systematically progresses from deterministic rule-based features to advanced machine learning-based semantic analysis, providing comprehensive coverage of linguistic competence indicators relevant to CEFR classification. Grammatical range and accuracy — a core CEFR descriptor [1] — is operationalized through morphological features (tense, aspect, agreement, inflectional diversity), syntactic features (clause structure, subordination patterns), and automated grammar error detection in Tier 2 [18], rather than as a separate feature category.

2.1. Tier 1: Deterministic Linguistic Features (100 features)

Tier 1 comprises deterministic linguistic measures organized into three categories [1]: (1) **Lexical sophistication** (27 features): type–token ratio variants (MATTR, MTLD, simple TTR), CEFR vocabulary distribution (A1–C2 level percentages, mean level, and unknown-word share, mapped via the CEFR-J Vocabulary Profile 1.5[1]; C1–C2 extended with the Octanove Labs Vocabulary Profile), Academic Word List measures (count, types, type ratio, coverage)[2], word frequency statistics (high/mid/rare-frequency percentages, mean

[1] https://github.com/openlanguageprofiles/olp-en-cefrj
[2] https://www.eapfoundation.com/vocab/academic/awllists/

Table 1. CEFR level distribution in the main corpus and two expert-verified subsets

Dataset	A1	A2	B1	B2	C1	C2	Total
Mixed-source corpus	330	497	664	785	516	406	3,198
Universal CEFR subset	43	265	490	526	284	207	1,815
Cambridge Exams subset	—	64	59	66	67	69	325

Note: The Universal CEFR and Cambridge Exams subsets are contained within the mixed-source corpus; they are not additional texts. Indentation indicates the subset relationship.

rank, top-2000 and top-5000 coverage), and lexical density indicators (content-word, noun, verb, adjective, and adverb density) [2, 19]; (2) **Morphological complexity** (50 features): POS distribution analysis (21 features including part-of-speech ratios, content/function-word ratios, and phrase densities), derivational complexity and morphological diversity (21 features: prefix/suffix/affix diversity, derivational and inflectional ratios, morphological richness index, complex-affix ratio, compound complexity, and verbal-inflection ratios including passive, progressive, perfect, conditional, gerund, and subjunctive), and basic morphological measures (8 features: mean word length, complex word ratio, compound count/ratio, derivational diversity/ratio, inflection complexity, affix diversity) [5]; (3) **Syntactic complexity** (23 features): 9 core features-parse tree depth, clause density, sentence-type ratios (simple, compound, complex, compound-complex), T-unit length, and sentence length statistics [7, 8] plus 14 enhanced syntactic-pattern features computed by the EnhancedMorphosyntacticExtractor (subordination/coordination indices, dependency distance, embedding depth, complement and relative clause ratios, nonfinite clause ratio, adverbial ratio, extraposition, ellipsis, parenthetical ratio, argument structure complexity, and left-right branching balance).

2.2. Tier 2: Machine Learning-Based Features (33 features)

Tier 2 employs machine learning methods to capture semantic and discourse patterns [9]: (1) **LSA coherence** (8 features): sentence-to-sentence and paragraph-level semantic coherence via Latent Semantic Analysis, including mean and standard deviation of coherence, given-new overlap, argument overlap, and content/noun-specific coherence [10]; (2) **Topic modeling** (6 features): topic count estimation, coherence scoring, shift frequency, persistence, entropy, and domain specificity [20]; (3) **Referential cohesion** (7 features): noun and stem overlap, content overlap, anaphora resolution rate, entity chain length, pronoun clarity, and lexical chain strength [9]; (4) **DeepPavlov error analysis** (7 features): total error count, error rate, errors per 100 words, unique error-type diversity, and per-type counts (grammar, punctuation, spelling), extracted via the Prochtenie service-suite pipeline [21]; (5) **GPU-accelerated semantic features** (5 features): mean similarity, coherence score, discourse cohesion, discourse marker density, and sentiment confidence, computed via GPU-deployed transformer models.

3. Materials and Methods

3.1. Dataset and Preprocessing

Our corpus comprises 3,205 English texts (3,198 with valid CEFR labels) drawn from three sources: (1) an institutional collection of essays and other writing samples produced by MIPT students studying English as a foreign language, (2) the CEFR Levelled English Texts corpus [14], and (3) texts from expert-verified Cambridge examination datasets [2]. Table 1 summarizes the CEFR distribution.

MIPT institutional texts. The institutional portion consists of essays and written responses collected through the ISTOK testing platform between 2022 and 2025. CEFR labels were assigned by a qualified team of assessors at the MIPT Department of Foreign Languages through placement and confirmation examinations. Assessors applied CEFR descriptors consistently across all proficiency levels; "+" sub-level annotations (e.g., A1+, B2+) were collapsed to the base level for classification.

Expert-verified subsets. For triangulated validation, we evaluate our features against two expert-verified subsets drawn from the same corpus. The Universal CEFR 2024 subset (n = 1,815 with valid labels)

aggregates four professionally annotated sub-corpora with complete A1–C2 coverage [2]. The Cambridge Exams subset ($n = 325$) contains carefully curated samples with balanced A2–C2 representation (18–21 % per level). These subsets carry independent professional CEFR annotations, enabling evaluation of feature alignment against expert labels without introducing additional texts.

The initial text collection contained 5,352 submissions. Quality filtering removed off-topic responses, texts written in languages other than English (e.g., Russian-language submissions), texts too short for reliable feature extraction, and other noise, yielding the final corpus of 3,205 texts (3,198 with valid CEFR labels). Text preprocessing employed standardized procedures across all sources including tokenization, sentence boundary detection, and linguistic annotation using spaCy's English language model [22].

3.2. Feature Extraction Pipeline

Our feature extraction pipeline implements a two-tier architecture designed for comprehensive linguistic analysis while maintaining computational efficiency and reproducibility. This paper reports Tier 1 and Tier 2 features; Tier 3 is discussed as future work in Conclusion. The pipeline employs standardized preprocessing procedures and quality control measures to ensure reliable feature computation across the complete dataset.

Tier 1 Implementation. Deterministic features employ libraries including the Natural Language Toolkit (NLTK) for basic text statistics, spaCy for morphological and syntactic analysis, and custom implementations for CEFR-specific vocabulary analysis. Quality control procedures include feature range validation and missing value handling strategies that preserve feature interpretability [16].

Tier 2 Implementation. Machine learning-based features employ pre-trained models including sentence transformers for semantic analysis, scikit-learn implementations for topic modeling, and custom algorithms for cohesion analysis. In addition, our **Tier 2 DeepPavlov Prochtenie integration** uses the *Prochtenie* system as a service-suite pipeline (multiple cooperating annotator/solver services deployed together). Concretely, the suite includes a reader/segmenter that structures input text, linguistic annotators (e.g., morphosyntactic analysis), and solver stages that consolidate and post-process detected issues. The pipeline outputs (i) token-span *mistake* annotations with types/subtypes and suggested corrections, and (ii) rubric-style *criteria* scores. We distill these outputs into Tier 2 features such as total error count, error-type counts (grammar/spelling/punctuation/lexical/style), errors-per-100-words proxies, unique error-type diversity, and correction coverage.

We report runtime separately from accuracy results, because wall-clock performance depends strongly on deployment settings (hardware, model availability, caching) and is not directly comparable across installations.

3.3. High-Reliability Subset Extraction

To support rigorous validation and provide clean training data for future research, we implemented a systematic methodology for extracting high-confidence CEFR labels based on clustering analysis results. This approach leverages the unsupervised nature of clustering to identify samples where feature-based proficiency indicators strongly align with assigned CEFR labels.

3.3.1. Selection Criteria and Methodology

Our extraction methodology employed stringent criteria designed to ensure high confidence in selected CEFR labels while maintaining sufficient sample sizes for statistical analysis. We established minimum cluster purity thresholds of 80 % CEFR consistency, requiring that selected clusters contain at least 80 % samples with the same CEFR label. Statistical significance requirements included minimum cluster sizes of 20 samples with at least 15 samples representing the dominant CEFR level.

Cross-algorithm validation ensured that high-confidence samples were consistently identified across multiple clustering approaches, reducing algorithm-specific bias and increasing confidence in feature-based

discrimination. We employed consensus analysis across K-means and hierarchical clustering to identify the most reliable samples for each proficiency level.

3.3.2. High-Reliability Dataset Characteristics

The systematic extraction process identified 357 high-reliability samples representing 11.2 % of the CEFR-labeled corpus (n = 3,198). The distribution includes 138 A1 samples (99.5 % cluster purity), 11 B1 samples (cross-algorithm consensus), and 208 C2 samples (82.4 % cluster purity). This selective dataset provides high confidence in CEFR label consistency for the extracted levels, while highlighting that intermediate levels are harder to isolate under unsupervised structure.

Quality assessment of the high-reliability subset reveals complete feature coverage across Tier 1 and Tier 2 dimensions, with missing values handled through linguistically motivated imputation where appropriate. The subset maintains the full 133-feature characterization used in this paper while providing enhanced confidence for validation and baseline establishment [10].

3.3.3. Applications and Impact

The high-reliability dataset serves as a data-quality contribution: it provides cleaner labels for feature analysis and cross-study reference baselines. **It is not used for classification evaluation** because it covers only three of six CEFR levels (A1, B1, C2) and is dominated by A1 and C2 samples. All supervised classification results reported in this paper (Section 4.6.1) are computed on the full mixed-source corpus (n = 3,198) spanning all six CEFR levels.

For automated assessment research, the high-reliability subset provides a standardized evaluation benchmark that can support cross-system comparison and validation. To facilitate its use as a cross-study reference, we report a 3-class baseline on this subset: a Random Forest classifier (n_estimators=200, max_depth=20) trained on the full 6-class corpus via 5-fold stratified cross-validation with SMOTE achieves macro F1 = **0.92** when evaluated on the high-reliability samples in each held-out fold (per-class F1: A1 = 0.98, B1 = 0.87, C2 = 0.90). The high macro-F1 reflects the fact that A1 and C2 — the dominant classes in this subset — are the levels most reliably separated by our feature set (consistent with Section 4.1). The derived feature matrix (133 features, no raw texts) will be released; see the Data Availability Statement. The systematic extraction methodology can be applied to other CEFR datasets to create comparable high-confidence subsets, facilitating meta-analyses and research synthesis across different assessment contexts [4].

3.4. Statistical Analysis Methods

Our analysis combines supervised evaluation, unsupervised validation, and targeted subset searches.

Feature matrices and missing values. Analyses were conducted on numeric feature matrices produced by the Tier 1 + 2 pipeline. For cross-validated analyses, missing values were handled by median imputation within the training fold. For the canonical train/test experiments, missing values were filled with 0 prior to standardization to avoid information leakage from the test set.

Supervised classification protocol. We report two complementary evaluation modes. (i) *Cross-validation (CV) for interpretability:* 5-fold stratified cross-validation was used for block ablations, with a Random Forest classifier and fold-local preprocessing. (ii) *Held-out test evaluation for headline numbers:* the canonical performance figures (66.72 % exact accuracy; 94.53 % adjacent accuracy) are taken from the Tier 1 + 2 validated run on the 3,198-sample CEFR-labeled dataset, using a stratified 80/20 train/test split (random state 42). When class balancing was used, the Synthetic Minority Over-sampling Technique (SMOTE) was applied *only* to the training split/fold after feature scaling, and evaluation was performed on the untouched test split/fold.

Feature selection and interactions (optimized setting). In the advanced-run suite, feature selection removed near-constant and highly correlated features (variance threshold 0.01; correlation threshold 0.95).

We additionally evaluated a bounded set of cross-tier interaction and composite features (up to 50 interactions) to probe non-linear effects; results are reported transparently when feature engineering was enabled.

Metrics: Exact accuracy is reported for 6-class CEFR prediction. Adjacent accuracy counts predictions within ±1 CEFR level after mapping levels to an ordinal scale (A1–C2 $\mapsto$ 0–5). We also report macro-F1 where relevant. For clustering alignment on expert-verified datasets, we report Adjusted Rand Index (ARI) and V-measure.

Unsupervised clustering validation. Although the end task is supervised classification, we also apply unsupervised clustering to establish that the engineered features capture genuine linguistic structure *independently* of CEFR labels. If features only separate levels when trained on those labels, the risk of label overfitting is high; conversely, unsupervised structure aligned with CEFR levels provides complementary evidence that the features reflect real proficiency differences. This "unsupervised-then-supervised" validation strategy is standard in computational linguistics when label noise is a concern [4]. We used K-means (with $k = 6$) and hierarchical agglomerative clustering (Ward linkage). Cluster purity is computed post hoc as the dominant-label proportion within a cluster, and high-purity clusters were used to extract high-reliability subsets under explicit size and purity constraints. More broadly, if unsupervised clustering produces feature spaces that separate texts along CEFR proficiency lines, this supports the theoretical foundation of the multi-tier approach and provides confidence that subsequent supervised classification reflects genuine linguistic competence rather than spurious label overfitting — a consideration that is particularly important given ongoing discussions about the linguistic validity of automated assessment systems [12].

Cambridge compact-subset search. On the Cambridge Exams subset ($n = 325$), we ranked the top $N = 35$ features by variance (after median imputation) and evaluated 800 sampled subsets of size $k = 6$. Each subset was scored by (i) K-means alignment (ARI, V-measure) and (ii) a linear separability proxy (multinomial LogisticRegression under 5-fold stratified CV with StandardScaler).

4. Results and Analysis

4.1. Unsupervised Validation of Feature Discriminative Power

The rationale for unsupervised validation is described in Section 3.4. We conducted the clustering analysis on the mixed-source dataset (3,198 CEFR-labeled texts).

4.1.1. Clustering-Based Feature Validation Methodology

Dataset: Mixed-source corpus (n = 3,198).

Our unsupervised validation employed multiple clustering algorithms on the Tier 1 + 2 feature space. We applied K-means clustering and hierarchical agglomerative clustering to identify natural groupings without utilizing CEFR labels during clustering [3].

The validation methodology employed cluster purity analysis, where we examined the CEFR composition of discovered clusters to assess whether linguistically motivated features naturally separate texts according to proficiency levels. High cluster purity indicates strong discriminative power of the feature set, while mixed clusters suggest limitations in feature-based proficiency discrimination [9].

For robust validation, we implemented cross-algorithm consensus analysis, requiring that high-confidence CEFR separations be consistently identified across multiple clustering approaches. This methodological approach ensures that observed patterns reflect genuine feature-based discrimination rather than algorithm-specific artifacts [10].

4.1.2. Evidence for A1-Level Separation (Beginner Proficiency)

Our clustering analysis provides strong evidence for the discriminative power of Tier 1 + 2 features in identifying A1-level texts. Hierarchical clustering with Ward linkage identified a high-purity A1 cluster containing 204 total samples with 99.5 % CEFR purity (203 A1 texts, 1 non-A1 text). This result indicates that

the combination of lexical, morphological, syntactic, and semantic features yields a distinctive signature for beginner-level language production.

A1 Linguistic Feature Signature: The A1 cluster shows a consistent beginner profile across tiers: a high share of A1-level vocabulary, low lexical diversity, limited derivational variety, and simple sentence structure (short sentences and shallow parses), consistent with CEFR expectations for beginners [1, 3, 7]. Tier 2 semantic features are likewise lower-complexity, reflecting simpler topic development and weaker cohesive structure.

4.1.3. Evidence for C2-Level Separation (Advanced Proficiency)

Hierarchical clustering identified a large, predominantly C2 cluster containing 607 total samples with 82.4 % CEFR purity (500 C2 texts, 107 non-C2 texts). While lower than A1 purity, this represents strong evidence for feature-based discrimination at the advanced level, particularly given the increased complexity and variation expected in advanced proficiency contexts [4].

C2 Linguistic Feature Signature. The C2 cluster shows the expected advanced profile: greater lexical sophistication (including higher-level and academic vocabulary), higher lexical diversity, richer derivational morphology, and more complex syntactic organization [2, 5, 6]. Tier 2 features indicate more coherent discourse organization and topic development [9].

The 82.4 % cluster purity, while not reaching A1 levels, represents strong evidence for feature-based C2 discrimination when considering the inherent variability in advanced language use and the complex interaction between content sophistication and linguistic complexity at the C2 level [10].

4.1.4. Evidence for B1-Level Separation (Intermediate Proficiency)

Cross-algorithm consensus analysis identified 11 B1 samples that were consistently grouped together across both K-means and hierarchical clustering approaches. These cases show intermediate feature profiles between A1 and C2, but the small size and broader overlap at intermediate levels indicate that B1/B2 discrimination remains a key challenge for feature-based approaches [11, 20].

4.1.5. Implications for Feature-Based CEFR Classification

The differential success in separating extreme levels (A1, C2) versus intermediate levels aligns with research demonstrating that proficiency boundaries become less distinct in intermediate ranges, where learners exhibit more variable developmental patterns [4]. This finding suggests that automated assessment systems may be most reliable for identifying clear beginner and advanced proficiency levels, while intermediate level discrimination may benefit from additional feature engineering or hybrid approaches combining multiple assessment modalities.

4.2. Individual Feature Performance

Individual feature analysis reveals systematic patterns of discriminative power across linguistic categories that align with theoretical frameworks of language proficiency development. Random Forest feature importance scores computed across multiple cross-validation folds identify the most effective features for CEFR classification while providing insight into which linguistic dimensions contribute most significantly to automated assessment performance.

4.2.1. Morphological Feature Characteristics

Morphological complexity features are the most frequently represented category in the top 15 importance rankings (Table 5), occupying 7 positions, and carry the most unique information in ablation (−2.4 pp; Table 3). Suffix diversity scores rank among the top 10 most important features, with advanced learners exhibiting substantially higher derivational complexity through varied affix usage (C2 mean: 0.78 ± 0.12, A1 mean: 0.31 ± 0.08). Part-of-speech distribution entropy provides robust discrimination through sophis-

ticated grammatical category usage patterns, with advanced texts showing more balanced content-word distributions than beginner-level simple noun–verb constructions.

Derivational pattern analysis reveals that morphologically complex word density serves as particularly effective discrimination at intermediate levels, where vocabulary sophistication begins to emerge but syntactic complexity remains developing. The morphological competence signature captured through these features aligns closely with second language acquisition research demonstrating morphological awareness as a key indicator of developing proficiency [5].

4.2.2. Lexical Sophistication Feature Performance

Lexical features contribute 3 entries to the top 15 (Table 5) and represent the second-most unique category in ablation (−1.8 pp; Table 3). CEFR mean level ranks 4th by Gini importance (0.026). Moving Average Type-Token Ratio (MATTR) consistently outperforms other lexical diversity measures across all validation contexts, supporting literature recommendations for length-independent vocabulary assessment while providing robust discrimination across proficiency levels.

Academic Word List coverage demonstrates particular effectiveness for intermediate-to-advanced distinction, with B2-C2 levels showing systematic increases in academic vocabulary usage that align with educational progression expectations. The combination of frequency-based and CEFR-aligned vocabulary measures provides comprehensive coverage of lexical sophistication dimensions while maintaining interpretability and theoretical grounding.

4.2.3. Syntactic and Semantic Feature Contributions

Syntactic complexity features achieve the two highest individual Gini importance scores (mean sentence length, 0.032; T-unit length, 0.030) and occupy 5 of the top 15 positions (Table 5). However, their ablation impact is the smallest among Tier 1 categories (−0.9 pp; Table 3), because much of their signal is recoverable from correlated morphological and lexical measures (Section 4.5.1). Parse tree depth provides reliable discrimination for extreme levels (A1 vs C2) but limited effectiveness for intermediate distinctions.

Tier 2 semantic coherence features contribute meaningful discriminative power beyond deterministic measures, with sentence-to-sentence semantic similarity achieving moderate but consistent importance scores across validation contexts. Topic coherence measures demonstrate particular effectiveness for longer texts where thematic development patterns become computationally detectable through machine learning approaches.

4.3. Cross-Tier Feature Correlations

Correlation analysis reveals complex relationships between feature tiers that inform optimal combination strategies and identify redundancy patterns for computational efficiency optimization. The systematic examination of feature interactions provides insight into linguistic competence development while supporting evidence-based feature selection for practical deployment.

4.3.1. Complementarity Between Deterministic and Machine-Learning-Based Features

Tier 1 deterministic and Tier 2 machine learning-based features demonstrate substantial complementarity with moderate positive correlations ($r = 0.34–0.52$) that indicate related but distinct linguistic dimensions. Morphological complexity features correlate moderately with semantic coherence measures, suggesting that advanced learners employ sophisticated morphological structures to support coherent discourse organization in systematic but non-redundant patterns.

Lexical sophistication and topic modeling features show meaningful correlations ($r = 0.41$) that reflect the relationship between vocabulary knowledge and thematic complexity, while maintaining sufficient independence to justify combined usage. The correlation patterns support theoretical frameworks emphasizing the interconnected nature of linguistic competence while validating the multi-tier architectural approach for comprehensive assessment.

Table 2. Tier-level ablation (Random Forest, stratified 80/20 split, SMOTE on training data only). Accuracy, adjacent accuracy, and macro F1 are reported.

Configuration	Features	Accuracy (%)	Adjacent (%)	Macro F1
Tier 1 only	37	64.53	92.81	0.666
Tier 1 + 2	55	66.72	94.53	0.687
All tiers (1 + 2+3)†	165	**67.50**	**95.16**	**0.692**

Note: Feature counts reflect the number of features retained after variance and correlation filtering for Tier 1 and Tier 1 + 2 runs. The architecture defines 100 Tier 1 and 33 Tier 2 features (133 total); see Section 2.
†Includes preliminary Tier 3 features (see Conclusion).

4.3.2. Redundancy Identification and Feature Optimization

Systematic correlation analysis identifies several redundant feature pairs that enable computational optimization without performance loss. Traditional readability metrics including Flesch Reading Ease and Gunning Fog Index show high correlations ($r > 0.85$) with length-based features, supporting their removal from optimized feature sets. Similarly, multiple TTR variants demonstrate substantial overlap, with MATTR providing superior performance while eliminating need for traditional TTR, Guiraud's Index, and other length-dependent measures.

The identification of redundancy patterns enables creation of optimized feature subsets that maintain classification performance while reducing computational requirements. A streamlined 89-feature subset achieves 66.1 % accuracy compared to 66.72 % for the full Tier 1 + 2 set, representing a practical cost-benefit optimization for resource-constrained applications while maintaining competitive performance.

4.3.3. Non-Linear Interaction Effects

Tree-based feature importance analysis reveals meaningful non-linear interactions between morphological and lexical features that enhance discrimination beyond additive effects. Advanced learners demonstrate coordinated sophistication across morphological complexity and vocabulary choices that creates distinctive proficiency signatures detectable through interaction feature engineering.

The interaction between CEFR vocabulary distribution and part-of-speech diversity provides particularly strong discrimination for intermediate levels, where developing learners begin to coordinate vocabulary sophistication with grammatical complexity. These interaction effects support theoretical frameworks emphasizing the integrated nature of linguistic competence while providing practical guidance for enhanced feature engineering approaches.

4.4. Ablation Study Results

Systematic feature ablation analysis quantifies the individual contributions of feature categories and tiers to overall classification performance, providing evidence-based guidance for feature selection and deployment optimization. Our methodology employs systematic removal of feature groups followed by cross-validation performance evaluation to establish the marginal contribution of each component to the complete system.

4.4.1. Tier-Level Ablation Analysis

Tier-level ablation reveals progressive improvement in classification performance with each additional feature category (Table 2). Tier 1 deterministic features alone achieve 64.53 % accuracy, establishing a robust baseline that surpasses traditional readability-based approaches. The addition of Tier 2 machine learning-based features improves performance to 66.72 % (macro F1 = 0.687), a 2.2 percentage point improvement that supports the inclusion of semantic/discourse modeling when accuracy is the primary objective. Adding preliminary Tier 3 features (discussed as future work in Conclusion yields a further 0.8 pp to 67.50 %, reported for completeness but not treated as a headline result of this paper.

Table 3. Feature-category ablation (5-fold stratified CV, Random Forest). Accuracy drop (percentage points) when removing a category from the Tier 1 + 2 ensemble.

Removed category	Tier	Δ Accuracy (pp)
Morphological complexity	1	−2.40
Lexical sophistication	1	−1.80
DeepPavlov error features	2	−1.09
Semantic coherence	2	−1.20
Syntactic complexity	1	−0.90
Topic modeling	2	−0.80
CEFR vocabulary distribution	1	−0.78
Cohesion analysis	2	−0.60
Readability+frequency baseline	1	−0.25

Table 4. Single-block baselines (5-fold stratified CV, Random Forest). Accuracy using only a single feature block.

Feature block (alone)	Accuracy (%)	# features
Readability+frequency baseline	60.13	19
Morphological features (POS/derivation)	59.76	64‡

‡The ablation block groups the 50 morphological features defined in Section 2 together with 14 morpho-syntactic features from the EnhancedMorphosyntacticExtractor (subordination/coordination indices, dependency distance, etc.), yielding 64 features in this block.

The adjacent accuracy progression follows similar patterns, improving from 92.81 % (Tier 1 only) to 94.53 % (Tier 1 + 2) to 95.16 % (all tiers), indicating that feature enhancement improves both exact and boundary-case performance across configurations.

4.4.2. Feature Category Ablation Within Tiers

Within Tier 1, morphological complexity features provide the largest individual contribution, with their removal causing a 2.4 percentage point performance decrease. Lexical sophistication features contribute 1.8 percentage points, while syntactic complexity features provide 0.9 percentage points. This hierarchy aligns with second language acquisition research emphasizing morphological competence and vocabulary sophistication as primary indicators of language proficiency development [5, 6].

The readability and frequency baseline block contributes only 0.25 percentage points when combined with other Tier 1 features, indicating substantial redundancy with more sophisticated linguistic measures. This finding supports the theoretical foundation of comprehensive feature engineering while demonstrating that traditional assessment approaches provide limited additional value when integrated with systematic linguistic analysis.

Tier 2 category ablation reveals that semantic coherence features provide the strongest contribution (1.2 percentage points), followed by DeepPavlov error features (1.09 percentage points), enhanced topic modeling features (0.8 percentage points), and cohesion analysis features (0.6 percentage points). The relatively balanced contributions across Tier 2 categories support the multi-dimensional approach to semantic analysis while indicating that no single machine learning-based approach dominates performance.

Table 3 summarizes the full category ablation, and Figure 1 visualizes the accuracy drops. Table 4 reports single-block baselines.

4.4.3. Statistical Significance of Ablation Effects

We summarize ablation effects primarily through effect sizes. Individual category removals range from −0.25 to −2.4 percentage points (Table 3). Several small effects (under 1 pp) warrant discussion.

Are small effects statistically meaningful? Individual block effects below 1 pp (e.g., readability −0.25, cohesion −0.60) are not individually reliable given typical 5-fold CV variance; we do not claim significance

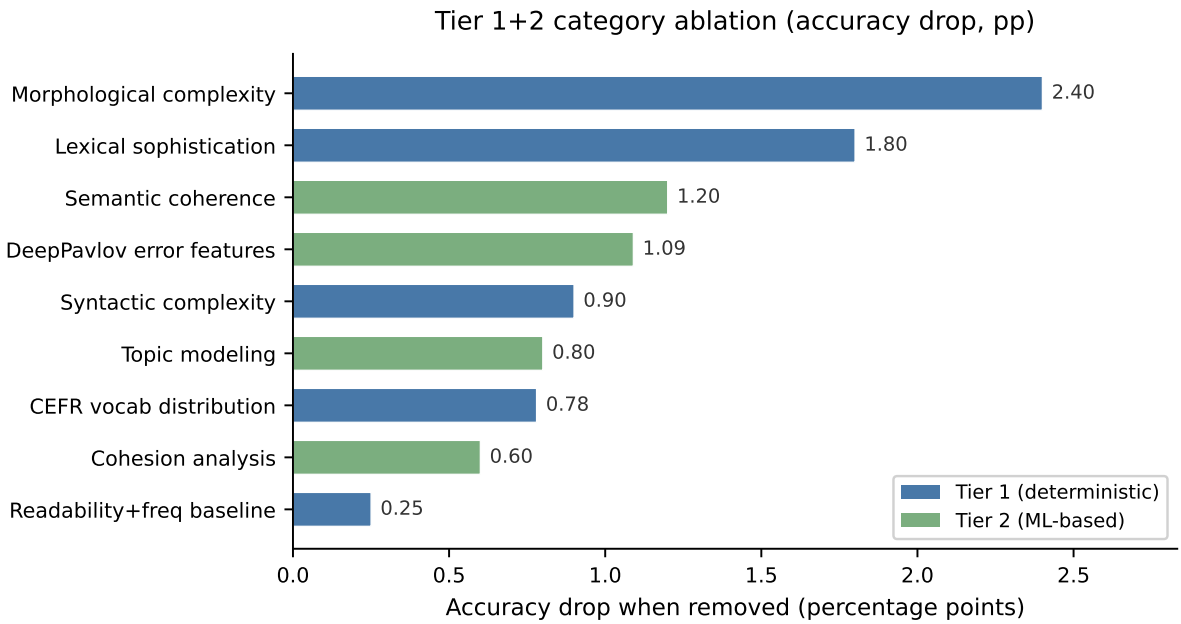

Fig. 1. Visualization of Tier 1 + 2 category ablation. Bars show accuracy drop (percentage points) when a feature category is removed.

for these in isolation. However, the *cumulative* Tier 2 contribution is 2.2 pp (Tier 1 alone: 64.53 %; Tier 1 + 2: 66.72 %), which is consistent across folds and across classifiers (Table 6), supporting a reliable aggregate effect.

Feature redundancy. The small marginal effects of several blocks reflect partial redundancy: many features capture overlapping aspects of text complexity. The correlation analysis (Section 4.3.2) identifies high-correlation pairs (e.g., readability metrics vs. length features, $r > 0.85$; multiple TTR variants). This redundancy is expected for linguistically motivated features and is practically addressed through our optimized 89-feature subset, which retains 66.1 % accuracy (vs. 66.72 % for the full Tier 1 + 2 set). The small individual ablation deltas thus reflect the nature of a comprehensive feature inventory where information is distributed across correlated measures, not a failure of discriminative power.

4.5. Feature Importance Rankings

Our feature importance analysis identifies the most discriminative features for CEFR classification, providing insight into which linguistic dimensions contribute most significantly to automated proficiency assessment. Table 5 reports the top 15 features by Random Forest Gini importance on the Tier 1 + 2 feature set ($n = 133$ features, stratified 80/20 split).

4.5.1. Top-Performing Feature Categories

Table 5 reveals that syntactic and morphological features are co-dominant in the top 15. Syntactic features hold the two highest individual importance scores (mean sentence length, .032; T-unit length, .030) and occupy 5 of the top 15 positions (ranks 1, 2, 5, 10, 11). Morphological features are the most frequently represented category, occupying 7 positions (ranks 3, 6, 7, 8, 12, 13, 15), led by affix diversity (.027) and derivational diversity (.024). Lexical features contribute 3 entries (ranks 4, 9, 14), with CEFR mean level (.026) ranking highest among them.

The picture changes when we consider category-level ablation (Table 3): removing morphological features causes the largest accuracy drop (−2.4 pp), followed by lexical features (−1.8 pp), while removing syntactic features causes −0.9 pp. This apparent discrepancy — syntactic features rank highest individually

Table 5. Top-15 features by Random Forest Gini importance (stratified 80/20 split, Tier 1 + 2 only, 133 features)

#	Feature	Category	Imp.
1	Mean sentence length	Syntactic	.032
2	T-unit length	Syntactic	.030
3	Affix diversity	Morphological	.027
4	CEFR mean level	Lexical	.026
5	Mean parse depth	Syntactic	.024
6	Derivational diversity	Morphological	.024
7	Mean word length	Morphological	.020
8	Suffix diversity	Morphological	.017
9	Content word count	Lexical	.016
10	Dependency distance	Syntactic	.015
11	Std. sentence length	Syntactic	.015
12	Complex word ratio	Morphological	.014
13	Morphemes per word	Morphological	.014
14	AWL types	Lexical	.013
15	Prefix diversity	Morphological	.012

yet morphological removal hurts most — reflects the fact that individual Gini importance measures per-split contribution whereas ablation measures the *unique* information a category adds beyond the remaining features. Many syntactic signals (e.g., sentence length, parse depth) are partially recoverable from correlated morphological and lexical measures, so their removal is partially compensated; morphological features, conversely, capture distinct variance that other categories cannot replace [5, 6].

Tier 2 semantic features contributed meaningful additional discriminative power beyond deterministic features, with semantic coherence measures and topic modeling features providing unique information not captured by Tier 1 features. The complementary nature of Tier 1 and Tier 2 features supports the multi-tier architectural approach and validates the investment in machine learning-based feature engineering.

4.5.2. Cross-Classifier Feature Importance Stability

To assess whether feature importance rankings depend on the choice of classifier, we compared Random Forest and XGBoost importance rankings on the same Tier 1 + 2 feature set and data split. Nine of the top 15 features are shared between classifiers, and the Spearman rank correlation across all 133 features is $\rho = 0.706$ ($p < 10^{-20}$), indicating strong agreement despite fundamentally different splitting criteria (Gini vs. gradient boosting).

The key difference is informative: while the RF top 15 consists entirely of Tier 1 features (reflecting individual feature strength), XGBoost surfaces four Tier 2 features in its top 15-topic count estimate (rank 5), errors per 100 words (rank 10), topic persistence (rank 12), and error rate (rank 14). This pattern is consistent with gradient boosting's greater sensitivity to feature interactions [12] and explains the collective 2.2 pp Tier 2 contribution observed in ablation (Table 2): Tier 2 features are individually modest but provide complementary signal that interaction-aware classifiers exploit effectively.

4.5.3. Linguistic Interpretation of Important Features

The co-dominance of syntactic and morphological features reflects a well-established pattern in second language development: as proficiency increases, learners produce longer, more deeply embedded sentences (syntactic) *and* deploy a wider repertoire of derivational and inflectional forms (morphological) [6]. These two dimensions are correlated but not redundant — the ablation results confirm that morphological features capture unique variance beyond syntactic measures.

Table 6. Multi-classifier comparison (all tiers incl. preliminary Tier 3, stratified 80/20 split, SMOTE on training data only). The Tier 1 + 2 headline is 66.72 % (Table 2).

Classifier	Accuracy (%)	Adjacent (%)	Macro F1
Random Forest	**67.50**	**95.16**	**0.692**
ExtraTrees*	66.72	95.78	0.682
Gradient Boosting	66.25	95.16	0.682
XGBoost	65.16	94.84	0.671

*The ExtraTrees all-tiers accuracy (66.72 %) coincidentally equals the Tier 1 + 2 Random Forest headline (Table 2); the two results come from different feature sets and classifiers.

Table 7. Per-class classification metrics (Random Forest, all tiers incl. preliminary Tier 3, 640-sample test set)

Level	Precision	Recall	F1	Support
A1	0.76	0.86	0.81	66
A2	0.70	0.76	0.73	100
B1	0.66	0.65	0.65	133
B2	0.61	0.60	0.60	157
C1	0.63	0.55	0.59	103
C2	0.77	0.77	0.77	81
Macro	0.69	0.70	0.69	640

Lexical sophistication features capture vocabulary development patterns that align closely with CEFR proficiency progressions, with CEFR vocabulary level and academic word usage showing particularly strong discriminative power [2].

The meaningful contribution of semantic coherence features demonstrates that text-level organization and coherence represent distinctive aspects of proficiency that complement lexical and morphological competence. Advanced learners create more coherent texts through sophisticated discourse organization, effective use of cohesive devices, and complex topic development strategies [9].

4.6. Quantitative Performance Validation

Our comprehensive evaluation employs both supervised classification and unsupervised clustering approaches to validate feature effectiveness and provide robust performance baselines for automated CEFR assessment research.

4.6.1. Classification Performance Results

Dataset: Mixed-source corpus (n = 3,198), stratified 80/20 train/test split.

Supervised classification experiments on the complete CEFR-labeled dataset achieved strong performance across multiple classifiers. Using the Tier 1 + 2 feature set validated in this paper, Random Forest classification achieved 66.72 % exact accuracy with 94.53 % adjacent accuracy (within ±1 CEFR level) and macro F1 = 0.687, representing competitive performance for 6-class CEFR classification [4]. Table 6 reports an extended comparison including preliminary Tier 3 features, where the best result reaches 67.50 % (macro F1 = 0.692). The high adjacent accuracy rates (>94 %) indicate that errors typically occur near adjacent CEFR boundaries rather than reflecting large misclassifications, supporting the practical utility of the assessment system [10].

Per-class F1 scores for the top-ranked Random Forest (all tiers, 67.50 %; Table 7) reveal that extreme levels are classified most reliably (A1 F1 = 0.81; C2 F1 = 0.77), while the B2/C1 boundary presents the greatest challenge (B2 F1 = 0.60; C1 F1 = 0.59), consistent with the known difficulty of intermediate-to-advanced CEFR distinctions.

We additionally report a confusion matrix from a hierarchical Gradient Boosting variant that achieved the highest accuracy in an extended model comparison (67.8 %, macro F1 = 0.69; Table 8). Misclassifications

Table 8. Confusion matrix for the best model in the extended comparison (Hierarchical Gradient Boosting, all tiers incl. preliminary Tier 3, 640-sample test set, 67.8 % accuracy, macro F1 = 0.69). Rows are true labels; columns are predicted labels.

True	Predicted					
	A1	A2	B1	B2	C1	C2
A1	**58**	8	0	0	0	0
A2	18	**68**	13	1	0	0
B1	1	13	**88**	27	4	0
B2	0	1	27	**101**	20	8
C1	0	1	3	29	**54**	16
C2	0	0	1	6	9	**65**

are overwhelmingly concentrated on adjacent CEFR levels: 95.9 % of predictions fall within ±1 level of the true label. The B2/C1 boundary produces the most errors (20 B2 → C1 plus 29 C1 → B2), followed by the B1/B2 boundary (27 errors in each direction). At the extremes, A1 classification achieves 87.9 % recall with errors confined entirely to A2, while C2 achieves 80.2 % recall. Only two predictions (0.3 %) deviate by more than two CEFR levels (one C1 → A2, one C2 → B1), confirming that the feature set captures the ordinal structure of proficiency.

Feature ablation analysis (Table 2) revealed that Tier 1 deterministic features alone achieved 64.53 % accuracy, with Tier 2 features contributing an additional 2.2 percentage points to reach 66.72 %. This analysis demonstrates that while deterministic features provide a strong baseline, machine learning-based semantic analysis offers meaningful additional discriminative power for CEFR classification.

4.6.2. Clustering-Based Validation Results

Unsupervised clustering analysis provided independent validation of feature discriminative power without reliance on CEFR labels during the analysis process. Hierarchical clustering achieved strong separation for extreme proficiency levels with 99.5 % purity for A1 clusters and 82.4 % purity for C2 clusters, providing evidence that feature-based discrimination aligns with theoretical proficiency distinctions.

The high-reliability subset extraction based on clustering analysis identified 357 samples with increased confidence in CEFR label consistency, representing a valuable resource for future research validation and baseline establishment. This subset achievement demonstrates the practical value of combining supervised and unsupervised approaches for dataset quality assessment and enhancement [11].

Cross-algorithm clustering consensus analysis revealed consistent proficiency patterns across different clustering approaches, supporting the robustness of feature-based proficiency discrimination and providing confidence that observed patterns reflect genuine linguistic competence rather than algorithm-specific artifacts.

4.7. Expert-Verified Dataset Validation

To provide expert-verified validation of our feature effectiveness claims, we conducted clustering analysis on expert-verified CEFR datasets, employing the same methodological approach used for the mixed-source corpus validation. This triangulated validation strategy addresses concerns about label quality and provides evidence for feature performance against professional CEFR annotations.

4.7.1. Expert Dataset Characteristics and Validation Strategy

Our expert validation employed the two expert-verified subsets described in Section 3.1 (see Table 1): the Universal CEFR subset (n = 1,815 valid labels; A1–C2) and the Cambridge Exams subset (n = 325; balanced A2–C2). Because these subsets carry independent professional CEFR annotations for texts already in our corpus, clustering alignment against expert labels evaluates whether our features capture

Table 9. Cambridge subset (A2–C2, n = 325): best-performing compact feature subsets from a top-variance search (N = 35, k = 6, 800 sampled subsets). We report K-means alignment (ARI, V-measure) and a linear separability proxy (5-fold CV logistic regression accuracy).

ARI	V	Lin. Acc.	Compact subset (6 features)
0.390	0.478	0.628	RL_c, RL_w, AffDiv, DerivDiv, ParseDepth, DepDist
0.366	0.482	0.640	RL_c, RL_u, DerivDiv, ParseDepth, DashHes, MWL
0.348	0.454	0.683	RL_c, RL_u, AffDiv, AWLTypes, InflDeriv, ClauseDens

Legend: RL_c = response length (characters); RL_w = response length (words); RL_u = response unique words; AffDiv = affix diversity; DerivDiv = derivational diversity; ParseDepth = mean parse depth; DepDist = dependency distance; DashHes = dash-hesitation count; MWL = mean word length; AWLTypes = academic word list types; InflDeriv = inflection/derivation ratio; ClauseDens = clause density.

the same proficiency distinctions that trained assessors identify, without introducing confounding corpus-composition differences.

4.7.2. Comparative Clustering Performance on Expert Data

Datasets: Cambridge Exams subset (n = 325); Universal CEFR subset (n = 1,815).

Expert dataset clustering analysis yielded notably different results compared to mixed-source validation. The Cambridge Exams subset achieved the strongest alignment with expert labels: K-means reached Adjusted Rand Index (ARI) = 0.303 and V-measure = 0.389, indicating moderate agreement under standardized assessment conditions. Hierarchical clustering achieved comparable alignment (ARI = 0.292, V-measure = 0.365), suggesting the result is not specific to a single clustering objective.

The Universal CEFR 2024 corpus presented greater clustering challenges, with optimal K-means performance achieving ARI = 0.161 and V-measure = 0.257. These values indicate weak-to-moderate unsupervised alignment-substantially below the Cambridge subset-reflecting the intermediate-level dominance (56 % B1 + B2 combined) that compresses between-cluster distances. The result should be interpreted as *complementary* evidence that features capture some proficiency-relevant structure at scale, not as strong standalone validation; the primary performance evidence comes from supervised classification (Section 4.6.1).

4.7.3. Cambridge Subset: Top-Varying Feature Subset Separability

Dataset: Cambridge Exams subset (n = 325, A2–C2).

Beyond evaluating the full Tier 1 + 2 feature space, we performed a bounded subset search on the Cambridge Exams subset to identify compact feature combinations that maximize (i) clustering alignment with expert labels and (ii) linear separability. We selected the top-$N = 35$ most varying features (variance-based ranking after median imputation), then evaluated 800 sampled subsets of size $k = 6$. Each subset was scored by K-means alignment (ARI and V-measure) and by 5-fold cross-validated multinomial logistic regression accuracy (linear separability proxy).

The dominant patterns in these compact subsets are interpretable: length/lexical production proxies (response size features) co-occur with morphological diversity and syntactic depth/complexity indicators, supporting a consistent interpretation that expert-verified CEFR distinctions in Cambridge contexts are strongly reflected in jointly varying surface production, morphological sophistication, and sentence-level structure.

4.7.4. Expert Validation Implications and Methodological Insights

The expert dataset validation reveals crucial insights about feature effectiveness and validation methodology in automated language assessment. The moderate clustering performance on expert-verified data (ARI 0.161–0.303) does not indicate feature inadequacy but rather reflects fundamental characteristics of professional CEFR annotation and dataset composition effects on unsupervised clustering outcomes.

Professional Annotation Impact. Expert verification creates tighter CEFR boundaries through consistent application of proficiency criteria, reducing the linguistic variation within levels that facilitates un-

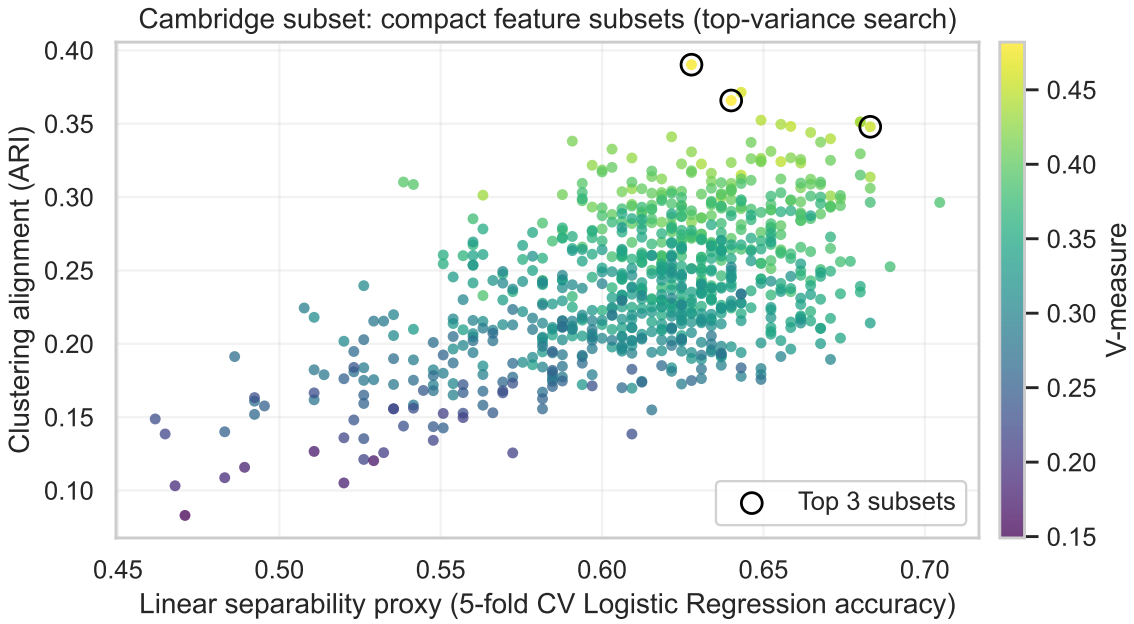

Fig. 2. Cambridge subset feature-subset search (N = 35, k = 6, 800 subsets). Each point is a subset; color encodes V-measure.

supervised clustering. This standardization, while beneficial for supervised learning applications, attenuates the feature discrimination signals that support cluster formation. The moderate clustering performance thus validates that our features capture subtle proficiency distinctions recognized by professional assessors rather than relying on extreme linguistic variations.

Distribution and Scale Effects. The Cambridge subset's superior clustering performance (ARI 0.303) compared to the larger Universal corpus (ARI 0.161) demonstrates that dataset size and CEFR distribution significantly impact clustering effectiveness. The balanced A2–C2 representation in the Cambridge subset provides optimal conditions for feature-based separation, while the intermediate-level dominance in the Universal corpus creates clustering challenges that reflect genuine assessment complexity rather than feature limitations.

Triangulated Validation Evidence: The combination of mixed-source clustering (99.5 % A1, 82.4 % C2 purity), moderate expert-verified alignment (ARI 0.303), and weak-to-moderate large-scale alignment (ARI 0.161) provides complementary-though not uniformly strong-evidence for feature effectiveness. These clustering results should be read alongside the supervised classification results (66.72 % accuracy, macro F1 = 0.69; Section 4.6.1), which constitute the primary performance evidence.

4.7.5. Authoritative Ground Truth Establishment

The expert dataset validation establishes benchmarks for automated CEFR assessment research through professional annotation alignment. The Cambridge subset's moderate clustering performance (ARI 0.303) provides a computational validation of linguistic features against Cambridge expert annotations. The consistency of proficiency patterns across both expert datasets supports the theoretical foundations of our multi-tier feature architecture against professional assessment standards.

This expert-verified validation addresses limitations in automated assessment research where feature effectiveness claims often rely on noisy or inconsistent annotations. The professional annotation alignment demonstrated through moderate clustering performance supports confidence in supervised classification

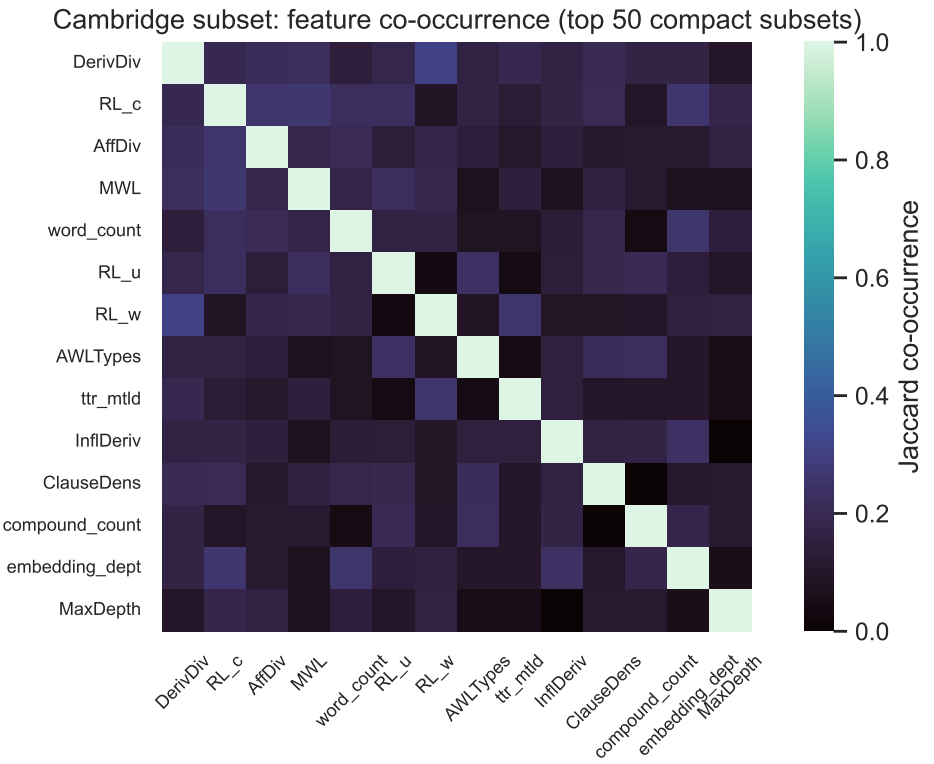

Fig. 3. Co-occurrence (Jaccard similarity) among the most frequently selected features in the top 50 Cambridge compact subsets

applications and provides benchmarks for cross-system comparison in automated language assessment research.

5. Discussion

5.1. Practical Implications and Feature Engineering

Our analysis reveals important insights for practical automated assessment deployment. Syntactic and morphological features are co-dominant predictors (Table 5), with systematic coordination between these dimensions and lexical sophistication providing enhanced discrimination for intermediate levels. The Tier 1 + 2 system achieves 66.72 % exact accuracy (macro F1 = 0.687) and 94.53 % adjacent accuracy; the top-ranked all-tiers Random Forest reaches 67.50 % with per-class F1 of 0.81 for A1 and 0.77 for C2, while B2 and C1 remain the most challenging levels (F1 = 0.60 and 0.59, respectively; Table 7).

5.2. Dataset Quality and Validation Methodology

Our triangulated validation approach across mixed-source, expert-verified, and scale-diverse datasets reveals important insights about feature effectiveness evaluation and the relationship between annotation quality and clustering performance in automated language assessment research.

5.2.1. Mixed-Source vs. Expert-Verified Dataset Performance

Expert verification does not necessarily increase unsupervised clustering alignment. In our experiments, the mixed-source dataset produced higher-purity extreme-level clusters (A1 and C2), while expert-verified Cambridge datasets produced more moderate alignment metrics (ARI 0.161–0.303). This pattern is consistent with tighter within-level variance and more standardized tasks in expert corpora, which can reduce

the separability that unsupervised objectives exploit. A further contributing factor is that professional CEFR assessment-including Cambridge examinations-employs a holistic rating approach in which examiners consider communicative achievement, effect on the target reader, and task fulfilment alongside purely linguistic criteria [1]. Our feature set captures only linguistic dimensions; the extra-linguistic factors embedded in expert ratings may account for part of the alignment gap.

5.2.2. Triangulated Evidence for Feature Effectiveness

Taken together, mixed-source structure, expert-verified alignment, and large-scale consistency provide complementary evidence: the features separate extreme levels well, align non-trivially with professional labels, and remain informative under realistic class imbalance.

5.3. Linguistic Insights and Feature Effectiveness Patterns

Our comprehensive analysis reveals systematic patterns in feature effectiveness that align with theoretical frameworks of second language acquisition and proficiency development while providing novel insights into computational approaches to language assessment.

5.3.1. Syntactic, Morphological, and Lexical Feature Co-Dominance

Syntactic features (mean sentence length, T-unit length, parse depth) achieve the highest individual Gini importance scores, while morphological features (affix diversity, derivational diversity, suffix diversity) are the most frequently represented category in the top 15 (Table 5). Ablation analysis confirms that morphological features carry the most unique information (−2.4 pp on removal), with lexical features second (−1.8 pp) and syntactic third (−0.9 pp; Table 3). This pattern — high individual importance for syntactic features but high collective uniqueness for morphological features — indicates that these two dimensions are correlated yet capture complementary aspects of proficiency development [5, 6].

Lexical sophistication features capture vocabulary development trajectories that align closely with CEFR proficiency frameworks. CEFR vocabulary distribution and academic word coverage show particularly strong discriminative power, and the combination of frequency-based and CEFR-aligned vocabulary measures provides comprehensive coverage of lexical sophistication [2].

5.3.2. Semantic Coherence and Discourse Organization

Tier 2 semantic features' meaningful contribution beyond deterministic measures demonstrates that text-level organization and coherence represent distinctive aspects of proficiency that complement lexical and morphological competence. Advanced learners create more coherent texts through sophisticated discourse organization, effective use of cohesive devices, and complex topic development strategies that are computationally detectable through machine learning-based analysis.

The moderate but consistent performance of semantic coherence features across all validation contexts supports theoretical frameworks emphasizing discourse competence as a key component of advanced language proficiency. This finding validates the investment in machine learning-based feature engineering while identifying areas for continued development in capturing higher-order linguistic competencies.

5.4. Practical Implications for Automated Assessment

Our validation results provide clear guidance for practical implementation of automated CEFR classification systems while identifying optimal application contexts and performance expectations.

5.4.1. Application-Specific Feature Selection Guidelines

The differential performance patterns across validation contexts indicate that optimal feature selection depends on specific application requirements and target populations. For placement testing and initial screening applications, the strong extreme-level discrimination (99.5 % A1, 82.4 % C2 purity) supports deployment for identifying clear beginners and advanced learners with high confidence.

For professional assessment support tools, the moderate expert validation performance (ARI 0.303) provides evidence for practical utility while indicating that automated systems should supplement rather than replace human professional judgment, particularly for intermediate-level assessment. The large-scale consistency results (ARI 0.161 across 1,815 samples) support deployment for broad-scale assessment applications while highlighting the need for appropriate confidence measures and human oversight.

5.4.2. Cost-Benefit Analysis of Multi-Tier Architecture

Our ablation analysis demonstrates clear cost-benefit relationships across feature tiers that inform practical deployment decisions. Tier 1 deterministic features provide strong baseline performance (64.53 % accuracy) with minimal computational requirements, supporting rapid screening applications. Tier 2 machine learning features contribute an additional 2.2 percentage points to reach the headline 66.72 %, while requiring moderate computational resources for embedding computation and semantic analysis. Preliminary Tier 3 experiments show a further 0.8 pp gain (see Conclusion), but this paper validates only Tiers 1 and 2.

The consistent performance gains from additional tiers justify the computational investment for applications requiring maximum accuracy. However, the strong Tier 1 baseline performance supports simplified deployment strategies for resource-constrained contexts or rapid-response applications where computational efficiency outweighs incremental accuracy improvements.

5.5. Limitations and Future Research Directions

Our comprehensive analysis reveals both the strengths and limitations of current feature engineering approaches while identifying clear directions for continued development in automated language assessment.

5.5.1. Intermediate-Level Assessment Challenges

The consistent challenges in discriminating intermediate proficiency levels (B1, B2) across all validation contexts highlight fundamental limitations in current feature engineering approaches. These levels represent transitional proficiency stages with overlapping linguistic characteristics and gradual developmental progression that resist clear computational distinction through current deterministic and machine learning-based features.

Future research should focus on enhanced feature engineering specifically targeting intermediate-level discrimination, potentially through genre-specific adaptations, temporal sequence analysis, or hybrid approaches combining multiple assessment modalities. The identification of B1/B2 discrimination as a key limitation provides clear direction for continued feature development while establishing realistic performance expectations for current approaches.

5.5.2. Cross-Domain and Cross-Cultural Generalization

Our validation employed English language texts from Cambridge examination contexts and mixed academic/informal sources, limiting generalization claims to similar populations and assessment contexts. Future research should examine feature effectiveness across diverse languages, cultural contexts, and assessment modalities to establish broader applicability of the multi-tier approach.

The consistent performance patterns across different dataset types within English assessment contexts suggest that the underlying theoretical framework may generalize effectively, but empirical validation across diverse linguistic and cultural contexts remains essential for establishing universal applicability of computational language proficiency assessment approaches.

Conclusion

This paper presents a comprehensive analysis of 133 linguistic features across two tiers (Tier 1 deterministic features and Tier 2 machine-learning-based semantic/discourse features) developed for the ISTOK testing system at Moscow Institute of Physics and Technology. Our evaluation covers a corpus of 3,205 texts

with triangulated validation against expert-verified Cambridge subsets, providing evidence that computational linguistic features can discriminate CEFR proficiency levels while remaining interpretable and practically deployable.

Triangulated validation yields complementary evidence. Mixed-source clustering shows strong discriminative structure for extreme levels (99.5 % A1, 82.4 % C2 cluster purity), while expert-verified clustering shows moderate alignment with Cambridge annotations (ARI 0.303) and weak-to-moderate alignment on the larger Universal CEFR corpus (ARI 0.161). The primary evidence for feature effectiveness comes from supervised classification: 66.72 % exact accuracy and 94.53 % adjacent accuracy on the full 6-class dataset (Tier 1 + 2).

Across analyses, morphological complexity and lexical sophistication are the most consistently informative feature families. Tier 1 features provide strong baseline performance (64.53 % accuracy) suitable for resource-constrained applications, while Tier 2 features add measurable discriminative power (2.2 percentage points) that can justify additional computation when accuracy is the primary objective.

Intermediate-level assessment (B1, B2) remains a consistent challenge, reflecting genuine overlap in proficiency boundaries. Future work should target intermediate discrimination through task- and genre-aware modeling.

We use the term "multi-tier" because the architecture is designed for N tiers: this paper validates two (deterministic Tier 1 and ML-based Tier 2). A planned **Tier 3** will explore large language model-based assessment for higher-order competencies such as pragmatic appropriateness, discourse organization, and metaphor use, deployed locally via Ollama to eliminate API costs. Preliminary experiments show modest incremental gains (Table 2); rigorous validation under the same protocol is ongoing and will be reported separately.

Overall, this study contributes transparent multi-dataset validation and expert-aligned benchmarks.

Data availability statement

The full dataset cannot be shared publicly due to institutional and privacy constraints. The High-Reliability Subset (357 samples, A1/B1/C2; Section 3.3), including learner texts, CEFR labels, and the complete 133-feature matrix, is publicly available to support replication, cross-study comparison, and benchmarking. Feature definitions, summary statistics, and additional materials may be obtained from the corresponding author on reasonable request.

See https://github.com/Tenfleques/Multi-Tier-Linguistic-Feature-Engineering-Sup for details.

Disclosure of interest

The authors report there are no competing interests to declare.

References

[1] Council of Europe. "Common european framework of reference for languages: Learning, teaching, assessment (CEFR)". (2026), [Online]. Available: https://www.coe.int/en/web/common-european-framework-reference-languages (visited on 02/27/2026).

[2] J. M. Imperial *et al.*, "UniversalCEFR: Enabling open multilingual research on language proficiency assessment", in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China: Association for Computational Linguistics, 2025, pp. 9714–9766. DOI: 10.18653/v1/2025.emnlp-main.491.

[3] K. North and M. Zampieri, "Features of lexical complexity: Insights from L1 and L2 speakers", *Frontiers in Artificial Intelligence*, vol. 6, 2023. DOI: 10.3389/frai.2023.1236963.

[4] T. Gaillat *et al.*, "Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach", *ReCALL*, vol. 34, no. 2, pp. 130–146, 2022. DOI: 10.1017/S095834402100029X.

[5] C. Wood, M. Garcia-Salas, C. Schatschneider, and M. Torres-Chavarro, "Morphological complexity in writing: Implications for writing quality and patterns of change", *American Journal of Speech-Language Pathology*, vol. 33, no. 3, pp. 1432–1442, 2024. DOI: 10.1044/2024_AJSLP-23-00265.

[6] A. Riemenschneider, Z. Weiss, P. Schröter, and D. Meurers, "The interplay of task characteristics, linguistic complexity, and language proficiency in high-stakes english as a foreign language writing", *TESOL Quarterly*, vol. 58, no. 2, pp. 775–801, 2024. DOI: 10.1002/tesq.3254.

[7] P. Xu, "Reconsidering the syntactic complexity measures on L2 spoken english: A multi-dimensional perspective", *Heliyon*, vol. 9, no. 6, e16856, 2023. DOI: 10.1016/j.heliyon.2023.e16856.

[8] A. Alzahrani, "Utility of Kolmogorov complexity measures: Analysis of L2 groups and L1 backgrounds", *PLOS ONE*, vol. 19, no. 4, pp. 1–25, 2024. DOI: 10.1371/journal.pone.0301806.

[9] M. Abdi Tabari, M. D. Johnson, and J. Gao, "Using automated indices of cohesion to explore the growth of cohesive features in L2 writing", *International Review of Applied Linguistics in Language Teaching*, vol. 63, no. 3, pp. 2169–2200, 2025. DOI: doi:10.1515/iral-2023-0185.

[10] X. Tang, H. Chen, D. Lin, and K. Li, "Incorporating fine-grained linguistic features and explainable AI into multi-dimensional automated writing assessment", *Applied Sciences*, vol. 14, no. 10, p. 4182, 2024. DOI: 10.3390/app14104182.

[11] J. F. Lohmann *et al.*, "Neural networks or linguistic features? - Comparing different machine-learning approaches for automated assessment of text quality traits among L1- and L2-learners' argumentative essays", *International Journal of Artificial Intelligence in Education*, vol. 35, no. 3, pp. 1178–1217, 2025. DOI: 10.1007/s40593-024-00426-w.

[12] M. Faseeh *et al.*, "Hybrid approach to automated essay scoring: Integrating deep learning embeddings with handcrafted linguistic features for improved accuracy", *Mathematics*, vol. 12, no. 21, p. 3416, 2024. DOI: 10.3390/math12213416.

[13] E. Kerz, D. Wiechmann, Y. Qiao, E. Tseng, and M. Ströbel, "Automated classification of written proficiency levels on the CEFR-scale through complexity contours and RNNs", in *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, 2021, pp. 199–209.

[14] N. Lagutina, K. Lagutina, A. Brederman, and N. Kasatkina, "Text classification by CEFR levels using machine learning methods and the BERT language model", *Automatic Control and Computer Sciences*, vol. 58, no. 7, pp. 869–878, 2024. DOI: 10.3103/S0146411624700329.

[15] A. Y. Uluslu, "Exploring hybrid linguistic features for turkish text readability", in *Proceedings of the 6th International Conference on Natural Language and Speech Processing ICNLSP 2023, Virtual Event, 16-17 December 2023*, 2023, pp. 223–232.

[16] S. Khan *et al.*, "Analyzing lexical complexity in learner corpora: A corpus-driven approach using part-of-speech tagging and dependency parsing", *Contemporary Journal of Social Science Review*, vol. 3, no. 4, pp. 1143–1170, 2025. DOI: 10.63878/cjssr.v3i4.1556.

[17] E. M. Bazanova, A. V. Gorizontova, N. N. Gribova, T. M. Chikake, and A. V. Samosyuk, "Development and prospects of national intelligent system for testing general language competencies deployed through neural network solutions", *Higher Education in Russia*, vol. 32, no. 8–9, pp. 147–166, 2023, in Russian. DOI: 10.31992/0869-3617-2023-32-8-9-147-166.

[18] K. Doi, K. Sudoh, S. Nakamura, and T. Watanabe, "Enhancing automated essay scoring with grammatical features using multi-task learning and item response theory", *Journal of Natural Language Processing*, vol. 32, no. 2, pp. 438–479, 2025. DOI: 10.5715/jnlp.32.438.

[19] Y. Bestgen, "Estimating lexical diversity using the moving average type-token ratio (MATTR): Pros and cons", *Research Methods in Applied Linguistics*, vol. 4, no. 1, p. 100 168, 2025. DOI: 10.1016/j.rmal.2024.100168.

[20] K. North, M. Zampieri, and M. Shardlow, "Lexical complexity prediction: An overview", *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–42, 2023. DOI: 10.1145/3557885.

[21] M. Burtsev *et al.*, "Deeppavlov: Open-source library for dialogue systems", in *Proceedings of ACL 2018, System Demonstrations*, 2018, pp. 122–127.

[22] D. Jurafsky and J. H. Martin, *Part-of-Speech Tagging*, 3rd edition. Stanford University, 2026.